Université Paris Cité
École doctorale de Sciences Mathématiques de Paris Centre (ED 386)
Laboratoire MAP5, UMR 8145 CNRS

Theory and Computation of Optimal Transport Variants

par Éloi TANGUY

Thèse de doctorat de Mathématiques Appliquées

Dirigée par Julie DELON et Agnès DESOLNEUX

Soutenue publiquement le 25 Novembre 2025 devant un jury composé de :

Julie DELON	Professeure, Université Paris Cité	Directrice
Agnès DESOLNEUX	Directrice de recherche CNRS, ENS Paris-Saclay	Directrice
Max FATHI	Professeur, Université Paris Cité	Examinateur
Kimia NADJAHI	Chargée de recherche CNRS, ENS Paris	Invitée
Rémi FLAMARY	Professeur, Ecole Polytechnique	Invité
Quentin MÉRIGOT	Professeur, Université Paris-Saclay	Rapporteur
Gabriel PEYRÉ	Directeur de recherche CNRS, ENS Paris	Examinateur
Gabriele STEIDL	Professeure, Technische Universität Berlin	Rapportrice

Université Paris Cité
45, rue des Saints-Pères
75006 Paris

Abstract

In this thesis, we tackle numerous theoretical and algorithmic challenges around variants of the Optimal Transport (OT) problem. We first investigate the use of the Sliced Wasserstein distance and the Mixture Wasserstein distances for different generative tasks, which lead to theoretical developments around their optimisation. Leveraging novel results from the field of non-smooth and non-convex optimisation, we show the convergence of (Stochastic) Gradient Descent methods for numerical resolution of the arising optimisation problems. We also introduce new OT frameworks which constrain the admissible set of OT maps and plans, first in the form of a problem finding a map that best transforms a source distribution into a target one under regularity constraints, and second as variants of the Sliced Wasserstein distance which constrain the admissible plans to satisfy some contraints related to their projections onto directions. Finally, we introduce generalisations of the Wasserstein barycentre problem, which defines barycentres of measures on different spaces and with respect to general costs, beyond the squared Euclidean distance. We provide a fixed-point algorithm for the numerical resolution of these barycentre problems with convergence guarantees.

We also provide contributions to the field of Reproducing Kernel Hilbert Spaces (RKHS), first studying the natural idea of representing gradients of convex functions within RKHS cones. Secondly, we introduce a novel explicit construction of universal kernels on compact metric spaces. We also define a new notion of approximate universality, and show that tractable discretisations of our universal kernels are approximately universal.

Keywords: Optimal Transport, Sliced Wasserstein Distance, Gaussian Mixture Models, Wasserstein Barycentres, Non-Smooth Non-Convex Optimisation, Reproducing Kernel Hilbert Spaces, Generative Modelling, Image Processing.

Résumé

Dans cette thèse, nous abordons de nombreux défis théoriques et algorithmiques autour des variantes du problème de Transport Optimal (TO). Nous étudions d'abord l'utilisation de la distance de *Sliced Wasserstein* et des distances *Mixture Wasserstein* pour différentes tâches génératives, ce qui nous conduit à des développements théoriques autour de leur optimisation. En nous appuyant sur des résultats récents dans le domaine de l'optimisation non-lisse et non-convexe, nous montrons la convergence des méthodes de descente de gradient (stochastique) pour la résolution numérique des problèmes d'optimisation qui en découlent. Nous introduisons également de nouvelles extensions du TO qui contraignent l'ensemble admissible des plans et des applications de transport, d'abord sous la forme d'un problème cherchant une application qui transforme au mieux une distribution source en une distribution cible sous des contraintes de régularité, et ensuite sous forme de variantes de la distance de *Sliced Wasserstein* qui contraignent les plans admissibles à satisfaire certaines contraintes liées à leurs projections sur des directions. Enfin, nous introduisons des généralisations du problème de barycentre de Wasserstein, qui définissent des barycentres de mesures sur différents espaces et par rapport à des coûts généraux, au-delà de la distance euclidienne au carré. Nous proposons un algorithme de point fixe pour la résolution numérique de ces problèmes de barycentre, avec des garanties de convergence.

Nous apportons également des contributions au domaine des espaces de Hilbert à noyau reproduisant (*RKHS* en Anglais), en étudiant en premier temps l'idée naturelle de représenter les gradients de fonctions convexes dans un cône d'un *RKHS*. Ensuite, nous introduisons une nouvelle construction explicite de noyaux universels sur des espaces métriques compacts. Nous définissons également une nouvelle notion d'universalité approximative et montrons que des discrétisations calculables de nos noyaux universels sont approximativement universelles.

Mots-Clefs: Transport Optimal, distance de *Sliced Wasserstein*, modèle de mélanges gaussiens, barycentres de Wasserstein, optimisation non-lisse non-convexe, espaces de Hilbert à noyau reproduisant, modélisation générative, traitement d'images.

Contents

0 Introduction 1

A Optimal Transport Discrepancies as Losses 43

A.I Reconstructing Discrete Measures from Projections 45

A.II Properties of Discrete Sliced Wasserstein Losses 53

A.III Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses 101

A.IV Differentiable Expectation-Maximisation and Applications to Gaussian Mixture Model Optimal Transport 125

B Variants of Optimal Transport Maps and Plans 161

B.I Constrained Approximate Optimal Transport Maps 163

B.II Sliced Optimal Transport Plans 203

B.III Sliced Gromov Wasserstein 251

C Optimal Transport Barycentres 261

C.I (Blind) Generalised Wasserstein Barycentres 263

C.II Computing Optimal Transport Barycentres 283

D Some Contributions to Kernel Methods 319

D.I A Gentle Introduction to RKHS 321

D.II On Gradients of Convex Functions in RKHS 329

D.III Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces 336

Future Directions 353

0

Introduction

0.1 Introduction to Optimal Transport and Beyond . . . 1
0.1.1 OT: a Modern and Multidisciplinary Field 1
0.1.2 Discrete OT Toolkit 2
0.1.3 General OT Toolkit 3
0.1.4 The Sliced Wasserstein Distance 5
0.1.5 Taming the Clarke Differential 7
0.2 Overview of the Thesis and Summary of Contributions . . . 9
0.2.1 Part A: Optimal Transport Discrepancies as Losses 9
0.2.2 Part B: Variants of Optimal Transport Maps and Plans 16
0.2.3 Part C: Optimal Transport Barycentres 21
0.2.4 Part D: Some Contributions to Kernel Methods 24
0.2.5 Summary of Open-Source Code Contributions 25
0.3 List of Preprints and Publications . . . 25
0.4 Résumé de la Thèse et des Contributions . . . 26
0.4.1 Partie A: Discrépances de Transport Optimal comme Fonctions de Perte 26
0.4.2 Partie B: Variantes de Plans et d'Applications de Transport Optimal 32
0.4.3 Partie C : Barycentres de Transport Optimal 37
0.4.4 Partie D : Quelques Contributions aux Méthodes à Noyaux 41
0.4.5 Résumé des Contributions en Code Source Ouvert 41

0.1 Introduction to Optimal Transport and Beyond

0.1.1 OT: a Modern and Multidisciplinary Field

Optimal Transport (OT) began as a practical optimisation problem in 1781 when [Mon81] posed the problem of optimally displacing heaps of soil. The problem remained largely unsolved until 1942 [Kan42] when Kantorovich (independently) studied an solved a relaxation of the problem. Kantorovich's work laid the foundation for modern OT and additionally introduced the powerful tool of duality in linear programming. Nowadays, the source and target objects in the OT problem are seen as probability distributions, and the optimal displacement cost defines a notion of discrepancy between probability measures, which respects the underlying geometry of the space. From a theoretical perspective, OT has been the subject of study of numerous monographs, including [AGS05; RR06; Vil09; San15; PC19b; Vil21; ABS+21; Pey25].

As a method for comparing probability distributions, OT has found applications in various fields, in particular in Machine Learning. In generative modelling, the objective is to generate samples that ressemble a given data distribution. To this end, OT and its variants have been used as a loss function quantifying the quality of the generations [BBR06; ACB17; Bou+17; DZS18; Des+19; Wu+19; Liu+19; RKB21; KSB22; KMK23].

More generally, OT has become a standard tool for practical tasks in Machine Learning and Statistics, including Domain Adaptation [Cou+16; Cou+17; MMS24a], Normalising Flows

[Ton+24], multi-label learning [Fro+15], omics data alignment [Dem+20; HPC22; Bun+24], Causal Inference [Wan+23; TGR24], Fairness constraints and testing [Gor+19; BYF20; Sil+20; Si+21] and two-sample tests [RGC17], just to name a few. Furthermore, OT theory has shed light on recent advances in deep learning, such as Diffusion Models [LS22], the Transformer architecture [FHP], The Resnet architecture [BPV24a], as well as Generative Adversarial Networks and Variational Auto-Encoders [GPC17].

OT theory is intimately linked to numerous fields, including Partial Differential Equations [Fig17; DF14; Bre18; San15], Stochastic Differential Equations [JKO98; Rig22; CLP23], Differential Geometry [Ott01; Vil09; AGS05; Mod17], Statistics [CNR25], Probability Theory [GJ20], Physics [Léo12; LBM24], as a non-exhaustive list.

In the following, we will introduce a few key concepts of OT theory germane to this work, which will be useful to present the contributions of this thesis. The rationale is to ensure the self-containedness of this thesis, as well as introduce notation and terminology for the summary of contributions and the following chapters. As a result, the reader familiar with OT may skip ahead.

0.1.2 Discrete OT Toolkit

We begin with a presentation of some OT concepts in the discrete setting, which is of paramount importance in applications. As introduced by [Mon81], the namesake Monge problem consists in finding an optimal permutation $\sigma^* \in \mathfrak{S}_n$ matching two families $(x_i)_{i=1}^n \in \mathcal{X}^n$ and $(y_i)_{i=1}^n \in \mathcal{Y}^n$ of n points in two spaces $\mathcal{X}$ and $\mathcal{Y}$, minimising the pairwise costs measured by a cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$:

$$\min_{\sigma \in \mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}). \tag{0.1}$$

We view this problem as minimising a transport cost between the two discrete measures $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. From a numerical perspective, this formulation requires a search over $n!$ possibilities, which is intractable. The Monge problem forbids separation of the mass units $\frac{1}{n}$, enforcing a one-to-one matching between (x_i) and (y_i). In Fig. 0.1, we illustrate the problem on a simple case in $\mathbb{R}^2$.

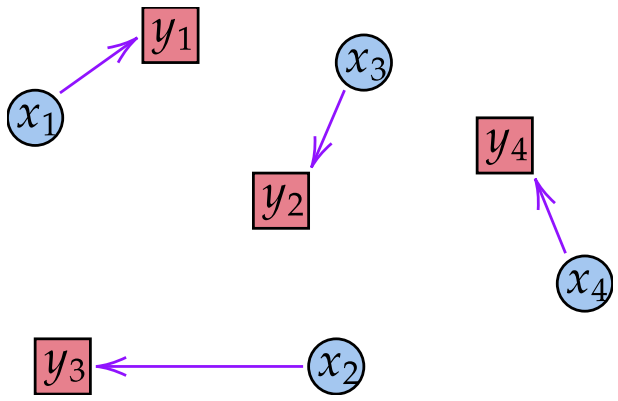


Figure 0.1: Example of the Monge problem in $\mathbb{R}^2$ for the squared Euclidean distance cost. The optimal permutation from the source points (blue circles) to the target points (red squares) is $\sigma^* = (1, 3, 2, 4)$.

We now consider two probability measures $\mu := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu := \sum_{j=1}^m b_j \delta_{y_j}$, with $a \in \triangle_n$ and $b \in \triangle_m$, where $\triangle_n$ is the n-simplex, i.e. the set of positive vectors of size n summing to one. The Kantorovich problem [Kan42] compares the two measures by finding an optimal transport plan $\pi^* \in \mathbb{R}_+^{n \times m}$, where $\pi^*_{i,j}$ represents the mass transported from x_i to y_j. This optimisation is performed under the constraint that all the mass of the source measure is transported, i.e. $\sum_{j=1}^m \pi^*_{i,j} = a_i$ for all i, and that the target receives all its mass, i.e. $\sum_{i=1}^n \pi^*_{i,j} = b_j$ for all j. We write the set of admissible transport plans as $\Pi(a, b)$. The Kantorovich problem is formulated

as follows:

$$\min_{\pi \in \Pi(a,b)} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j)\pi_{i,j}. \tag{0.2}$$

We provide an example in dimension two in Fig. 0.2. The Kantorovich problem is a linear

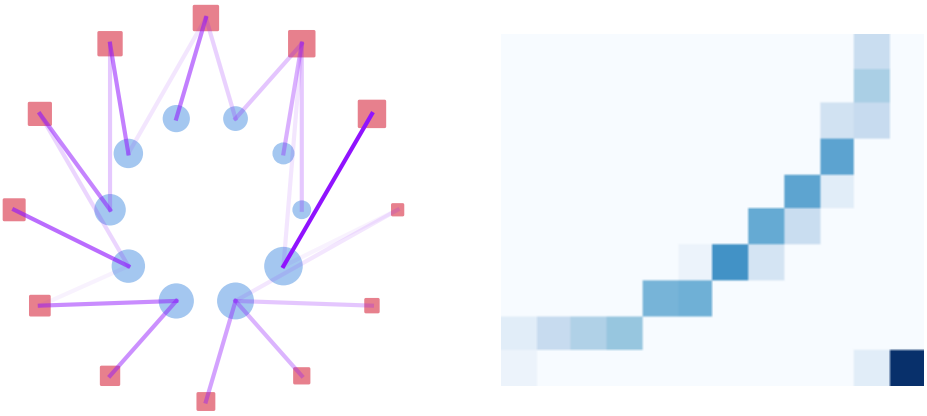


Figure 0.2: Example of the Kantorovich problem in $\mathbb{R}^2$. On the left, an optimal transport plan π^* from the source points (blue circles) to the target points (red squares) is represented by the purple lines. The size of the points represent their weight in the discrete measure. On the right, we represent the transport plan matrix π^* as a heatmap, where darker values signify more weight. The points are numbered by increasing size.

program (see [BT97]), in that the objective is linear in the transport plan π, and the constraints are linear equalities and inequalities ($\Pi(a,b)$ is a convex and compact polytope of $\mathbb{R}^{n\times m}$). The Network Simplex algorithm ([Orl97; Tar97], see also [PC19b, Section 3.5]) allows the numerical resolution of Eq. (0.2) in

$$\mathcal{O}\Big((n+m)nm\log(n+m)\log((n+m)\|C\|_\infty)\Big)$$

time, where $C := [c(x_i, y_j)]_{i,j}$. The elements $\pi \in \Pi(a,b)$ that cannot be written as $\pi = \frac{1}{2}\pi' + \frac{1}{2}\pi''$ for some $\pi' \neq \pi'' \in \Pi(a,b)$ are called *extreme points* of $\Pi(a,b)$. They are of particular interest since by [BT97, Theorem 2.7], there exists solutions that are extreme points. In particular, in the case of uniform measures with $n = m$, the Birkhoff Von Neumann Theorem [Bir46] states that the extreme points of $\Pi(a,b)$ are permutation matrices, which is why we say that the Kantorovich problem is a relaxation of the Monge problem.

0.1.3 General OT Toolkit

In this section, we present the OT problem in the so-called continuous setting, where the measures are not assumed to be discrete. To avoid excessive technicalities, we will work on Polish spaces.

Definition 0.1. A **Polish space** is a metric space $(\mathcal{X}, d_\mathcal{X})$ that is separable (there exists a countable family $(y_n) \in \mathcal{X}^\mathbb{N}$ such that for any $\varepsilon > 0$, $\forall x \in \mathcal{X}$, $\exists n \in \mathbb{N} : d_\mathcal{X}(x, y_n) \leq \varepsilon$) and complete (every Cauchy sequence converges in $\mathcal{X}$: any sequence $(x_n) \in \mathcal{X}^\mathbb{N}$ such that $\forall \varepsilon > 0$, $\exists N \in \mathbb{N} :$ $\forall n, m \geq N$, $d_\mathcal{X}(x_n, x_m) \leq \varepsilon$, then (x_n) converges to some $x \in \mathcal{X}$).

Given a Polish space $(\mathcal{X}, d_\mathcal{X})$, we denote by $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$, i.e. the set of probability measures defined on the Borel σ-algebra of $(\mathcal{X}, d_\mathcal{X})$. A useful notion of convergence in $\mathcal{P}(\mathcal{X})$ is the so-called **weak convergence of measures**, which corresponds to the weak* topology of $\mathcal{P}(\mathcal{X})$, seen as a subset of the dual of space $\mathcal{C}_b^0(\mathcal{X})$ of bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$ (see [AGS05, Remark 5.1.2] for details).

Definition 0.2. A sequence of measures $(\mu_n) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$ converges weakly to $\mu \in \mathcal{P}(\mathcal{X})$ if:

$$\forall \phi \in \mathcal{C}_b^0(\mathcal{X}),\ \int_{\mathcal{X}} \phi \mathrm{d}\mu_n \xrightarrow[n\longrightarrow+\infty]{} \int_{\mathcal{X}} \phi \mathrm{d}\mu,$$

and in this case we write $\mu_n \xrightarrow[n\longrightarrow+\infty]{w} \mu$.

We now fix two Polish spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$. The Kantorovich formulation of OT compares probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ by lifting a lower semi-continuous cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ to an OT cost $\mathcal{T}_c : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \longrightarrow \mathbb{R}_+$, defined as follows:

$$\mathcal{T}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y), \tag{0.3}$$

where $\Pi(\mu, \nu)$ is the set of measures $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ whose marginals are μ and ν, i.e. $\pi(A \times \mathcal{Y}) = \mu(A)$ and $\pi(\mathcal{X} \times B) = \nu(B)$ for all Borel sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. In our setting, [San15, Theorem 1.7] ensures the existence of optimal transport plans $\pi^* \in \Pi(\mu, \nu)$ attaining the cost $\mathcal{T}_c(\mu, \nu)$, and we write $\Pi_c^*(\mu, \nu)$ the set of such plans (the OT cost may be infinite). The OT cost has an alternative probabilistic expression:

$$\mathcal{T}_c(\mu, \nu) = \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{E}\left[c(X, Y)\right], \tag{0.4}$$

where the infimum is taken over all couplings (X, Y) of μ and ν.

In the particular case where $\mathcal{X} = \mathcal{Y}$ and for a cost $c(x, y) = d_{\mathcal{X}}(x, y)^p$ for some $p \geq 1$, the quantity $\mathrm{W}_p(\mu, \nu) := (\mathcal{T}_c(\mu, \nu))^{1/p}$ is referred to as the **Wasserstein distance** of order p, and W_p is a distance on the set $\mathcal{P}_p(\mathcal{X}) := \{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)^p \mathrm{d}\mu(x) < +\infty\}$ (see [Vil09, Definition 6.1]).

Of course, when considering discrete measures $\mu := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu := \sum_{j=1}^m b_j \delta_{y_j}$, the OT cost reduces to the discrete Kantorovich problem Eq. (0.2), where we identify a transport plan measure $\pi \in \Pi(\mu, \nu)$ as a transport plan matrix $P \in \mathbb{R}^{n\times m}$ defined by $P_{i,j} := \pi(\{(x_i, y_j)\})$.

To introduce the Monge problem in the "continuous setting", we will require the notion of push-forward measure:

Definition 0.3. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $f : \mathcal{X} \longrightarrow \mathcal{Y}$ be a measurable map. The **push-forward measure** of μ by f is the measure $f\#\mu \in \mathcal{P}(\mathcal{Y})$ defined by: $(f\#\mu)(B) := \mu(f^{-1}(B))$ for all Borel sets $B \subset \mathcal{Y}$.

In terms of random variables, if $X \sim \mu$, then $f\#\mu$ is the law of the random variable $Y := f(X)$. We illustrate the notion of push-forward measure in the discrete and absolutely continuous setting in Fig. 0.3.

The Monge problem consists in finding a measurable map $T : \mathcal{X} \longrightarrow \mathcal{Y}$ such that the push-forward measure $T\#\mu$ equals ν, and minimising the transport cost:

$$\min_{T:T\#\mu=\nu} \int_{\mathcal{X}} c(x, T(x)) \mathrm{d}\mu(x). \tag{0.5}$$

The Monge problem may not have a solution, for instance if $\mu = \delta_x$, then $T\#\mu = \delta_{T(x)}$ and cannot equal a measure ν that is not also a Dirac mass. We see the Kantorovich problem as a relaxation of the Monge problem, since if T^* is a solution of Eq. (0.5), then the transport plan $\pi := (I, T^*)\#\mu = \mathrm{Law}_{X\sim\mu}[(X, T^*(X))]$ is a solution of the Kantorovich problem Eq. (0.2). We say that π is *induced* by the map T^*. Extensive study of the Monge problem has been conducted across the years [Bre91; GM96; Pra07], and we focus on the celebrated Brenier Theorem [Bre91], specifically a simplification of the statement in [San15, Theorem 1.22]. We write $\mathcal{P}_p(\mathbb{R}^d)$ the set of probability measures $\mu \in \mathcal{P}(\mathbb{R}^d)$ such that $\int_{\mathbb{R}^d} \|x\|^p \mathrm{d}\mu(x) < +\infty$, $\mathscr{L}^d$ denotes the Lebesgue measure on $\mathbb{R}^d$, and "$\alpha \ll \beta$" means that α is absolutely continuous with respect to β (i.e. $\beta(A) = 0 \implies \alpha(A) = 0$ for all Borel sets A).

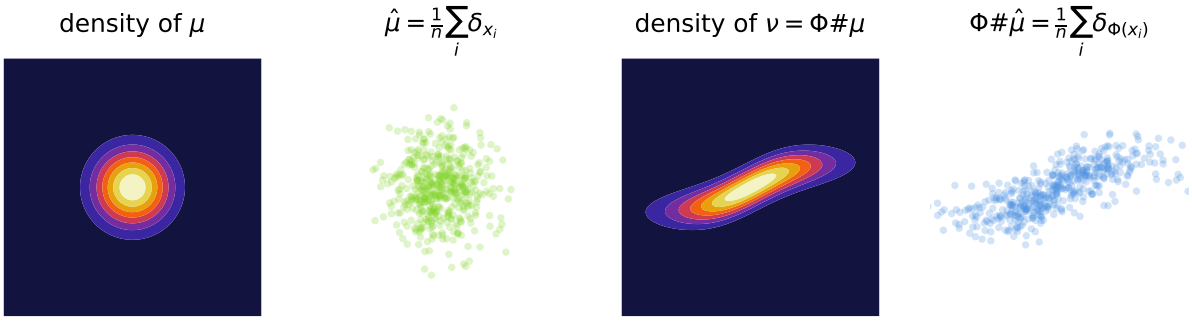
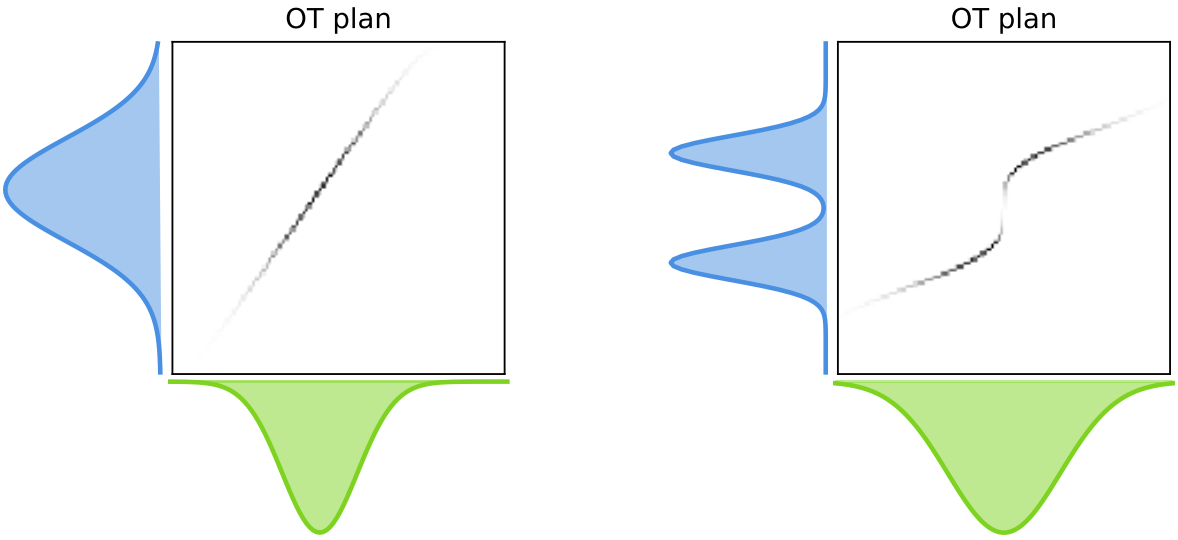


Figure 0.3: We consider the source measure $\mu := \mathcal{N}(0_{\mathbb{R}^2}, I_2)$ and represent its density p_μ on the first plot. We take $n = 500$ i.i.d. samples $(x_i)_{i=1}^n \sim \mu$ and show the discrete measure $\hat{\mu} := \frac{1}{n}\sum_i \delta_{x_i}$ on the second plot. Consider the diffeomorphism $\Phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by $\Phi(x_1, x_2) := (2x_1, \tanh(x_1) + \frac{1}{2}x_2)$. The measure $\nu =: \Phi\#\mu$ has the density $p_\nu(y) = |\det(J_\Phi(\Phi^{-1}(y)))|^{-1} p_\mu(\Phi^{-1}(y))$, where J_Φ is the Jacobian matrix of Φ, and we represent it on the third plot. Finally, $\hat{\nu} := \Phi\#\hat{\mu}$ has the expression $\hat{\nu} = \frac{1}{n}\sum_i \delta_{\Phi(x)}$ and is shown on the fourth plot.

Theorem 0.1 (Brenier's Theorem)**.** Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two probability measures such that $\mu \ll \mathcal{L}^d$ and consider the cost $c(x, y) := \|x - y\|_2^2$. Then the Kantorovich problem Eq. (0.4) between μ and ν has a unique solution, and it writes $\pi^* = (I, T^*)\#\mu$, with T^* the unique minimiser of the Monge problem Eq. (0.5). The map T^* is of the form $T^* = \nabla\varphi^*$, where $\varphi^* : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a convex function.

In Fig. 0.4b, we illustrate Brenier's Theorem in the one-dimensional case.



(a) Gaussian OT: the plan is induced by an affine map.

(b) The OT plan induced by an increasing function.

Figure 0.4: Example of the Monge problem in $\mathbb{R}$, with the source measure μ (green) on the bottom and target measure ν (blue) on the left.

0.1.4 The Sliced Wasserstein Distance

The Sliced Wasserstein (SW) distance [Rab+12] is a modification of the Wasserstein distance that leverages the simplicity of one-dimensional OT. Indeed, for $p \geq 1$ and between measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ for the cost $c(x, y) = |x - y|^p$ on $\mathbb{R}$, the "monotone plan" is optimal ([San15,

Theorems 2.9 and Proposition 2.17]), yielding the following closed-form expression:

$$\mathrm{W}_p^p(\mu,\nu) = \int_0^1 |F_\mu^{[-1]}(t) - F_\nu^{[-1]}(t)|^p \mathrm{d}t, \tag{0.6}$$

with $F_\alpha^{[-1]}$ the quantile function of $\alpha \in \mathcal{P}(\mathbb{R})$, defined as $F_\alpha^{[-1]}(t) := \inf\{x \in \mathbb{R} : \alpha((-\infty, x]) \geq t\}$ for $t \in [0,1]$. We define $\mathcal{P}^n(\mathbb{R})$ as the set of discrete uniform probability measures on n points in $\mathbb{R}$, and take $\mu := \frac{1}{n}\sum_i \delta_{s_i}$, $\nu := \frac{1}{n}\sum_i \delta_{t_i} \in \mathcal{P}^n(\mathbb{R})$. The expression in Eq. (0.6) simplifies to:

$$\mathrm{W}_p^p(\mu,\nu) = \frac{1}{n}\sum_{i=1}^n |s_{\sigma(i)} - t_{\tau(i)}|^p, \tag{0.7}$$

where $\sigma, \tau \in \mathfrak{S}_n$ are permutations of the indices of the (s_i) and (t_i) such that the $(s_{\sigma(i)})$ and $(t_{\tau(i)})$ are ordered increasingly.

Given now $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the idea of the SW distance is to compute the sliced distances $\mathrm{W}_p^p(P_\theta\#\mu, P_\theta\#\nu)$, where $\theta \in \mathbb{S}^{d-1}$ is a unit vector and $P_\theta : \mathbb{R}^d \longrightarrow \mathbb{R}$ is the linear map $P_\theta(x) := x \cdot \theta$. Since the sliced measures $P_\theta\#\mu$ and $P_\theta\#\nu$ are one-dimensional, the expression of $\mathrm{W}_p^p(P_\theta\#\mu, P_\theta\#\nu)$ is given explicitly by Eq. (0.6), which we illustrate in Fig. 0.5 in the case of discrete uniform measures with the same amount of points.

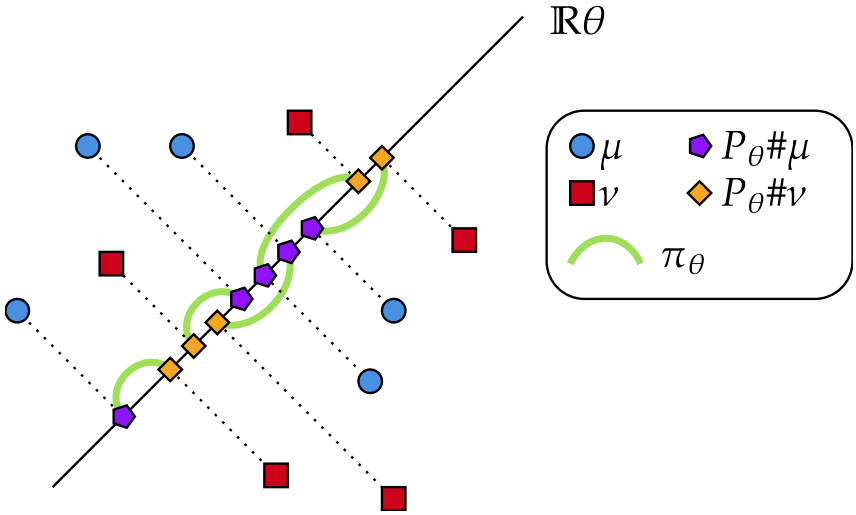


Figure 0.5: Illustration of the quantity $\mathrm{W}_2^2(P_\theta\#\mu, P_\theta\#\nu)$ for $\mu = \frac{1}{5}\sum_i \delta_{x_i}, \nu = \frac{1}{5}\sum_j \delta_{y_j}$ two discrete uniform measures in $\mathbb{R}^2$ with 5 points. The measure μ (blue circles) is projected onto the line $\mathbb{R}\theta$, yielding the one-dimensional measure $P_\theta\#\mu$ (purple pentagons). Likewise, the measure ν (red squares) is projected, giving $P_\theta\#\nu$ (orange diamonds). To compute the OT plan π_θ between the projected measures (thick green lines), we sort the projections $(P_\theta x_i)$ into $s_1 \leq \cdots \leq s_5$ and likewise $(P_\theta y_j)$ into $t_1 \leq \cdots \leq t_5$, and assign s_1 to t_1 and so on until s_5 to t_5.

The SW distance is then the expectation of this sliced distance over the unit sphere $\mathbb{S}^{d-1}$:

$$\mathrm{SW}_p^p(\mu,\nu) := \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(P_\theta\#\mu_\theta, P_\theta\#\nu_\theta)\mathrm{d}\sigma(\theta), \tag{0.8}$$

where σ is the uniform measure on $\mathbb{S}^{d-1}$. The SW distance is indeed a distance on $\mathcal{P}_p(\mathbb{R}^d)$ [Bon13], and its expression for discrete measures $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ is given by:

$$\mathrm{SW}_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n |P_\theta x_{\sigma_\theta(i)} - P_\theta y_{\tau_\theta(i)}|^p \mathrm{d}\sigma(\theta), \tag{0.9}$$

where $\sigma_\theta, \tau_\theta \in \mathfrak{S}_n$ sort the families $(P_\theta x_i)_{i=1}^n$ and $(P_\theta y_j)_{j=1}^n$ respectively. The integrals in Eqs. (0.8) and (0.9) are not tractable in general, and thus one resorts to Monte-Carlo estimation, sampling $\theta_1, \cdots, \theta_K \in \mathbb{S}^{d-1}$ uniformly, and approximating the SW distance by:

$$\mathrm{SW}_p^p(\mu,\nu) \approx \widehat{\mathrm{SW}}_K := \frac{1}{K}\sum_{k=1}^K \frac{1}{n}\sum_{i=1}^n |P_{\theta_k} x_{\sigma_{\theta_k}(i)} - P_{\theta_k} y_{\tau_{\theta_k}(i)}|^p. \tag{0.10}$$

The time complexity of the expression in Eq. (0.10) is $\mathcal{O}(Kn\log n + Knd)$, which is a substantial improvement over the $\mathcal{O}(n^3\log(n) + n^2 d)$ complexity of the Wasserstein distance from Eq. (0.2). This improvement comes at the cost of approximating the true SW distance, and the associated cost does not provide a transport plan, unlike usual OT costs.

0.1.5 Taming the Clarke Differential

Throughout this thesis, we will study optimisation problems minimising functions that are both non-convex and non-smooth. In this section, we explain the basics of a generalised (sub-)gradient called the Clarke differential, which is a powerful tool to study such problems [Cla90]. First, we present the notion of sub-differential for convex functions. Recall that a convex function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is differentiable almost-everywhere by [Roc97, Theorem 25.5]. At a point $a \in \mathbb{R}^d$ where f is differentiable, we have the inequality:

$$f(x) \geq f(a) + \nabla f(a) \cdot (x - a),$$

which is the first-order condition for convexity. The **convex sub-differential** of a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ at an arbitrary point $a \in \mathbb{R}^d$ is the compact convex set:

$$\partial_{\text{cvx}} f(a) := \{v \in \mathbb{R}^d : \forall x \in \mathbb{R}^d,\ f(x) \geq f(a) + v \cdot (x - a)\},$$

and by definition, any minimiser $x^\star$ of f must verify $0 \in \partial_{\text{cvx}} f(x^\star)$. Note that this definition and property do not require f to be convex, nor even differentiable (anywhere). When f is convex, at any point a of differentiability, we have $\partial_{\text{cvx}} f(a) = \{\nabla f(a)\}$, and that $\partial_{\text{cvx}} f(a)$ is always non-empty.

For non-convex functions, the convex sub-differential may be too restrictive (in the sense that it is often empty), which is why Clarke [Cla90] introduced another set-valued differential called the **Clarke differential**. First, recall that a **locally Lipschitz** function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a function that verifies:

$$\forall R > 0,\ \exists L_R > 0 : \forall x, y \in \overline{B}(0, R),\ |f(x) - f(y)| \leq L_R \|x - y\|,$$

where $\overline{B}(0, R)$ is the closed ball of radius R centred at the origin. This property can be understood as f being Lipschitz on all compact sets of $\mathbb{R}^d$, and by Rademacher's Theorem [Eva18, Theorem 3.2], it implies differentiability $\mathscr{L}^d$-almost everywhere. The **Clarke (sub-)differential** of a locally Lipschitz function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ at a point $a \in \mathbb{R}^d$ is defined[1] as the convex hull of limits of gradients of f:

$$\partial_C f(a) := \text{conv}\left\{v \in \mathbb{R}^d : \exists (x_n) \in D_f^{\mathbb{N}} : x_n \xrightarrow[n\longrightarrow+\infty]{} a \text{ and } \nabla f(x_n) \xrightarrow[n\longrightarrow+\infty]{} v\right\},$$

where D_f is the set of points of differentiability of f. If f is also convex, then by [Cla90, Proposition 4.3], $\partial_{\text{cvx}} f = \partial_C f$. Another formulation, which is convenient to relate to convex sub-differentials, is given by [Cla90, Proposition 1.5]: for a locally Lipschitz function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$:

$$\forall a \in \mathbb{R}^d,\ \partial_C f(a) = \left\{v \in \mathbb{R}^d : \forall w \in \mathbb{R}^d,\ f^\circ(a; w) \geq v \cdot w\right\},$$

where $f^\circ(a; w)$ is the **generalised directional derivative** of f at a in the direction w:

$$f^\circ(a; w) := \limsup_{\substack{y \longrightarrow a \\ t \downarrow 0}} \frac{f(y + tw) - f(y)}{t}.$$

In particular, this implies the intuitive relationship $\partial_{\text{cvx}} f \subset \partial_C f$. The strength of the Clarke differential for the study of local optima shines through the study of **Clarke critical points**, which are points $a \in \mathbb{R}^d$ such that $0 \in \partial_C f(a)$. The following result ([Cla90, Theorem 1.5]) states that non-critical points are not local minima:

[1] Our definition of the Clarke differential is specific to finite-dimensional spaces (we use the formulation from [Cla90, Theorem 8.1])

Theorem 0.2. Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a locally Lipschitz function, and $a \in \mathbb{R}^d$ be a point such that $0 \notin \partial_C f(a)$. Then any element $v \in \partial_C f(a)$ of minimal norm is such that:

$$\exists \varepsilon > 0 : \forall t \in (0, \varepsilon) : f(a - tv) < f(a).$$

We conclude this introduction to the Clarke differential with an example of a non-smooth and non-convex function $f : \mathbb{R} \longrightarrow \mathbb{R}$, representing its convex sub-differential and Clarke differential in Fig. 0.6.

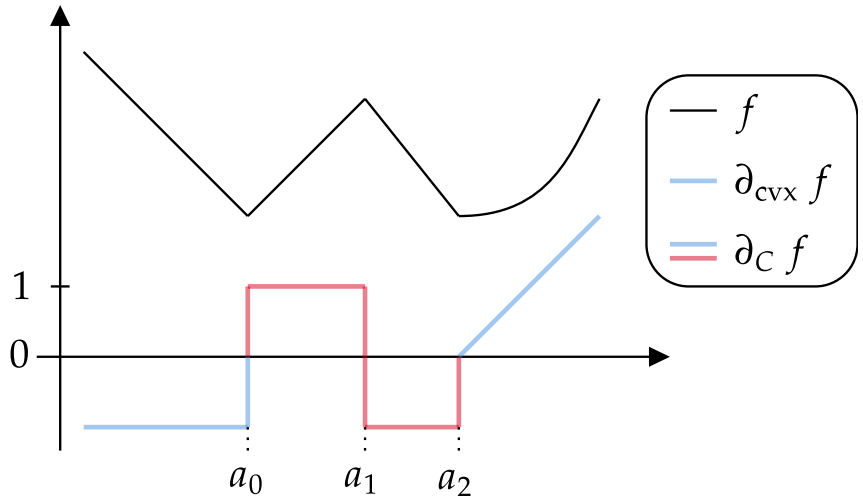


Figure 0.6: Example of a non-smooth and non-convex function f with its convex sub-differential (in blue) and Clarke differential (blue and red). Concerning the differentiability regions, first for all $a \in (-\infty, a_0)$, we have $\partial_{\text{cvx}} f(a) = \partial_C f(a) = \{f'(a)\}$ and likewise for $a \in (a_2, +\infty)$. However, for $a \in (a_0, a_1) \cup (a_1, a_2)$, the gradients of f are not sub-gradients, hence $\partial_{\text{cvx}} f(a) = \varnothing$ and $\partial_C f(a) = \{f'(a)\}$. At a_0, we have a local minimum and $\partial_C f(a_0) = [-1, 1]$, which we represent as a vertical segment. For the convex sub-differential, only the elements of $[-1, 0]$ are sub-gradients, thus $\partial_{\text{cvx}} f(a_0) = [-1, 0]$. At a_1, we have a local maximum, and $\partial_{\text{cvx}} f(a_1) = \varnothing$ while $\partial_C f(a_1) = [-1, 1]$. At a_2, we have another local minimum, and $\partial_{\text{cvx}} f(a_2) = \{0\}$ while $\partial_C f(a_2) = [-1, 0]$. The points a_0, a_1 and a_2 are Clarke critical.

Beyond the seminal works of Clarke (see [Cla90] for a monographic compilation), the Clarke differential is a starting point for convergence results in non-smooth, non-convex optimisation. Building on this tool allows the study of convergence of (stochastic) gradient descent algorithms and beyond for specific classes of "tame" non-convex non-smooth functions, see [BDL07; Bol+07; BDL09; MMM18; Dav+20; BP21; BHS22; BLP23; BHS23b] for an insight into this field. In this thesis, we will use the Clarke differential and results from this literature to study the convergence of (stochastic) gradient descent algorithms for the optimisation of non-smooth non-convex functions arising from OT problems. To illustrate, we consider two discrete measures $\mu_X = \sum_i a_i \delta_{x_i}, \nu_Y = \sum_j b_j \delta_{y_j}$ in $\mathbb{R}^2$ with $n = 200$ points $(X, Y \in \mathbb{R}^{200 \times 2})$, and study the landscape of the discrete 2-Wasserstein distance $\mathrm{W}_2^2(\mu_X, \nu_Y)$ as a function of the first point $x_1 \in \mathbb{R}^2$. We represent the illustration setting in Fig. 0.7a and the associated landscape surface in Fig. 0.7b, and see that the landscape is both non-smooth and non-convex, with numerous local minima.

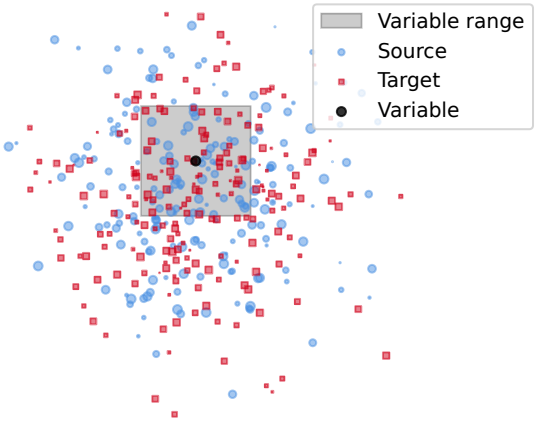
We mention a particular class of "simple" functions that are well suited for non-smooth non-convex analysis, called **semi-algebraic functions**. They are defined as functions $f : \mathbb{R}^a \longrightarrow \mathbb{R}^b$ whose graph

$$G := \{(x, f(x)) : x \in \mathbb{R}^a\} \subset \mathbb{R}^a \times \mathbb{R}^b$$

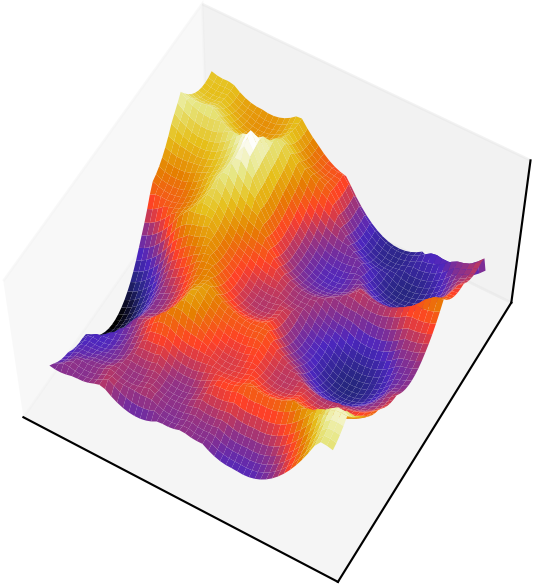
can be written with a finite union and intersection of sets S_k of the form:

$$S_k = \{y \in \mathbb{R}^{a+b} : p_k(y) \geq 0\},$$

where p_k is a $(a+b)$-variate polynomial. Such functions arise commonly in OT, for example it is easy to see that $X \longmapsto \mathrm{W}_2^2(\mu_X, \nu_Y)$ is a minimum of quadratic functions and is semi-algebraic.



(a) Measures $\mu_X = \sum_i a_i \delta_{x_i}$ (blue circles) and $\nu_Y = \sum_j b_j \delta_{y_j}$ (red squares) for the 2-Wasserstein surface illustration. The first point x_1 (large black circle) of μ_X will evolve in the highlighted grey square.



(b) Surface $x_1 \in \mathbb{R}^2 \longmapsto \mathrm{W}_2^2(\mu_X, \nu_Y)$ with the setting from Fig. 0.7a.

Figure 0.7: Energy landscape of the 2-Wasserstein distance with respect to the support of one of the measures.

0.2 Overview of the Thesis and Summary of Contributions

This thesis is primarily focused on the theoretical study and practical resolution of optimisation problems arising in the context of Optimal Transport and its applications. Part A studies the minimisation of OT discrepancies with respect to discrete measures, with theoretical consequences and practical applications in generative modelling. Part B introduces several variants of OT problems that constrain candidate transport maps or plans. Finally, Part C defines generalisations of Wasserstein barycentres and introduces algorithms for their computation. The last part of the thesis (Part D) is dedicated to an unrelated field of research, namely kernel methods, where we investigate representations of gradients of convex functions, and present novel explicit universal kernels on compact metric spaces. To provide a broad overview of the content of this thesis, we begin with a summary in Fig. 0.8, respecting the colour code of each part and chapter.

0.2.1 Part A: Optimal Transport Discrepancies as Losses

Machine Learning (ML) tasks are predominantly formulated as minimisation problems over some loss function. Chapters A.I to A.III study the minimisation of the discrete 2-Sliced Wasserstein distance from different viewpoints, leading to its study as a ML loss. The proverbial red tape of these three chapters is the two following energy functions:

$$\mathcal{E} := Y \in \mathbb{R}^{n\times d} \longmapsto \int_{\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n} (P_\theta y_{\sigma_\theta(i)} - P_\theta z_{\tau_\theta(i)})^2 \mathrm{d}\varpi(\theta) = \mathrm{SW}_2^2(\gamma_Y, \gamma_Z);$$

$$\mathcal{E}_p := Y \in \mathbb{R}^{n\times d} \longmapsto \frac{1}{p}\sum_{k=1}^{p} \frac{1}{n}\sum_{i=1}^{n} (P_{\theta_k} y_{\sigma_{\theta_k}(i)} - P_{\theta_k} z_{\tau_{\theta_k}(i)})^2 = \widehat{\mathrm{SW}}_p(\gamma_Y, \gamma_Z),$$

where $\gamma_Y := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ and $\gamma_Z := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ are the empirical measures associated to the data $Y = (y_1, \cdots, y_n) \in \mathbb{R}^{n\times d}$ and $Z = (z_1, \cdots, z_n) \in \mathbb{R}^{n\times d}$. The two energies $\mathcal{E}$ and $\mathcal{E}_p$ correspond to the discrete 2-Sliced Wasserstein distance $\mathrm{SW}_2(\gamma_Y, \gamma_Z)$ and its empirical Monte-Carlo approximation as a function of the support Y of one the measures.

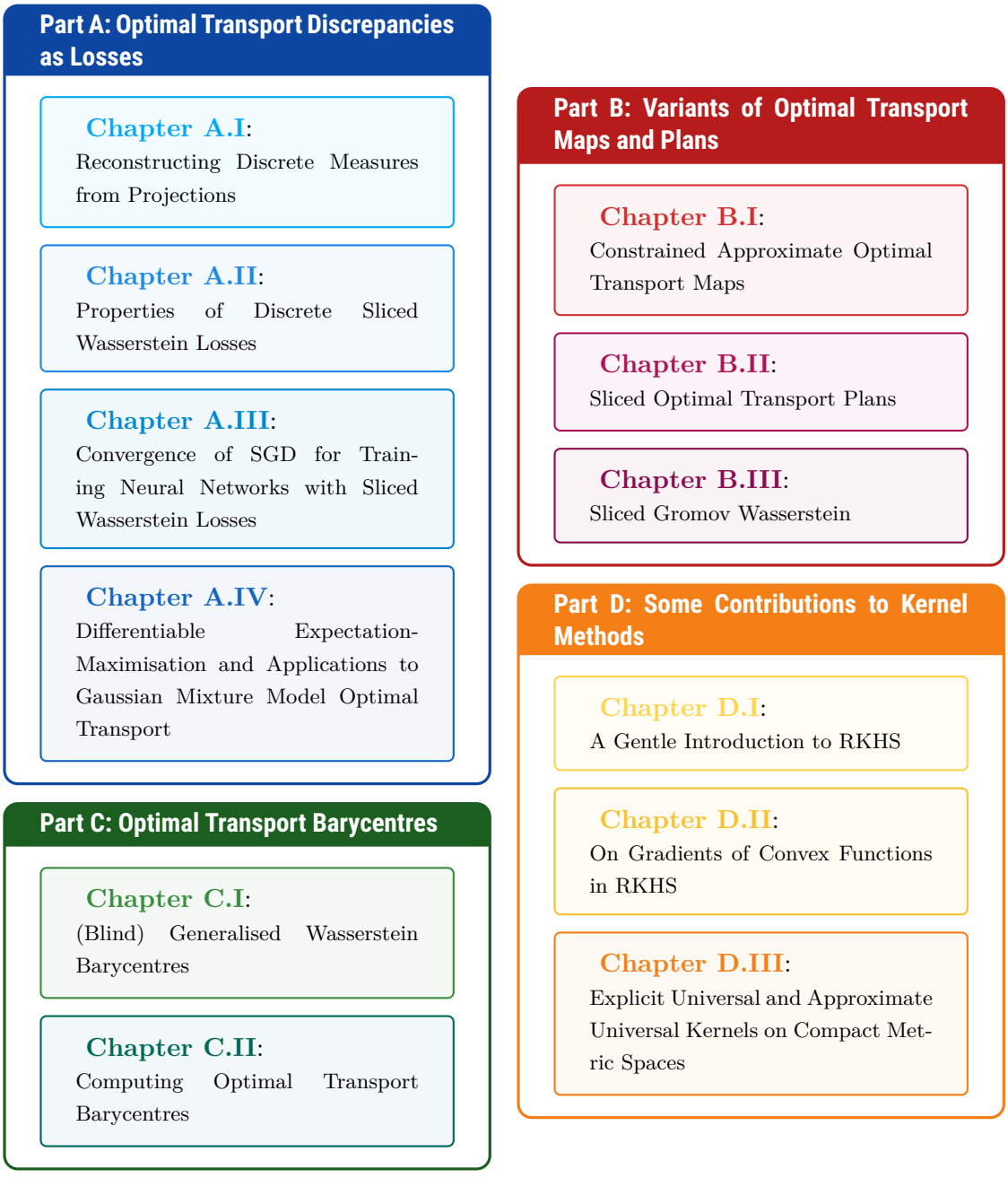


Figure 0.8: Overview of this thesis's parts and chapters.

0.2.1.1 Chapter A.I: Reconstructing Discrete Measures from Projections

Chapter A.I studies the inverse problem of recovering all probability measures $\gamma \in \mathcal{P}(\mathbb{R}^d)$ that have the same images by linear maps $P_i : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_i}$ as a fixed target measure $\gamma_Z := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. More formally, we are interested in the following solution set:

$$\mathcal{S} = \left\{ \gamma \in \mathcal{P}(\mathbb{R}^d) \mid \forall i \in [\![1, r]\!],\ P_i \# \gamma = P_i \# \gamma_Z \right\}.$$

The most important result states that when $\sum_i d_i > d$ and the P_i are taken randomly (each line being chosen independently following a law with density on $\mathbb{R}^d$), then almost-surely the original measure is the unique solution: $\mathcal{S} = \{\gamma_Z\}$. We also present pathological cases for specific maps P_i and show that when $\sum_i d_i \leq d$, there are always undesired solutions. Taking in particular p projections $P_i := \theta_i^\top$ for $\theta_i \in \mathbb{S}^{d-1}$, the reconstruction problem is exactly the problem of finding the global minima of $\mathcal{E}_p$, which leads to the following result:

Theorem. For $Z \in \mathbb{R}^{n\times d}$ with the (z_i) distinct and $(\theta_1, \cdots, \theta_p) \sim \mathfrak{w}^{\otimes p}$, then a.s.:

- If $p \leq d$, there are an infinity of $Y \in \mathbb{R}^{n\times d}$ such that $\gamma_Y \neq \gamma_Z$ and $\mathcal{E}_p(Y) = 0$.
- If $p > d$, then $\operatorname{argmin} \mathcal{E}_p = \{Y \in \mathbb{R}^{n\times d} : \gamma_Y = \gamma_Z\}$.

This intuitive property is of great use for the study of the discrete 2-Sliced Wasserstein distance, in particular it discourages strongly the use of $\mathcal{E}_p$ for a small number of projections p. Furthermore, it determines the set of global optima of $\mathcal{E}_p$ when $p > d$, which corresponds to the target measure.

0.2.1.2 Chapter A.II: Properties of Discrete Sliced Wasserstein Losses

Convergence of $\mathcal{E}_p$ to $\mathcal{E}$. Chapter A.II is focused on the energies $\mathcal{E}$ and $\mathcal{E}_p$, and begins with statistical properties about the convergence as $p \longrightarrow +\infty$. First, uniform convergence of $\mathcal{E}_p$ to $\mathcal{E}$ is established, extending the pointwise Monte-Carlo property to uniform convergence on compact sets, and showing a uniform central-limit convergence:

Theorem. For any compact set $\mathcal{K} \subset \mathbb{R}^{n\times d}$, we have:

- almost-surely, $\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{} 0$;
- $\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} G$, where G is a centred Gaussian process on $\mathcal{K}$,

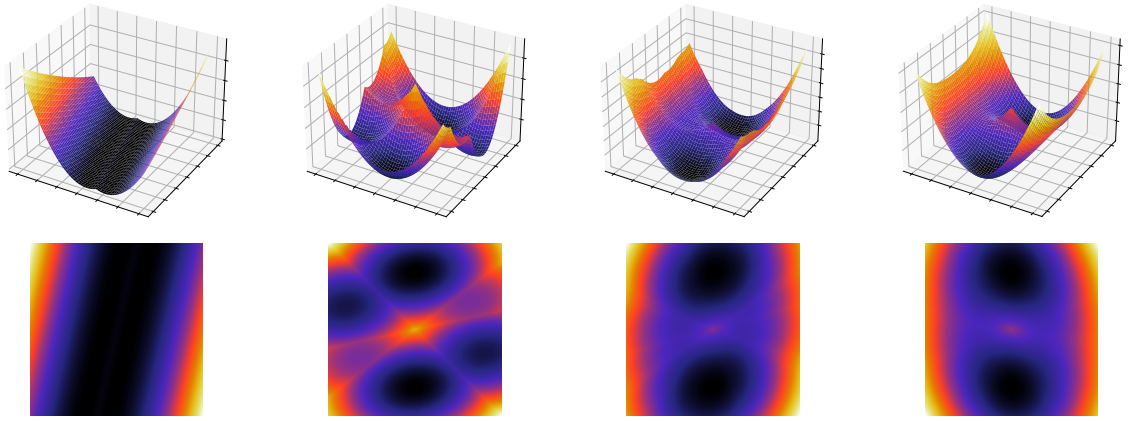
where $\|\cdot\|_{\ell^\infty(\mathcal{K})}$ is the supremum norm on $\mathcal{K}$, and "$\mathcal{L}$" denotes convergence in law.

Regularity questions. To study optimisation properties of $\mathcal{E}$ and $\mathcal{E}_p$, we first establish regularity results on these energies. We begin with a useful result on the Lipschitzness of the Kantorovich problem value $\mathrm{W}(\alpha, \beta; C) := \min_{\pi\in\Pi(\alpha,\beta)} \sum_{i,j} C_{i,j}\pi_{i,j}$ with respect to the weights α, β and cost matrix C:

Lemma. Given weights $\alpha, \overline{\alpha} \in \triangle_n$ and $\beta, \overline{\beta} \in \triangle_m$ (with positive entries) and cost matrices $C, \overline{C} \in \mathbb{R}^{n\times m}$, we have:

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C - \overline{C}\|_\infty + \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1).$$

This result not only yields local Lipschitzness of $\mathcal{E}$ and $\mathcal{E}_p$, it will also be used repeatedly throughout this thesis to establish regularity of OT energies. We further show that $\mathcal{E}$ and $\mathcal{E}_p$ are semi-concave (which is unfortunate for minimisation), and detail the structure of $\mathcal{E}_p$, writing it as a minimum of quadratic functions. In Fig. 0.9 we illustrate the landscapes of $\mathcal{E}_p$ and $\mathcal{E}$ in a



(a) $\mathcal{E}_p,\ p=1$ (b) $\mathcal{E}_p,\ p=3$ (c) $\mathcal{E}_p,\ p=10$ (d) $\mathcal{E}$

Figure 0.9: Landscapes $\mathcal{E}_p$ and $\mathcal{E}$ in a simplified case.

simple case, to showcase the convergence as p increases, and to illustrate the expression of $\mathcal{E}_p$ as a minimum of quadratics, which we refer to as a "cell structure". This property implies in particular that $\mathcal{E}_p$ is semi-algebraic, which will be useful repeatedly to study Stochastic Gradient Descent (SGD). As for $\mathcal{E}$, differentiability on a certain open set is known since [Bon+15a], with an explicit expression of the gradient.

Minimising the energies. As for optimisation properties, thanks to Chapter A.I, the global optima of $\mathcal{E}_p$ are almost-surely permutations of the target Z if p is large enough. Since SW_2 is a distance, the only global optima of $\mathcal{E}$ are the re-arrangements of Z as well. As for local optima, we leverage the cell structure of $\mathcal{E}_p$ to show that $\mathcal{E}_p$ is only differentiable in cell interiors, and that its points Y of differentiability verifying $\nabla\mathcal{E}(Y)=0$ (called "critical points") correspond to cell global minima. Furthermore, we explicit a function Ψ such that within the set of differentiability of $\mathcal{E}$, we have $\nabla\mathcal{E}(Y)=0 \iff \Psi(Y)=Y$. Given $(Y_p)_{p>d}$ critical points of $\mathcal{E}_p$ for $p>d$, they approach critical points of $\mathcal{E}$ in the following weak sense:

$$Y_p - \Psi(Y_p) \xrightarrow[p\longrightarrow+\infty]{\mathbb{P}} 0,$$

where the rate of decrease is controlled by an explicit upper-bound.

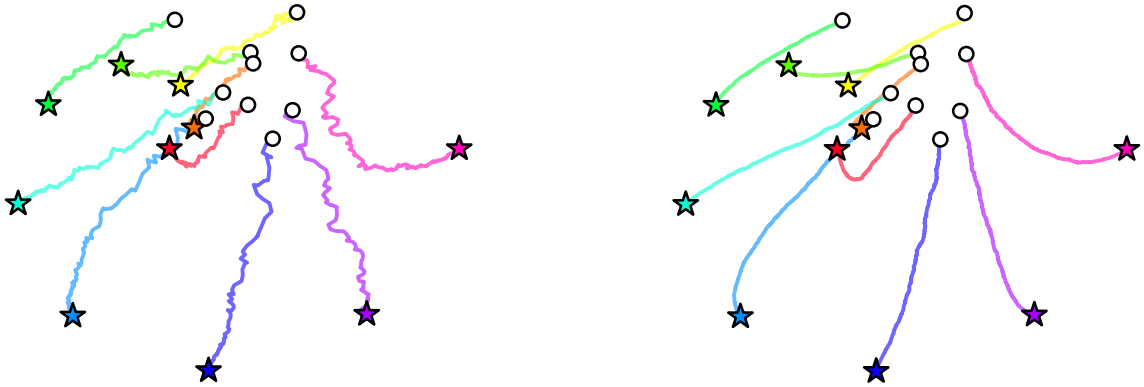
Convergence of SGD. Thanks to our regularity results, we were able to study the convergence of Stochastic Gradient Descent (SGD) on the energy $\mathcal{E}$, which consists in taking a random batch of p projections at each iteration, and computing a stochastic (sub-)gradient of $\mathcal{E}_p$. This study is a stepping stone to the understanding of the training of generative neural networks with the 2-Sliced Wasserstein loss, which is further studied in Chapter A.III. Furthermore, it provides a theoretical framework for the sliced matching algorithm proposed in [Rab+12]. This method consists in performing SGD on $\mathcal{E}$, and at convergence assigning to each point x_i from the initialisation to the closest point z_j it converged to. Using our results, we have in a weak sense some guarantees of convergence of this procedure. We summarise an informal and simplified formulation of our convergence results below, which rely on theoretical works on SGD convergence [BHS22; Dav+20].

Theorem.

- SGD iterations $(Y_\alpha^{(t)})_{t\in\mathbb{N}}$ with a fixed step α are such that their piecewise affine interpolations $(Y_\alpha(\cdot))$ approach the set of solutions of the differential inclusion $\dot{Y}(s) \in -\partial_C\mathcal{E}(Y(s))$ as the step size approaches 0.
- Noised SGD iterations $(Y_\alpha^{(t)})_{t\in\mathbb{N}}$ with a fixed step α are such that long-time subsequential limits approach the set of Clarke critical points of $\mathcal{E}$ as the step size approaches 0.

- Conditionally to their boundedness, decreasing-step SGD trajectories converge subsequentially to Clarke critical points of $\mathcal{E}$.

The first result signifies that for low (constant) step sizes, the trajectories are similar to solutions of the differential inclusion, which in the differentiable case would be equivalent to a gradient flow. The two other results justify convergence as $t \longrightarrow +\infty$ in a weak sense to a set of generalised critical points of $\mathcal{E}$, which in practice is observed to be global optima, as is the case in the illustration proposed in Fig. 0.10.



(a) SGD with step $\alpha = 0.3$.

(b) SGD with step $\alpha = 0.01$.

Figure 0.10: SGD on the energy $\mathcal{E}$ where the target measure is a 2d spiral (stars) with a random initialisation (circles).

0.2.1.3 Chapter A.III: Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses

Chapter A.III is an extension of the SGD studies from Chapter A.II to the training of generative neural networks with the 2-Sliced Wasserstein loss. Instead of minimising $\mathcal{E}$ with respect to data, we now would like to minimise the energy: $F^* := u \longmapsto \mathrm{SW}_2^2(T_u\#\mu, \nu)$, where μ is a fixed input data distribution (often chosen to be a Gaussian or uniform distribution) and ν is a target data distribution (typically a discrete dataset), with T_u a Neural Network (NN) parametrised by $u \in \mathbb{R}^{d_u}$. In practice (such as in [DZS18]), this minimisation is done by SGD, sampling at each iteration a batch of n input data points $X \sim \mu^{\otimes n}$ and n target data point $Y \sim \nu^{\otimes n}$, as well as a a random direction $\theta \sim \varpi$ (or a batch thereof), computing a gradient step of sample loss function:

$$f(u, X, Y, \theta) := \mathrm{W}_2^2(P_\theta\#T_u\#\gamma_X, P_\theta\#\gamma_Y)$$

with respect to u. We remind that $\gamma_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. This corresponds to minimising a mini-batch surrogate of F^*, since the associated population loss is:

$$F(u) := \int \mathrm{W}_2^2(P_\theta\#T_u\#\gamma_X, P_\theta\#\gamma_Y)\mathrm{d}\varpi(\theta)\mathrm{d}\mu^{\otimes n}(X)\mathrm{d}\nu^{\otimes n}(Y).$$

Under various technical assumptions and largely relying on the results and techniques of Chapter A.II, we show in Chapter A.III an analogue of the convergence results of Chapter A.II for the SGD of F. In some sense, these conclusions provide convergence guarantees for training generative models with the 2-Sliced Wasserstein loss.

0.2.1.4 Chapter A.IV: Differentiable Expectation-Maximisation and Applications to Gaussian Mixture Model Optimal Transport

Chapter A.IV focuses on another OT variant called the Mixture Wasserstein distance [DD20] which compares two Gaussian Mixtures Models (GMMs) $\mu := \sum_{k=1}^K w_k \mathcal{N}(m_k, \Sigma_k)$ and $\nu := \sum_{\ell=1}^L \overline{w}_\ell \mathcal{N}(\overline{m}_\ell, \overline{\Sigma}_\ell)$. This comparison is done by solving a Kantorovich problem between $\mu, \nu \in$

$\mathcal{P}_2(\mathbb{R}^d)$ for the square-Euclidean cost, restricting the couplings π to be themselves GMMs. An equivalent formulation is a discrete Kantorovich problem between the Gaussian components of μ and ν, where the cost between the components is the (Gaussian) 2-Wasserstein distance:

$$\mathrm{MW}_2^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)\cap\mathrm{GMM}_{2d}} \int_{\mathbb{R}^{2d}} \|x-y\|_2^2 \mathrm{d}\pi(x,y) = \min_{\pi\in\Pi(w,\overline{w})} \sum_{k,\ell} \pi_{k,\ell} \mathrm{W}_2^2(\mathcal{N}(m_k,\Sigma_k),\mathcal{N}(\overline{m}_\ell,\overline{\Sigma}_\ell)),$$

where GMM_d is the set of Gaussian Mixture probability measures on $\mathbb{R}^d$ with any (finite) number of components. In practical settings, one often works with point clouds of the form $X \in \mathbb{R}^{n\times d}$, and to use MW_2^2 to compare such data, one must first fit GMMs. As suggested by [DD20], the most natural method is the Expectation-Maximisation (EM) algorithm. In practical Machine Learning applications, model training relies on automatic differentiation, which motivates the study of the differentiability of the EM algorithm.

EM differentiation methods. In Chapter A.IV, we introduce three methods to compute automatic gradients of the output of the EM algorithm with respect to the input data $X \in \mathbb{R}^{n\times d}$. We view the EM steps as a fixed-point iteration scheme $\theta_{t+1} = F(\theta_t, X)$, where $\theta_t := (w^{(t)}, (m_k^{(t)})_k, (\Sigma_k^{(t)})_k)$ represents the parameters of the K-component GMM at step t. The first method is *Full Automatic Differentiation (AD)* , which simply performs automatic differentiation through all the EM steps. While this method can be costly when the number of iterations T is large, it is theoretically exact (barring the accumulation of numerical errors). We also propose two approximate methods which rely on the assumption that the EM algorithm converges to a fixed point $\theta^*(\theta_0, X)$ (which depends on the initialisation θ_0 and the data X). Using the implicit function theorem, we show that θ^* is a well-defined and differentiable function of X (under technical assumptions). Differentiating the property $\theta^*(X) = F(\theta^*, X)$ with respect to X yields the following expressions:

$$\frac{\partial\theta^*}{\partial X} = \frac{\partial F}{\partial\theta}(\theta^*,X)\frac{\partial\theta^*}{\partial X} + \frac{\partial F}{\partial X}(\theta^*,X) \quad\Longleftrightarrow\quad \frac{\partial\theta^*}{\partial X} = \left(I - \frac{\partial F}{\partial\theta}(\theta^*,X)\right)^{-1}\frac{\partial F}{\partial X}(\theta^*,X). \tag{0.11}$$

The first approximate method, called *Approximate Implicit gradient (AI)*, consists in replacing the unknown θ^* in Eq. (0.11) by the last iteration θ_T. This method is computationally extensive due to the matrix inversion, but provides a good approximation of the gradient in practice. The second approximate method, called *One-Step gradient (OS)*, consists in additionally neglecting the term $\partial_\theta F(\theta^*, X)$ in Eq. (0.11), which yields a much simpler expression, but the underlying assumption is not satisfied by EM in practice. We summarise the three methods in Table 1.

Method	Expression
Full Automatic Differentiation (AD)	compute $\partial_X[F_X^T(\theta_0)]$ with `auto-diff`
Approximate Implicit gradient (AI)	$J_{\mathrm{AI}} := (I - \partial_\theta F(\theta_T, X))^{-1}\,\partial_X F(\theta_T, X)$
One-Step gradient (OS)	$J_{\mathrm{OS}} := \partial_X F(\theta_T, X)$

Table 1: Summary of EM gradient approximation methods.

Importance of fixing uniform weights. To begin with, to justify the practical use of EM with MW_2^2, we used the discrete OT stability lemma from Chapter A.II to bound the error on the estimation of MW_2^2 when using GMM parameters estimated by EM. In practical applications, we observed that a variant of EM which does not update the GMM weights performed substantially better for optimisation. Delving in detail in the theory of certain particular cases, we show that this *fixed-weights EM* variant is better suited for problems which optimise the input of the EM algorithm in order to minimise MW_2^2. To illustrate this phenomenon, we consider a particle flow that transports data such as to fit a target GMM, namely minimising the energy $\mathcal{E}(X) := \mathrm{MW}_2^2(F_X^T(\theta_0, X), \nu)$, where ν is a target GMM, and F_X^T is the EM algorithm applied to X for T iterations with the initialisation θ_0. In Fig. 0.11, we observe that the flow converges to an unsatisfactory local minimum when the source and target GMMs have (different) non-uniform weights, while the case with uniform weights leads to convergence to the target GMM.

Based on these observations, we recommend the use of EM with GMM-OT for GMMs with uniform weights, leveraging a variant of EM which does not update the weights.

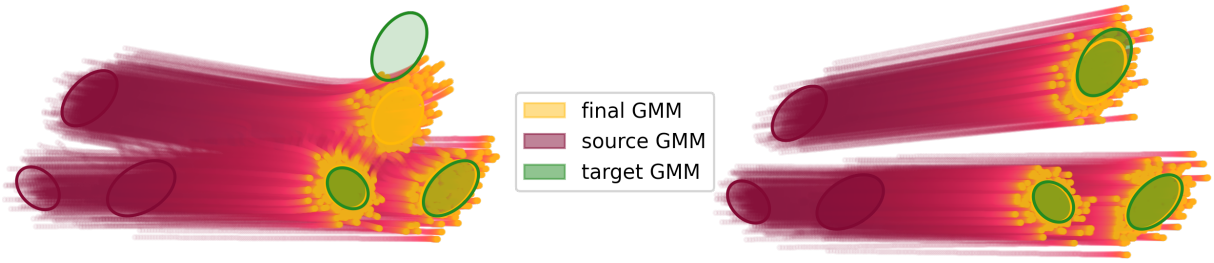
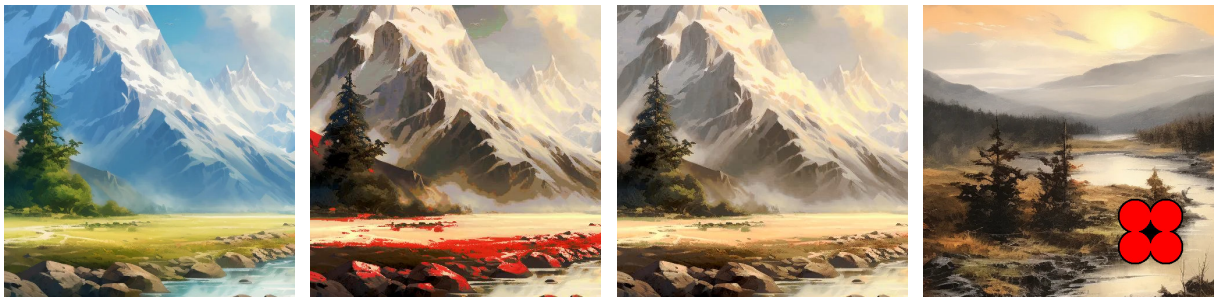


Figure 0.11: Illustration of the particle flow minimising $\mathcal{E}$ using AD for two different GMMs, with different non-uniform weights (Left) and uniform weights (Right).

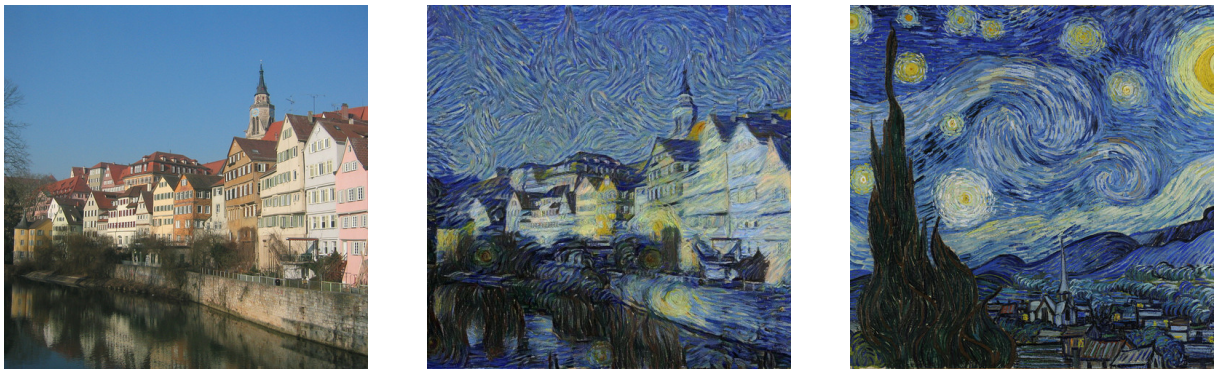
Unbalanced GMM-OT. As an extension of GMM-OT and as a possible approach to circumvent the weight optimisation issues, we introduce a new variant of MW_2^2 which penalises the marginal constraints instead of enforcing them. This Unbalanced Mixture Wasserstein discrepancy is closely inspired by Unbalanced OT [LMS18] and likewise allows the comparison of GMMs with different masses, in addition to providing robustness to outlier GMM components. To illustrate this, in Fig. 0.12, we present a toy colour transfer example displacing the colour distributions of a source image using GMM-OT and Unbalanced GMM-OT. The outlier object in the target image is successfully ignored by Unbalanced GMM-OT, while GMM-OT is forced to take it into account, yielding a leakage of red in the result.



(a) Source image (b) Balanced result (c) Unbalanced result (d) Corrupted target

Figure 0.12: Balanced and Unbalanced colour transfer.

Larger-scale applications. The chapter includes numerous illustrative examples, such as barycenter computation, other MW_2^2 flows with different gradient methods, and a quantitative study of EM convergence and the gradient approximation methods. We also delve into larger-scale applications, introducing an EM-MW_2^2-regularised Generative Adversarial Network [Goo+14; ACB17]. Moreover, we present a style transfer application in the spirit of [GEB15] which transforms a source image into the style of a target image by matching its features in a pre-trained VGG-19 network [SZ15] to the features of the target image using the EM-MW_2^2. In Fig. 0.13, we illustrate this method, given the style of Vincent van Gogh's *Starry Night* painting to a picture of the city of Tuebingen, Germany. Finally, inspired by [GLR18; LDD23], we introduce a novel texture synthesis method based on EM-MW_2^2 in patch spaces at different scales, see Fig. 0.14 for some results. Given the simplicity of the method and the fact that it does not require any training nor neural networks, the results are surprisingly good.



(a) Source image (b) Result image (c) Style target

Figure 0.13: EM-MW$_2^2$ neural style transfer example.

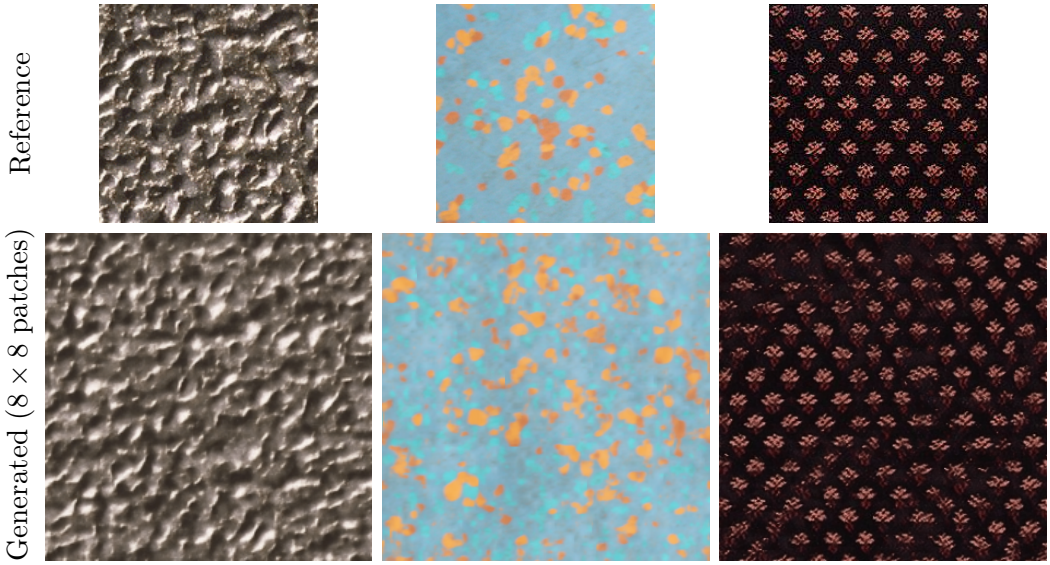


Figure 0.14: Multi-scale texture synthesis with $K = 4$ components.

0.2.2 Part B: Variants of Optimal Transport Maps and Plans

0.2.2.1 Chapter B.I: Constrained Approximate Optimal Transport Maps

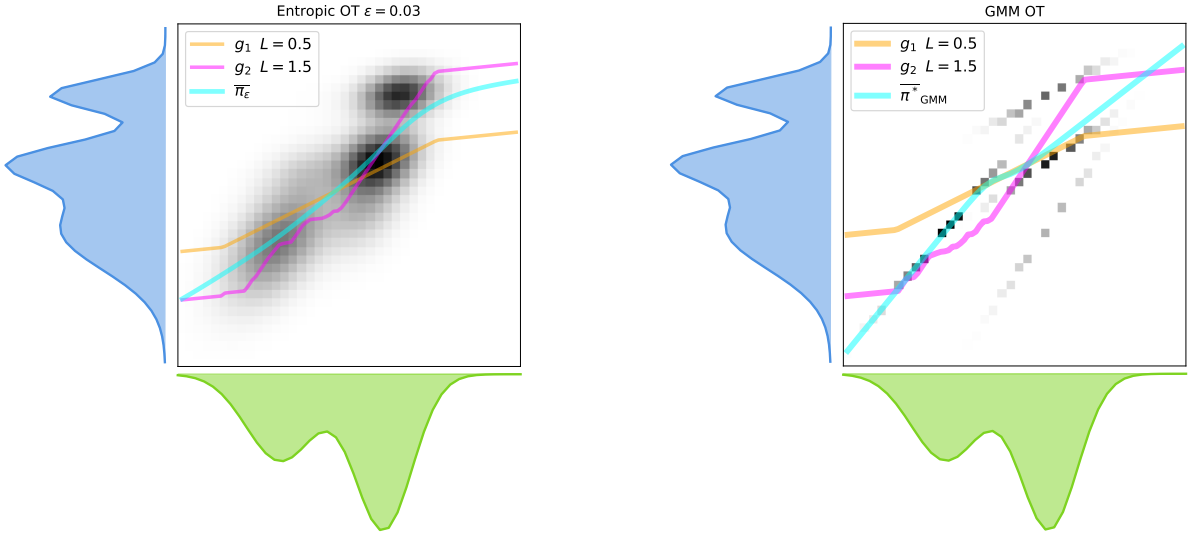
Many generative modelling and registration tasks can be formulated as searching for a function f that transports a source measure μ to a target measure ν as well as possible in some sense. As in [BBR06; ACB17; DZS18], this criterion can be chosen as minimising an OT discrepancy. In Chapter B.I, we study the problem of minimising $\mathcal{T}_c(g\#\mu, \nu)$ within a function class G of candidate maps g, which we see as a natural theoretical problem to consider when studying OT methods for generative modelling. In some sense, this problem consists in a relaxed Monge problem which penalises the constraint $g\#\mu = \nu$ instead of enforcing it. In practice, it is not desirable to have $g\#\mu = \nu$ (this corresponds to sampling known data from ν and not to generation), and in many cases, regularity of g is sought after to ensure robustness and interpretability. These constraints come into play within the function class G, which is typically a class of Neural Networks. The special case where G is the set of L-Lipschitz gradients of ℓ-strongly convex functions and $c(x, y) = \|x-y\|_2^2$ was studied in [PdC20], where optimal potentials were called Smooth Strongly Convex Nearest Brenier Potentials: in some sense, this problem can be seen as a relaxed Monge problem, where the constraint $g\#\mu = \nu$ is penalised, but the property of being the gradient of a convex function is enforced. A strength of this approach is that the learned map can be evaluated at any point of the source space, even outside the support of μ, making the strategy amenable to transfer learning and out-of-distribution generalisation.

Existence. The first natural question that we tackle in Chapter B.I is the existence of minimisers, which was surprisingly difficult. We show that existence holds under several technical assumptions, that we formulate loosely below:

> **Theorem.** Let $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ be lower semi-continuous and $(\mathcal{X}, d_{\mathcal{X}})$ be a locally compact Polish space. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ such that there exists $g \in G$ such that $\mathcal{T}_c(g\#\mu, \nu) < +\infty$. Suppose further that c is coercive and that G is a subset of the L-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}^d$ for some $L > 0$, such that G is stable under uniform limits on compact sets. Then there exists a minimiser $g^* \in G$ of the problem $\operatorname{argmin}_{g \in G} \mathcal{T}_c(g\#\mu, \nu)$.

In particular, we show that if G is a class of neural networks with weights confined to a compact set and with Lipschitz activation functions, then existence holds. Similarly, we show that the class of L-Lipschitz gradients of ℓ-convex functions satisfies the assumptions. An extension of this result to classes of functions which verify these conditions only on each part of a fixed partition of $\mathcal{X}$ is also provided, hence our results fully generalise that of [PdC20].

A plan approximation problem. Another motivation for this study is the problem of finding a map that best approaches a transport plan π between $\mu \in \mathcal{P}(\mathbb{R}^k)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$. For instance, given only a Gaussian Mixture Model (GMM) plan π between GMMs μ and ν ([DD20]), it is convenient in applications to approximate this plan by a map g. A natural way to do so is to solve the problem $\min_{g \in G} \mathcal{T}_C((I, g)\#\mu, \pi)$ with C a cost on $(\mathbb{R}^k \times \mathbb{R}^d)^2$, which is a particular case of the constrained map problem studied in Chapter B.I. We illustrate this idea to approximate an entropically regularised [Cut13] plan and a GMM [DD20] plan in Fig. 0.15.



(a) Plan approximation solutions for the Entropic-OT plan [Cut13].

(b) Illustration of plan approximation solutions for the GMM-OT plan [DD20].

Figure 0.15: Illustration of solutions of plan approximation problems for two different plans between Gaussian Mixtures. We compare the plans with $L = 1/2$ and $L = 3/2$-Lipschitz solutions, as well as to the barycentric projection of the given plans.

Perhaps counter-intuitively, for many ground costs C commonly used in OT, the cost $\mathcal{T}_C((I, g)\#\mu, \pi)$ in fact does not depend on the coupling π but only on the marginals μ and ν. This disappointing observation leads us to a result of great applicative interest, providing a broad sufficient condition on C for the cost $\mathcal{T}_C((I, g)\#\mu, \pi)$ to be independent on π. In particular, if C is chosen as a power of a p-norm, such as the ubiquitous squared Euclidean cost, then the cost $\mathcal{T}_C((I, g)\#\mu, \pi)$ is independent of π, thus the plan approximation problem is of little practical interest.

Focus on the L^2 case. In the case of the squared Euclidean cost and $\mathcal{X} = \mathbb{R}^d$, a natural candidate for a map g is an approximation of the barycentric projection of an OT plan π^* between μ and ν. We consider the following problem equivalence question:

$$\operatorname*{argmin}_{g\in G} \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x)-y\|_2^2 \mathrm{d}\pi(x,y) \stackrel{?}{=} \operatorname*{argmin}_{g\in G} \|g-\overline{\pi^*}\|^2_{L^2(\mu)}, \tag{0.12}$$

where π^* is a fixed OT plan between μ and ν, and $\overline{\pi^*}$ is its barycentric projection (a.k.a conditional expectation): $\overline{\pi^*}(x) = \mathbb{E}_{(X,Y)\sim\pi}[Y|X=x]$. [PdC20] showed that when $d=1$, the class of functions is that of L-Lipschitz gradients of ℓ-strongly convex functions, and μ is either absolutely continuous or discrete, then the two problems are equivalent. We extend this result to the case where G is any set of non-decreasing functions such that $g\#\mu \in \mathcal{P}_2(\mathbb{R})$, without assumptions on μ. We also present a counter-example to the problem equivalence question in dimension $d=2$ for the favourable case where G is the set of *monotone* continuous functions (which is a necessary property of Lipschitz gradients of convex functions).

Practical optimisation questions. To study the convergence of practical methods solving the map approximation problem, we build upon the discrete OT regularity results from Chapter A.II and provide details about the Clarke differential of the discrete Kantorovich cost with respect to the cost matrix:

Proposition. Consider weights $a\in\triangle_n$, $b\in\triangle_m$, and the discrete Kantorovich cost:

$$W(a,b,\cdot) := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times m} & \longrightarrow & \mathbb{R} \\ M & \longmapsto & \min\limits_{\pi\in\Pi(a,b)} \pi\cdot M \end{array} \right. .$$

The map $W(a,b,\cdot)$ is semi-algebraic, Lipschitz, and its Clarke sub-gradient is semi-algebraic and writes for $M\in\mathbb{R}^{n\times m}$:

$$\partial_C W(a,b,\cdot)(M) = \operatorname*{argmin}_{\pi\in\Pi(a,b)} \pi\cdot M.$$

Thanks to this result and to the recent advances of [BLP23], we were able to use similar techniques to Chapter A.III to study SGD on the following minibatch loss:

$$F(\theta) := \int \mathcal{T}_c(\delta_{h(\theta,X^{(n)})}, \delta_{Y^{(m)}}) \mathrm{d}\mu^{\otimes n}(X^{(n)}) \mathrm{d}\nu^{\otimes m}(Y^{(m)}),$$

with the goal of training a Neural Network $h(\theta,\cdot)$ with semi-algebraic and Lipschitz activations, and with parameters θ in a compact set $\Theta\subset\mathbb{R}^p$. Glossing over technical details, we show that the decreasing-step SGD iterates (θ_t) almost-surely converge sub-sequentially to a Clarke critical point of F. We also propose a simple method to look for maps g within a Reproducing Kernel Hilbert Space, and our regularity results along with usual non-convex Gradient Descent (GD) theory from [EN97] allowed us to show that decreasing-step GD iterates (θ_t) converge sub-sequentially to a Clarke critical point of the associated energy.

0.2.2.2 Chapter B.II Sliced Optimal Transport Plans

The computational efficiency of the Sliced Wasserstein distance comes at the cost of not obtaining a transport plan nor map between the measures. Some suggestions to define sliced plans have been proposed in the discrete case by [Rab+12; Mah+23; Liu+24], however numerous theoretical and practical questions remained open. In Chapter B.II, we introduce different formalisations of the ideas introduced in [Mah+23; Liu+24], with an extensive theoretical study.

Intuition on sliced transport plans. The two objects that this chapter focuses on are the (Min-)Pivot Sliced and Expected Sliced discrepancies. To introduce them, the first step is to see

how to define a transport plan between two measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ based on their projections $P_\theta \# \mu_i$ on an axis $\theta \in \mathbb{S}^{d-1}$. In the discrete case, it would be natural to simply operate the same assignment between the points in $\mathbb{R}^d$ as was done on the line, as illustrated in Fig. 0.16.

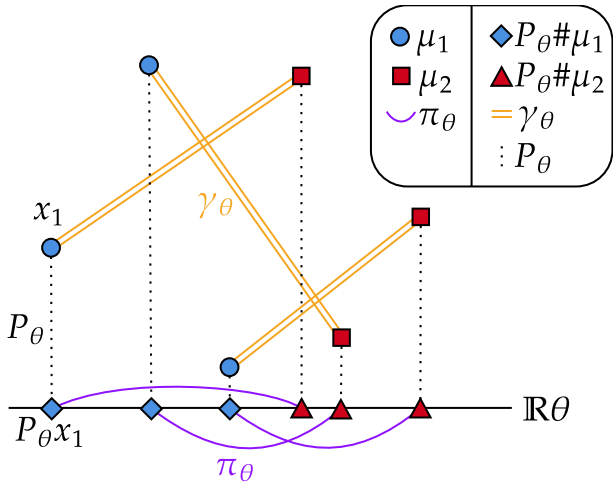


Figure 0.16: Given discrete measures with uniform weights supported on points with distinct projections on the line $\mathbb{R}\theta$, the one-dimensional OT plan between the projections can be lifted to a plan between the two measures without ambiguity.

Introduced by [Mah+23], the lifting technique from Fig. 0.16 is ill-defined when some points have the same projection, or for general measures in $\mathcal{P}_2(\mathbb{R}^d)$. The idea of the Pivot Sliced discrepancy is to resolve this ambiguity by selecting plans that have minimal cost using a pivot measure (as in [Mah+23; NP23]). Contrastingly, [Liu+24] propose to enforce an independent coupling to resolve ambiguity. Before turning to formal mathematical definitions, we illustrate the two approaches in Fig. 0.17.

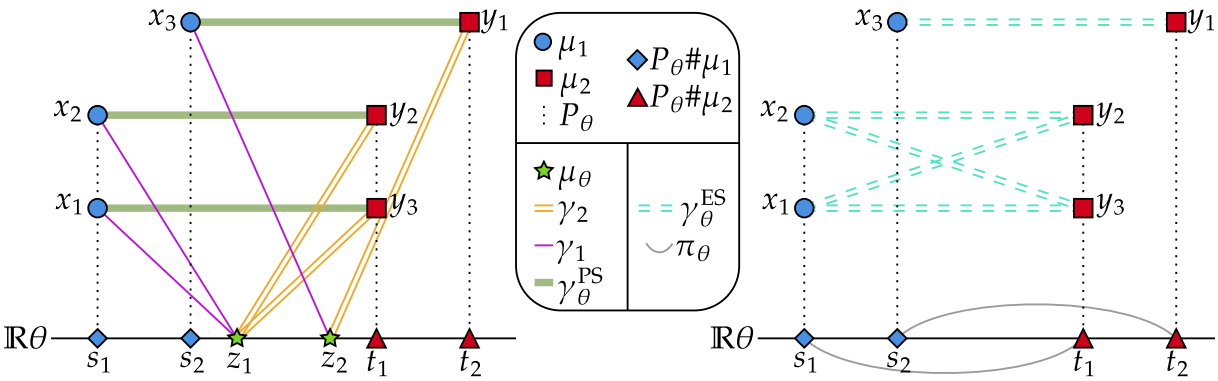


Figure 0.17: Comparison of the Pivot Sliced (left) and Expected Sliced (right) discrepancies for two measures μ_1, μ_2 with uniform weights supported three points with non-distinct projections on the line $\mathbb{R}\theta$. The projections are $P_\theta \# \mu_1 = \frac{2}{3}\delta_{s_1} + \frac{1}{3}\delta_{s_2}$ and $P_\theta \# \mu_2 = \frac{2}{3}\delta_{t_1} + \frac{1}{3}\delta_{t_2}$. **Pivot Sliced:** first, we compute the middle measure, here $\mu_\theta = \frac{2}{3}\delta_{(s_1+t_1)/2} + \frac{1}{3}\delta_{(s_2+t_2)/2}$, then for $i \in \{1, 2\}$ we take γ_i an optimal plan between μ_θ and μ_i (in this case, both are unique). We now determine a plan $\gamma_\theta^{\mathrm{PS}} \in \Pi(\mu_1, \mu_2)$ that minimises the transport cost $\int \|x - y\|_2^2 \mathrm{d}\gamma_\theta^{\mathrm{PS}}(x, y)$ amongst plans that are consistent with γ_1 and γ_2: here, this means that x_3 must be matched to y_1 because z_2 is matched only to x_3 in γ_1 and z_2 is only matched to y_1 in γ_2. For z_1, there is ambiguity, since it is assigned to both x_1 and x_2 in γ_1 and to both y_2 and y_3 in γ_2: here we take the assignement that matches x_1 to y_3 and x_2 to y_2, since it minimises the transport cost. **Expected Sliced:** we compute the 1D OT plan π_θ between the projections $P_\theta \# \mu_1$ and $P_\theta \# \mu_2$, then lift it: first, s_1 is assigned to t_1 by π_θ, hence we match x_1 and x_2 to y_2 and y_3 with the product (independent) coupling. Continuing, s_2 is matched to t_2, hence we link x_3 to y_1 (no ambiguity this time).

Formal definition of the Pivot Sliced discrepancy. Formally, the Pivot Sliced discrepancy between $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as follows:

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) := \min_{\rho \in \mathcal{P}_2(\mathbb{R}^{3d}):\ \rho_{0,1} \in \Pi^*(\mu_\theta, \mu_1),\ \rho_{0,2} \in \Pi^*(\mu_\theta, \mu_2)} \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2),$$

where for $\rho \in \mathcal{P}_2(\mathbb{R}^{3d})$ and $i \in \{1, 2\}$, $\rho_{0,i} \in \mathcal{P}_2(\mathbb{R}^{2d})$ refers to the bi-marginal $P_{0,i}\#\rho$ with $P_{0,i} := (y, x_1, x_2) \longmapsto (y, x_i)$. The measure $\mu_\theta \in \mathcal{P}_2(\mathbb{R}^d)$ is the 2-Wasserstein middle between $Q_\theta\#\mu_1$ and $Q_\theta\#\mu_2$ with $Q_\theta := x \longmapsto (\theta^\top x)\theta$. Finally, $\Pi^*(\mu_\theta, \mu_i)$ denotes the set of optimal transport plans between μ_θ and μ_i for the squared Euclidean cost. We show that PS_θ verifies all the axioms to be a distance apart from the triangle inequality and that it is an upper-bound of W_2. After proving novel results on the related ν-based Wasserstein distance [NP23], we demonstrate that PS_θ is lower semi-continuous (but not continuous) with respect to the weak convergence of measures. We show that PS_θ is equal to another discrepancy CW_θ that we refer to as the Constrained Wasserstein distance, which is defined by adding a constraint along the slice $\mathbb{R}\theta$ in the Kantorovich problem:

Theorem. For any $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, we have:

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) = \mathrm{CW}_\theta^2(\mu_1, \mu_2) := \min_{\substack{\omega \in \Pi(\mu_1, \mu_2) \\ (P_\theta, P_\theta)\#\omega = \pi_\theta}} \int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\omega(x_1, x_2),$$

where $P_\theta := x \longmapsto \theta^\top x$ and π_θ is the unique OT plan between $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$.

Monge formulation of the Pivot Sliced discrepancy in the discrete case. In the case where $\mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mu_2 = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, it is natural to look for a Monge formulation for the problem defining the Pivot Sliced discrepancy. The question is the following: in this setting, is an optimal plan for PS_θ necessarily a permutation? By showing a constrained version of the Birkhoff-von Neumann theorem [Bir46] (which states that the extreme points of the polytope of doubly stochastic matrices are the permutation matrices), we were able to answer positively to this question, as stated below.

Theorem. For any $(x_1, \cdots, x_n) \in \mathbb{R}^{n \times d}$, $(y_1, \cdots, y_n) \in \mathbb{R}^{n \times d}$ and $\theta \in \mathbb{S}^{d-1}$, it holds:

$$\mathrm{PS}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{(\sigma,\tau) \in \mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2,$$

where $\mathfrak{S}_\theta(X, Y) \subset \mathfrak{S}_n^2$ is the set of permutation pairs (σ, τ) such that:

$$P_\theta x_{\sigma(1)} \leq \cdots \leq P_\theta x_{\sigma(n)}, \quad P_\theta y_{\tau(1)} \leq \cdots \leq P_\theta y_{\tau(n)}.$$

Min-Pivot Sliced discrepancy. Following the ideas of [Mah+23], we introduce the Min-Pivot Sliced discrepancy $\min \mathrm{PS}$ as the minimum of PS_θ over all $\theta \in \mathbb{S}^{d-1}$. We show that this discrepancy inherits the properties of PS_θ and that it equals W_2 in certain favourable discrete cases. We observe numerically that the triangle inequality does not hold for $\min \mathrm{PS}$ either on a well-chosen example.

Expected Sliced discrepancy. We were able to generalise the lifting technique from [Liu+24] illustrated in Fig. 0.17 to generic mesures, thereby defining a discrepancy LS_θ on $\mathcal{P}_2(\mathbb{R}^d)$ and lifted plans $\gamma_\theta[\mu_1, \mu_2]$ for any $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. This discrepancy LS_θ is an upper-bound of W_2 and verifies all the axioms of a distance apart from the fact that there exists $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathrm{LS}_\theta(\mu, \mu) > 0$. The discrete case appears to be better suited: if μ has countable discrete support, then $\mathrm{LS}_\theta(\mu, \mu) = 0$ for almost-every $\theta \in \mathbb{S}^{d-1}$. The Expected Sliced discrepancy is then

defined as the cost of the *average* for $\mathfrak{s} \in \mathcal{P}(\mathbb{S}^{d-1})$ (a probability measure on the hypersphere) of the lifted plans $\gamma_\theta[\mu_1, \mu_2]$, which evaluates to:

$$\mathrm{ES}^2_{\mathfrak{s}}(\mu_1, \mu_2) := \int_{\mathbb{S}^{d-1}} \mathrm{LS}^2_\theta(\mu_1, \mu_2) \mathrm{d}\mathfrak{s}(\theta) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^{2d}} \|x - y\|_2^2 \mathrm{d}\gamma_\theta[\mu_1, \mu_2](x, y) \mathrm{d}\mathfrak{s}(\theta).$$

We show that when $\mathfrak{s}$ is absolutely continuous with respect to the uniform measure on $\mathbb{S}^{d-1}$, then $\mathrm{ES}_{\mathfrak{s}}$ is a distance on the set of measures with countable support. However, we also show that for *any* $\mathfrak{s} \in \mathcal{P}(\mathbb{S}^{d-1})$, taking μ as the uniform measure on the unit ball of $\mathbb{R}^d$ yields $\mathrm{ES}_{\mathfrak{s}}(\mu, \mu) > 0$, hence $\mathrm{ES}_{\mathfrak{s}}$ is never a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

0.2.2.3 Chapter B.III: Sliced Gromov-Wasserstein

In Chapter B.III, we begin with reminders on the Frank-Wolfe (FW) algorithm [FW+56] to solve the notoriously difficult Gromov-Wasserstein (GW) problem [Mém11]. This approach was popularised by [Vay+20] for a variant of GW. We then present an alternate minimisation method for GW which solves a lower-bound, and explain how each iteration can be seen as an OT problem. We then investigate the heuristic of replacing the inner OT problems with sliced OT problems using the Pivot Sliced plans introduced in Chapter B.II. Motivated by computational advantages, this heuristic appears to perform unsatisfactorily in practice.

0.2.3 Part C: Optimal Transport Barycentres

As a metric on the space of probability measures, the 2-Wasserstein distance provides a natural way of defining barycentres in the sense of Fréchet, as introduced by [CE10; AC11]: a 2-Wasserstein barycentre of measures $(\nu_1, \cdots, \nu_K) \in \mathcal{P}_2(\mathbb{R}^d)^K$ for weights $(\lambda_1, \cdots, \lambda_K) \in \triangle_K$ is any measure minimising the squared distances to the other measures:

$$\operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\mu, \nu_k).$$

Numerically, computing a barycentre of discrete measures is challenging [Kro+19; AB22], and the most commonly used approach consists in minimising with respect to a support of fixed size, yielding the following Lagrangian functional:

$$X \in \mathbb{R}^{n \times d} \longmapsto \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\gamma_X, \nu_k),$$

where $\gamma_X := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. An iterative approach to this minimisation consisting in using barycentric projections of OT plans was proposed in [CD14], and is currently one of the most popular for problems which require optimisation of the support (thanks to the implementation in `POT` [Fla+21]).

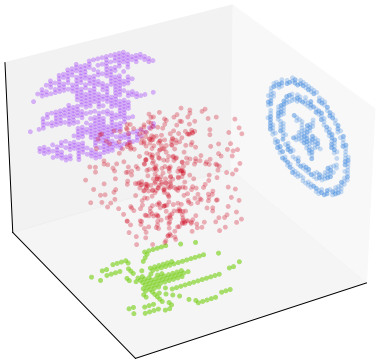
Generalisations of the notion of 2-Wasserstein barycentres are surprisingly scarce in literature: the W_1 case was studied recently in [CCE24], and the case W_p^p even more recently in [BFR25], with an extension to costs $c(x, y) = h(x - y)$ with h smooth and strongly convex in [BFR24]. Another generalisation was proposed in [DGS21], where the barycentre is sought in $\mathcal{P}_2(\mathbb{R}^d)$ between measures $\nu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$, using a cost $c_k(x, y) := \|P_k x - y_k\|_2^2$ with $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$ fixed linear maps. This yields the following variational expression, called the Generalised Wasserstein Barycentre (GWB) problem:

$$\operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k).$$

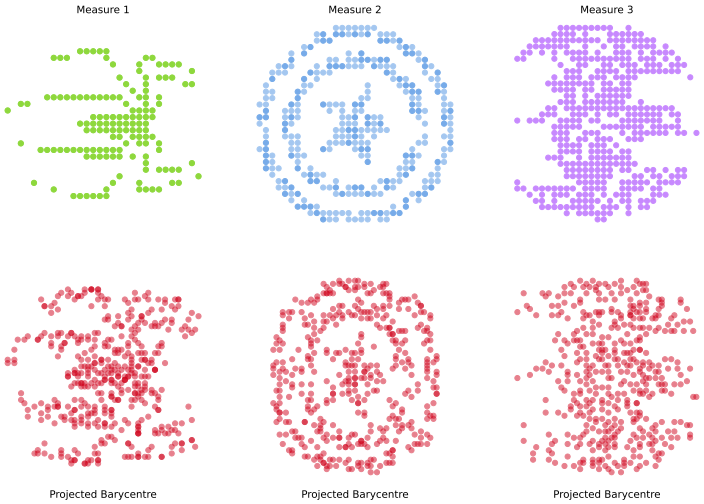
In Chapter C.I, we study numerical methods for solving the GWB problem in the discrete case, and introduce an extension of the problem which also optimises over the linear maps P_k. In Chapter C.II, we introduce a very general notion of barycentre between measures in different metric spaces. We propose a fixed-point algorithm which generalises the method introduced by [Álv+16] for 2-Wasserstein barycentres between absolutely continuous measures, and provide numerical algorithms as well as ties to the popular "free-support" method of [CD14].

0.2.3.1 Chapter C.I: (Blind) Generalised Wasserstein Barycentres

In Chapter C.I, we study the Generalised Wasserstein Barycentre (GWB), providing three additional numerical methods for its numerical resolution: GD, SGD, and Block Coordinate Descent (BCD). Using OT regularity results from Chapter B.I, we were able to apply a theorem from [EN97] to show that GD iterates sub-sequentially converge to Clarke critical points. Likewise, for SGD using again Chapter B.I, we use [Dav+20] to show that SGD trajectories sub-sequentially converge to Clarke critical points almost-surely, conditionally to their boundedness. In practice, the BCD method converges much faster towards a local optimum (which may not be a global optimum), but no guarantees could be established. We illustrate the GWB problem in Fig. 0.18.



(a) 3D representation of the GWB solution.



(b) Projections of the GWB solution.

Figure 0.18: Results of a GWB which determines a barycentre in $\mathbb{R}^3$ whose projections match three given discrete measures in $\mathbb{R}^2$. The target measures are represented in blue, green and purple, while the barycentre is represented in red. On the left, we observe the barycentre in $\mathbb{R}^3$ along with embeddings of the two-dimensional measures. On the right, for each of the three target measures ν_k, we compare the target measure with the projection $P_k\#\mu$ of the barycentre μ in $\mathbb{R}^3$.

We also extend the GWB problem to the case where the linear maps P_k are also optimised. After showing existence of minimisers to this ill-posed problem, we adapt our algorithms to this case, and show that the GD and SGD methods converge in the same sense as for the GWB resolution. Even though the problem is substantially more difficult to solve numerically, a few simple numerical tricks allowed us to obtain good results in practice.

0.2.3.2 Chapter C.II: Computing Optimal Transport Barycentres

Generalising barycentres to any cost. In Chapter C.II, we define a notion of barycentre of measures $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$ where for $k \in [\![1, K]\!]$, the space $(\mathcal{Y}_k, d_{\mathcal{Y}_k})$ is a compact metric space. A barycentre is sought in the space of measures on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ using continuous costs $c_k : \mathcal{X} \times \mathcal{Y}_k \longrightarrow \mathbb{R}_+$, and the barycentre problem is the following generalised Fréchet mean of measures:

$$\operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \sum_{k=1}^{K} \mathcal{T}_{c_k}(\mu, \nu_k) =: V(\mu).$$

Fixed-point algorithm. In order to formulate an iterative method for this problem, we work under the assumption that the generalised Fréchet mean between points $(y_k) \in \Pi_k \mathcal{Y}_k$ is well-defined (we call this the "ground barycentre"), which we formulate as the assumption that the

following problem has a unique solution:

$$B(y_1, \cdots, y_K) := \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k).$$

Our fixed-point algorithm consists in an extensive generalisation of the method proposed by [Álv+16] for the 2-Wasserstein barycentre of absolutely continuous measures $(\nu_k) \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)^K$. At step t, their algorithm starts with the current measure $\mu_t \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and computes the OT maps T_k from μ_t to ν_k. This provides a local linearisation of the 2-Wasserstein space in the sense of [MDC20], and an L^2 barycentre is computed in this linearised space, namely $S := \sum_{k=1}^{K} \lambda_k T_k$. The next step is then $\mu_{t+1} := S\#\mu_t$, which in terms of random variables can be written $X_{t+1} := \sum_k \lambda_k T_k(X_t)$ for $X_t \sim \mu_t$. We summarise this linearisation intuition in Fig. 0.19.

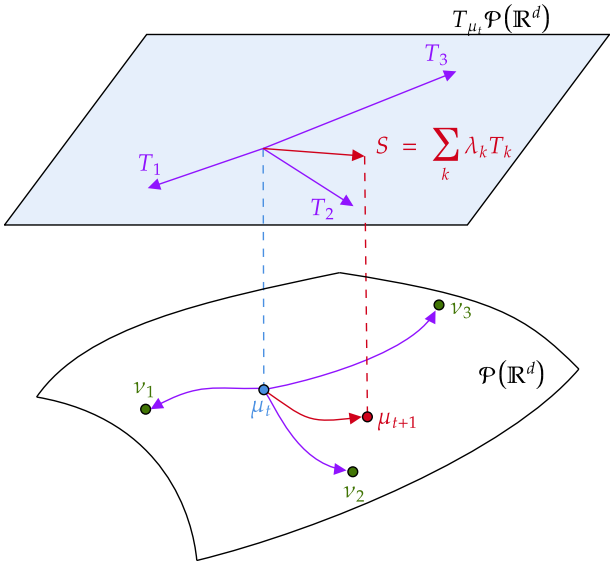


Figure 0.19: Visualisation of the linearisation interpretation of the fixed-point algorithm of [Álv+16] for 2-Wasserstein barycentres of absolutely continuous measures. The space $L^2(\mu_t)$ is informally seen as a the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ_t.

In our setting, we may not have OT maps for the costs c_k, so we combine OT plans instead of OT maps. To this end, we introduce a set-valued iteration functional $G : \mathcal{P}(\mathcal{X}) \rightrightarrows \mathcal{P}(\mathcal{X})$ as follows:

$$\forall \mu \in \mathcal{P}(\mathcal{X}),\ G(\mu) := \left\{ \mathrm{Law}[B(Y_1, \cdots, Y_K)] : \forall k \in [\![1, K]\!],\ \mathrm{Law}[(X, Y_k)] \in \Pi^*_{c_k}(\mu, \nu_k) \right\},$$

where $\Pi^*_{c_k}(\mu, \nu_k)$ is the set of OT plans between μ and ν_k for the cost c_k. Note that if each $\Pi^*_{c_k}(\mu, \nu_k)$ is reduced to a single map-induced plan $(I, T_k)\#\mu$, then $G(\mu)$ is reduced to the single measure $\mathrm{Law}_{X\sim\mu}[B(T_1(X), \cdots, T_K(X))]$ which corresponds to taking the ground barycentre B of the OT maps T_k. Using this set-valued functional, we define our fixed-point algorithm as

$$\mu_0 \in \mathcal{P}(\mathcal{X}),\ \forall t \in \mathbb{N},\ \mu_{t+1} \in G(\mu_t),$$

which is a generalisation of [Álv+16], since we work without regularity assumptions on the measures and with general costs c_k. In particular, our framework allows for barycentres between discrete measures, which is of paramount interest in practice.

Convergence. We show that the convergence results of [Álv+16] hold in our setting, in particular leveraging technical results about the regularity of the set-valued functional G. First, we show that the iterates have non-increasing energy:

Proposition. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\overline{\mu} \in G(\mu)$. There exists a function $\delta : \mathcal{X}^2 \longrightarrow \mathbb{R}_+$ such that $V(\mu) \geq V(\overline{\mu}) + \mathcal{T}_\delta(\mu, \overline{\mu})$. If μ^* is a barycentre, then $G(\mu^*) = \{\mu^*\}$.

The function δ at play is explicit and enjoys several useful properties that we do not detail here, but note that in the W_2^2 case, it is simply $\|x - y\|_2^2$. Thanks to this result and to technical regularity results on G, we obtain the following convergence result, which mirrors that of [Álv+16]:

Theorem. For any $\mu_0 \in \mathcal{P}(\mathcal{X})$, let (μ_t) verifying $\mu_{t+1} \in G(\mu_t)$. Then (μ_t) has converging subsequences, and any weakly converging subsequence necessarily converges towards a $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mu \in G(\mu)$.

Numerics. We provide an implementation of this algorithm, in particular constructing a suitable coupling $(Y_1, \cdots, Y_K)$ using a generalised North-West Corner Rule, that is sufficiently sparse to avoid memory issues. We also investigate a heuristic simplification of G which enforces the support size and the weight of the barycentre to remain constant, which can be seen as a direct generalisation of the "free-support" method of [CD14]. We also show that our theory extends to barycentres for entropically regularised costs, and detail the application of our methods to the computation of barycentres of Gaussian Mixture Models (GMMs) in the sense of [DD20]. In Fig. 0.20, we illustrate a numerical application to a barycentre for the costs $c_k(x, y) = \|P_k(x) - y\|_2^2$ where P_k are (non-linear) projections onto circles, which can be seen as a non-linear generalisation of the GWB problem [DGS21] studied in Chapter C.I.

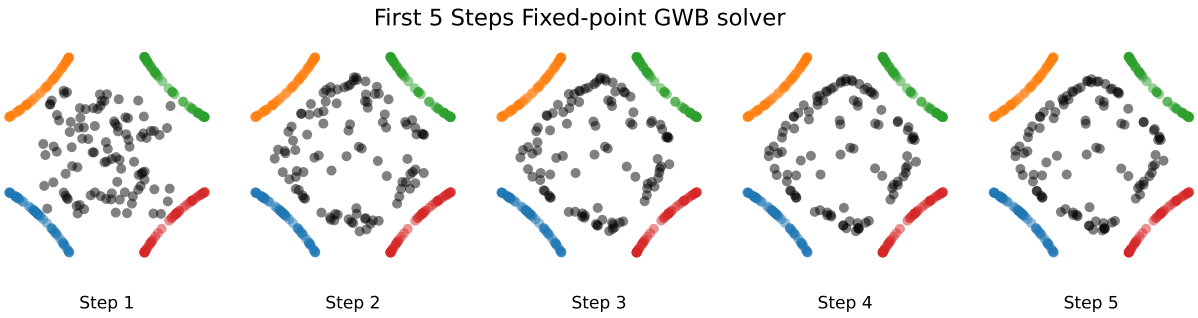


Figure 0.20: First 5 iterations of the fixed-point algorithm for costs $c_k(x, y) = \|P_k(x) - y\|_2^2$, where P_k are projections onto four different circles on which the ν_k are supported (plotted in colour).

0.2.4 Part D: Some Contributions to Kernel Methods

In the last part of this thesis, we present some contributions to a relatively unrelated field of research, that of kernel methods. The genesis of these projects were the attempt at solving the constrained map problem from Chapter B.I using kernel methods. Unfortunately, we show in Chapter D.II that natural ideas to tackle this problem using kernels cannot succeed, since there are no kernels that satisfy the required properties. To introduce the field of Reproducing Kernel Hilbert Spaces (RKHS), we begin Part D with a reminder of the theory of RKHS in Chapter D.I, which is a transcription of introductory talks given by the author at the MAP5 laboratory. Finally, in Chapter D.III, we present an explicit construction of universal kernels on compact metric spaces, which at some point was of interest in the project of Chapter C.II, but ended up not being a requirement, and stemmed into a separate research direction. We also introduce a notion of approximate universality, and show that two discretisations of the proposed universal kernel verify this property, which is useful in practice given their tractability.

0.2.5 Summary of Open-Source Code Contributions

Beyond the theoretical considerations in this thesis, we also provide numerous algorithms with open-source implementations in Python. Most of this code has been contributed to the Python Optimal Transport (`POT`) library [Fla+21] of which the author is a maintainer.

0.2.5.1 Public Reproducible Code

All the code for the algorithms and experiments presented in Chapter C.I is publicly available in the Github repository https://github.com/eloitanguy/bgwb. Likewise, for Chapter C.II, all the implementations can be found in https://github.com/eloitanguy/ot_bar. The latter was awarded the "Silver Reproducible Label" during the submission of a communication on Chapter C.II to the GRETSI 2025 conference. A public implementation for Chapter A.IV is in progress, as is one for Chapter B.II.

0.2.5.2 Contributions to the POT library

Throughout this thesis, the author has contributed to the `POT` library [Fla+21] through several pull requests (PRs) implementing algorithms presented by - or related to - this work.

- Generalised Wasserstein Barycentres ([DGS21], Chapter C.I): PR 372
- Smooth Strongly Convex Nearest Brenier Potentials ([PdC20], Chapter B.I): PR 526
- Gaussian Mixture Model Optimal Transport ([DD20], Chapters C.II and A.IV): PR 649
- Fixed-point solvers for OT barycentres (Chapter C.II): PR 715

0.3 List of Preprints and Publications

[TFD24b] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Reconstructing discrete measures from projections.
Consequences on the empirical Sliced Wasserstein Distance".
Comptes Rendus. Mathématique 362 (Jun. 2024), pp. 1121-1129.

[TFD24a] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Properties of Discrete Sliced Wasserstein Losses".
Mathematics of Computation (Jun. 2024).

[Tan23] Eloi Tanguy.
"Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses".
Transactions on Machine Learning Research (Oct. 2023).

[Boï+25] Samuel Boïté*, Eloi Tanguy*, Julie Delon, Agnès Desolneux and Rémi Flamary.
"Differentiable Expectation-Maximisation
and Applications to Gaussian Mixture Model Optimal Transport".
arxiv preprint 2509.02109 (Sept. 2025). (*: equal contribution)

[TDD25] Eloi Tanguy, Agnès Desolneux, and Julie Delon.
"Constrained Approximate Optimal Transport Maps".
ESAIM: Control, Optimisation and Calculus of Variations. (Aug. 2025)

[TCD25] Eloi Tanguy, Laetitia Chapel and Julie Delon.
"Sliced Optimal Transport Plans".
arxiv preprint 2508.01243 (Aug. 2025).

[TDG24] Eloi Tanguy, Julie Delon and Nathaël Gozlan.
"Computing Barycentres of Measures for Generic Transport Costs".
arxiv preprint 2501.04016 (Dec. 2024).

[Tan25] Eloi Tanguy.
"Explicit Universal and Approximate-Universal Kernels
on Compact Metric Spaces".
arxiv preprint 2506.03661 (Jun. 2025).

0.4 Résumé de la Thèse et des Contributions

Cette thèse est principalement consacrée à l'étude théorique et à la résolution pratique de problèmes d'optimisation issus du contexte du Transport Optimal (TO) et de ses applications. La Partie A étudie la minimisation de discrépances de TO par rapport à des mesures discrètes, avec des conséquences théoriques et des applications pratiques en modélisation générative. La Partie B introduit plusieurs variantes de problèmes de TO qui contraignent les applications ou plans de transport candidats. Enfin, la Partie C définit des généralisations des barycentres de Wasserstein et introduit des algorithmes pour leur calcul. La dernière partie de la thèse (Partie D) est consacrée aux méthodes à noyaux. Dans cette partie au thème très différent, nous étudions des représentations des gradients de fonctions convexes, et présentons de nouveaux noyaux universels explicites sur des espaces métriques compacts. Pour donner un aperçu général du contenu de cette thèse, nous attirons l'attention à la Fig. 0.8 de la partie anglophone, qui respecte le code couleur de chaque partie et chapitre.

0.4.1 Partie A: Discrépances de Transport Optimal comme Fonctions de Perte

Les tâches de *Machine Learning* (ML) sont généralement formulées comme des problèmes de minimisation sur une certaine fonction de perte. Les Chapitres A.I to A.III étudient la minimisation de la distance 2-*Sliced Wasserstein* entre mesures discrètes selon différents points de vue, dans le but de son étude comme une perte en ML. Le fil conducteur de ces trois chapitres est les deux fonctions d'énergie suivantes :

$$\mathcal{E} := Y \in \mathbb{R}^{n\times d} \longmapsto \int_{\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^{n}(P_\theta y_{\sigma_\theta(i)} - P_\theta z_{\tau_\theta(i)})^2 \mathrm{d}\sigma(\theta) = \mathrm{SW}_2^2(\gamma_Y,\gamma_Z);$$

$$\mathcal{E}_p := Y \in \mathbb{R}^{n\times d} \longmapsto \frac{1}{p}\sum_{k=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(P_{\theta_k} y_{\sigma_{\theta_k}(i)} - P_{\theta_k} z_{\tau_{\theta_k}(i)})^2 = \widehat{\mathrm{SW}}_p(\gamma_Y,\gamma_Z),$$

où $\gamma_Y := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ et $\gamma_Z := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ sont les mesures empiriques associées aux données $Y = (y_1,\cdots,y_n) \in \mathbb{R}^{n\times d}$ et $Z = (z_1,\cdots,z_n) \in \mathbb{R}^{n\times d}$. Les deux énergies $\mathcal{E}$ et $\mathcal{E}_p$ correspondent à la distance 2-*Sliced Wasserstein* discrète $\mathrm{SW}_2(\gamma_Y,\gamma_Z)$ et à son approximation empirique de Monte-Carlo en fonction du support Y de l'une des mesures.

0.4.1.1 Chapitre A.I: Reconstruction de Mesures Discrètes à partir de Projections

Le Chapitre A.I étudie le problème inverse déterminant toutes les mesures de probabilité $\gamma \in \mathcal{P}(\mathbb{R}^d)$ ayant les mêmes images par des applications linéaires $P_i : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_i}$ qu'une mesure cible fixée $\gamma_Z := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. Plus formellement, nous nous intéressons à l'ensemble de solutions suivant :

$$\mathcal{S} = \left\{\gamma \in \mathcal{P}(\mathbb{R}^d) \mid \forall i \in [\![1,r]\!],\ P_i\#\gamma = P_i\#\gamma_Z\right\}.$$

Le résultat central énonce que lorsque $\sum_i d_i > d$ et que les P_i sont choisis aléatoirement (chaque ligne étant choisie indépendamment selon une loi ayant une densité sur $\mathbb{R}^d$), alors presque sûrement la mesure originale est la solution unique : $\mathcal{S} = \{\gamma_Z\}$. Nous présentons aussi des cas pathologiques pour certaines applications P_i et montrons que lorsque $\sum_i d_i \leq d$, il existe toujours des solutions indésirables. En considérant en particulier p projections $P_i := \theta_i^\top$ pour $\theta_i \in \mathbb{S}^{d-1}$, le problème de reconstruction coïncide exactement avec le problème de recherche des minima globaux de $\mathcal{E}_p$, ce qui conduit au résultat suivant :

Théorème. Pour $Z \in \mathbb{R}^{n\times d}$ avec les (z_i) distincts et $(\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p}$, alors presque sûrement :

- Si $p \leq d$, il existe une infinité de $Y \in \mathbb{R}^{n\times d}$ tels que $\gamma_Y \neq \gamma_Z$ et $\mathcal{E}_p(Y) = 0$.
- Si $p > d$, alors $\operatorname{argmin} \mathcal{E}_p = \{Y \in \mathbb{R}^{n\times d} : \gamma_Y = \gamma_Z\}$.

Cette propriété intuitive est d'une grande utilité pour l'étude de la distance 2-*Sliced Wasserstein* en discret ; en particulier, elle décourage fortement l'utilisation de $\mathcal{E}_p$ pour un petit nombre de projections p. De plus, elle détermine l'ensemble des optima globaux de $\mathcal{E}_p$ lorsque $p > d$, qui correspondent à la mesure cible.

0.4.1.2 Chapitre A.II: Propriétés des Pertes Sliced Wasserstein en Discret

Convergence de $\mathcal{E}_p$ vers $\mathcal{E}$. Le Chapitre A.II se concentre sur les énergies $\mathcal{E}$ et $\mathcal{E}_p$, et commence par des propriétés statistiques concernant la convergence lorsque $p \longrightarrow +\infty$. Tout d'abord, la convergence uniforme de $\mathcal{E}_p$ vers $\mathcal{E}$ est établie, étendant la convergence ponctuelle de Monte-Carlo à la convergence uniforme sur tout compact, et montrant une convergence uniforme au sens du théorème centrale-limite :

Théorème. Pour tout ensemble compact $\mathcal{K} \subset \mathbb{R}^{n\times d}$, nous avons :

- presque sûrement, $\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{} 0$;
- $\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} G$, où G est un processus gaussien centré sur $\mathcal{K}$,

où $\|\cdot\|_{\ell^\infty(\mathcal{K})}$ est la norme du supremum sur $\mathcal{K}$, et "$\mathcal{L}$" désigne la convergence en loi.

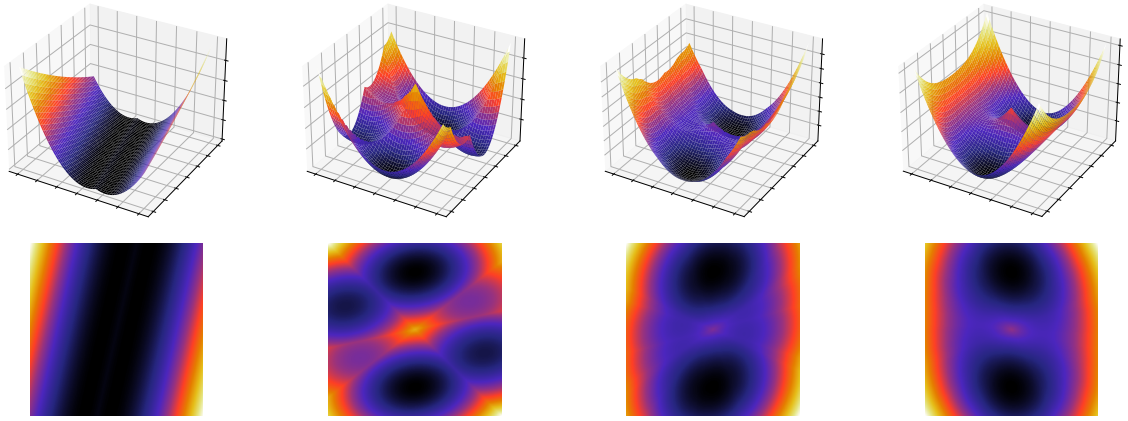
Questions de régularité. Pour étudier les propriétés d'optimisation de $\mathcal{E}$ et $\mathcal{E}_p$, nous établissons d'abord des résultats de régularité sur ces énergies. À cette fin, nous montrons un résultat utile sur le caractère Lipschitz de la valeur du problème de Kantorovich $\mathrm{W}(\alpha, \beta; C)$ par rapport aux poids α, β et à la matrice de coût C :

Lemme. Étant donné des poids $\alpha, \overline{\alpha} \in \triangle_n$ et $\beta, \overline{\beta} \in \triangle_m$ (à entrées > 0) et des matrices de coût $C, \overline{C} \in \mathbb{R}^{n\times m}$, nous avons :

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C - \overline{C}\|_\infty + \|C\|_\infty(\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1).$$

Ce résultat fournit non seulement le caractère localement Lipschitz de $\mathcal{E}$ et $\mathcal{E}_p$, mais sera également utilisé à plusieurs reprises dans cette thèse pour établir la régularité d'énergies issues du TO. Nous montrons en outre que $\mathcal{E}$ et $\mathcal{E}_p$ sont semi-concaves (ce qui est défavorable pour la minimisation), et détaillons la structure de $\mathcal{E}_p$, en l'écrivant comme un minimum de fonctions quadratiques. Dans le Fig. 0.21, nous illustrons les paysages de $\mathcal{E}_p$ et $\mathcal{E}$ dans un cas simple, pour montrer la convergence lorsque p augmente, et pour illustrer l'expression de $\mathcal{E}_p$ comme minimum de quadratiques, ce que nous appelons une "structure en cellules". Cette propriété implique en particulier que $\mathcal{E}_p$ est semi-algébrique, ce qui sera utile à plusieurs reprises pour l'étude de la Descente de Gradient Stochastique (SGD en Anglais). Quant à $\mathcal{E}$, sa différentiabilité sur un certain ouvert est connue depuis [Bon+15a], avec une expression explicite du gradient.

Minimisation des énergies. En ce qui concerne les propriétés d'optimisation, grâce à au Chapitre A.I, les optima globaux de $\mathcal{E}_p$ sont presque sûrement des permutations de la cible Z si p est suffisamment grand. Puisque SW_2 est une distance, les seuls optima globaux de $\mathcal{E}$ sont également les réarrangements de Z. Quant aux optima locaux, nous exploitons la "structure en cellules" de $\mathcal{E}_p$ pour montrer que $\mathcal{E}_p$ est différentiable uniquement en dehors des frontières des



(a) $\mathcal{E}_p,\ p=1$ (b) $\mathcal{E}_p,\ p=3$ (c) $\mathcal{E}_p,\ p=10$ (d) $\mathcal{E}$

Figure 0.21: Paysages $\mathcal{E}_p$ et $\mathcal{E}$ dans un cas simple.

cellules, et que ses points Y de différentiabilité vérifiant $\nabla\mathcal{E}(Y)=0$ (appelés "points critiques") ce qui correspond aux minima globaux des cellules. De plus, nous explicitons une fonction Ψ telle que, dans l'ensemble de différentiabilité de $\mathcal{E}$, nous avons $\nabla\mathcal{E}(Y)=0 \iff \Psi(Y)=Y$. Étant donné $(Y_p)_{p>d}$ points critiques de $\mathcal{E}_p$ pour $p>d$, ils approchent les points critiques de $\mathcal{E}$ au sens faible suivant :

$$Y_p - \Psi(Y_p) \xrightarrow[p\longrightarrow+\infty]{\mathbb{P}} 0,$$

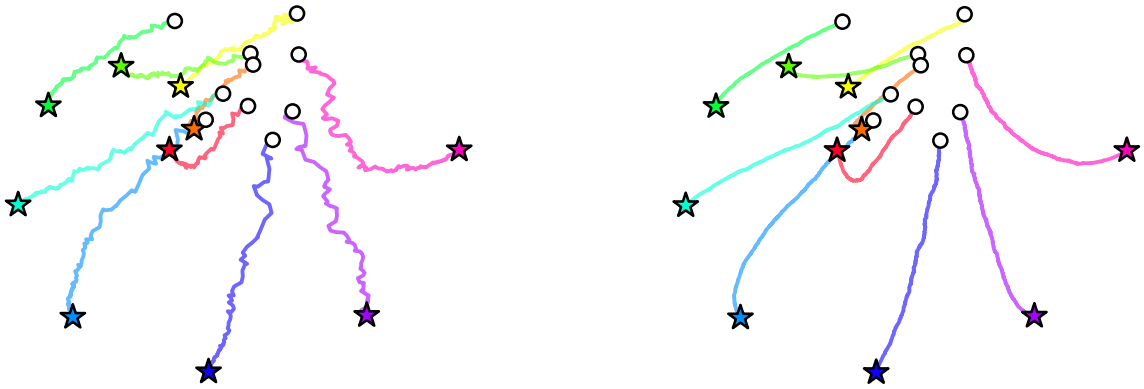
où le taux de décroissance est contrôlé par une majoration explicite.

Convergence de la SGD. Grâce à nos résultats de régularité, nous avons pu étudier la convergence de la SGD sur l'énergie $\mathcal{E}$, qui consiste à prendre un échantillon aléatoire de p projections à chaque itération et à calculer un (sous-)gradient stochastique de $\mathcal{E}_p$. Cette étude constitue une étape vers la compréhension de l'entraînement des réseaux de neurones génératifs avec la perte 2-*Sliced Wasserstein*, qui est étudiée plus en détail dans le Chapitre A.III. De plus, elle fournit un cadre théorique pour l'algorithme de correspondance *sliced* proposé dans [Rab+12]. Cette méthode consiste à exécuter une SGD sur $\mathcal{E}$, et à la convergence d'assigner à chaque point x_i de l'initialisation le point z_j le plus proche vers lequel il a convergé. En utilisant nos résultats, nous obtenons dans un sens faible certaines garanties de convergence de cette procédure. Nous résumons ci-dessous une formulation informelle et simplifiée de nos résultats de convergence, qui s'appuient sur des travaux théoriques concernant la convergence de la SGD [BHS22; Dav+20].

Théorème.

- Les itérations SGD $(Y_\alpha^{(t)})_{t\in\mathbb{N}}$ avec un pas fixé α sont telles que leurs interpolations affines par morceaux $(Y_\alpha(\cdot))$ approchent l'ensemble des solutions de l'inclusion différentielle $\dot{Y}(s)\in -\partial_C\mathcal{E}(Y(s))$ lorsque le pas tend vers 0.
- Les itérations SGD bruitées $(Y_\alpha^{(t)})_{t\in\mathbb{N}}$ avec un pas fixé α sont telles que les limites sous-séquentielles en temps long approchent l'ensemble des points critiques de Clarke de $\mathcal{E}$ lorsque le pas tend vers 0.
- Conditionnellement à leur caractère borné, les trajectoires de la SGD à pas décroissants convergent sous-séquentiellement vers des points critiques de Clarke de $\mathcal{E}$.

Le premier résultat signifie que pour des petits pas (constants), les trajectoires sont similaires aux solutions de l'inclusion différentielle, ce qui, dans le cas différentiable, serait équivalent à un flot de gradient. Les deux autres résultats justifient la convergence lorsque $t\longrightarrow+\infty$ vers un ensemble de points critiques généralisés de $\mathcal{E}$, qui en pratique s'avèrent être des optima globaux, comme c'est le cas dans l'illustration proposée dans la Fig. 0.22.



(a) SGD avec pas $\alpha = 0.3$.

(b) SGD avec pas $\alpha = 0.01$.

Figure 0.22: SGD sur l'énergie $\mathcal{E}$ où la mesure cible est une spirale 2d (étoiles) avec une initialisation aléatoire (cercles).

0.4.1.3 Chapitre A.III: Convergence de la SGD pour l'Entraînement de Réseaux de Neurones avec des Pertes Sliced Wasserstein

Le Chapitre A.III est une extension des études de la SGD du Chapitre A.II à l'entraînement de réseaux de neurones génératifs avec la perte 2-*Sliced Wasserstein*. Au lieu de minimiser $\mathcal{E}$ par rapport aux données, nous souhaitons maintenant minimiser l'énergie : $F^* := u \longmapsto \mathrm{SW}_2^2(T_u\#\mu, \nu)$, où μ est une distribution de données source fixée (souvent choisie comme une distribution gaussienne ou uniforme) et ν est une distribution de données cible (typiquement un jeu de données discret), avec T_u un réseau de neurones (*Neural Network* NN) paramétré par $u \in \mathbb{R}^{d_u}$. En pratique (comme dans [DZS18]), cette minimisation est effectuée par SGD, en échantillonnant à chaque itération un lot de n points de données d'entrée $X \sim \mu^{\otimes n}$ et de n points de données cibles $Y \sim \nu^{\otimes n}$, ainsi qu'une direction aléatoire $\theta \sim \varpi$ (ou un lot de celles-ci), puis en calculant un pas de gradient de la fonction de perte entre échantillons :

$$f(u, X, Y, \theta) := \mathrm{W}_2^2(P_\theta\#T_u\#\gamma_X, P_\theta\#\gamma_Y)$$

par rapport à u. Nous rappelons que $\gamma_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Cela correspond à la minimisation d'un substitut *mini-batch* de F^*, puisque la perte globale associée est :

$$F(u) := \int \mathrm{W}_2^2(P_\theta\#T_u\#\gamma_X, P_\theta\#\gamma_Y)\mathrm{d}\varpi(\theta)\mathrm{d}\mu^{\otimes n}(X)\mathrm{d}\nu^{\otimes n}(Y).$$

Sous diverses hypothèses techniques et en s'appuyant largement sur les résultats et techniques du Chapitre A.II, nous montrons dans le Chapitre A.III un analogue des résultats de convergence du Chapitre A.II pour la SGD de F. En un certain sens, ces conclusions fournissent des garanties de convergence pour l'entraînement de modèles génératifs avec la perte 2-*Sliced Wasserstein*.

0.4.1.4 Chapitre A.IV: Espérance-Maximisation Différentiable et Applications au Transport Optimal entre Mixtures Gaussiennes

Le Chapitre A.IV se concentre sur une autre variante du TO appelée la distance *Mixture-Wasserstein distance* [DD20], qui compare deux mélanges de gaussiennes (*Gaussian Mixture Models* GMMs) $\mu := \sum_{k=1}^K w_k \mathcal{N}(m_k, \Sigma_k)$ et $\nu := \sum_{\ell=1}^L \overline{w}_\ell \mathcal{N}(\overline{m}_\ell, \overline{\Sigma}_\ell)$. Cette comparaison se fait en résolvant un problème de Kantorovich entre $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ pour le coût euclidien quadratique, en restreignant les couplages π à être eux-mêmes des GMMs. Une formulation équivalente est un problème de Kantorovich discret entre les composantes gaussiennes de μ et ν, où le coût entre les composantes est la distance de 2-Wasserstein gaussienne :

$$\mathrm{MW}_2^2(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu) \cap \mathrm{GMM}_{2d}} \int_{\mathbb{R}^{2d}} \|x-y\|_2^2 \mathrm{d}\pi(x, y) = \min_{\pi \in \Pi(w,\overline{w})} \sum_{k,\ell} \pi_{k,\ell} \mathrm{W}_2^2(\mathcal{N}(m_k, \Sigma_k), \mathcal{N}(\overline{m}_\ell, \overline{\Sigma}_\ell)),$$

où GMM_d est l'ensemble des GMMs sur $\mathbb{R}^d$ avec un nombre (fini) quelconque de composantes. En pratique, on travaille souvent avec des nuages de points de la forme $X \in \mathbb{R}^{n\times d}$, et pour utiliser MW_2^2 afin de comparer de telles données, il faut d'abord déterminer des GMMs les approchant. Comme suggéré par [DD20], la méthode la plus naturelle est l'algorithme Espérance-Maximisation (EM). Dans les applications pratiques de *Machine Learning*, l'entraînement des modèles repose sur la différentiation automatique, ce qui motive l'étude de la différentiabilité de l'algorithme EM.

Méthodes de différentiation d'EM. Dans le Chapitre A.IV, nous introduisons trois méthodes pour calculer les gradients automatiques de la sortie de l'algorithme EM par rapport aux données d'entrée $X \in \mathbb{R}^{n\times d}$. Nous considérons les étapes d'EM comme un procédé d'itération $\theta_{t+1} = F(\theta_t, X)$, où $\theta_t := (w^{(t)}, (m_k^{(t)})_k, (\Sigma_k^{(t)})_k)$ représente les paramètres du GMM à K composantes à l'étape t. La première méthode est la *Différentiation Automatique Complète (AD)*, qui consiste simplement à effectuer la différentiation automatique à travers toutes les étapes d'EM. Bien que cette méthode puisse être coûteuse lorsque le nombre d'itérations T est grand, elle est théoriquement exacte (en dehors de l'accumulation d'erreurs numériques). Nous proposons également deux méthodes approchées qui reposent sur l'hypothèse que l'algorithme EM converge vers un point fixe $\theta^*(\theta_0, X)$ (qui dépend de l'initialisation θ_0 et des données X). En utilisant le théorème des fonctions implicites, nous montrons que θ^* est une fonction bien définie et différentiable de X (sous des hypothèses techniques). Différencier la propriété $\theta^*(X) = F(\theta^*, X)$ par rapport à X donne les expressions suivantes :

$$\frac{\partial \theta^*}{\partial X} = \frac{\partial F}{\partial \theta}(\theta^*, X)\frac{\partial \theta^*}{\partial X} + \frac{\partial F}{\partial X}(\theta^*, X) \quad \Longleftrightarrow \quad \frac{\partial \theta^*}{\partial X} = \left(I - \frac{\partial F}{\partial \theta}(\theta^*, X)\right)^{-1} \frac{\partial F}{\partial X}(\theta^*, X). \tag{0.13}$$

La première méthode approchée, appelée *gradient Implicite Approché (AI)*, consiste à remplacer l'inconnu θ^* dans l' Eq. (0.13) par la dernière itération θ_T. Cette méthode est coûteuse en calcul à cause de l'inversion de matrice, mais fournit en pratique une bonne approximation du gradient. La seconde méthode approchée, appelée *gradient One-Step (OS)*, consiste à également négliger le terme $\partial_\theta F(\theta^*, X)$ dans l' Eq. (0.13), ce qui donne une expression beaucoup plus simple, mais l'hypothèse sous-jacente n'est pas satisfaite par EM en pratique. Nous résumons les trois méthodes dans le Table 2.

Méthode	Expression
Différentiation Automatique Complète (AD) (AD)	calculer $\partial_X[F_X^T(\theta_0)]$ avec `auto-diff`
gradient Implicite Approché (AI)	$J_{\mathrm{AI}} := (I - \partial_\theta F(\theta_T, X))^{-1}\, \partial_X F(\theta_T, X)$
gradient *One-Step* (OS)	$J_{\mathrm{OS}} := \partial_X F(\theta_T, X)$

Table 2: Résumé des méthodes d'approximation du gradient d'EM.

Importance de fixer des poids uniformes. Pour commencer, afin de justifier l'utilisation pratique d'EM avec MW_2^2, nous avons utilisé le lemme de stabilité du transport optimal discret du Chapitre A.II pour borner l'erreur sur l'estimation de MW_2^2 lorsqu'on emploie des paramètres de GMM estimés par EM. Dans les applications pratiques, nous avons observé qu'une variante d'EM qui n'actualise pas les poids des GMM offrait des performances nettement meilleures pour l'optimisation, et qu'il est préférable de se placer dans un contexte de poids uniformes. En examinant en détail la théorie de certains cas particuliers, nous montrons que cette variante *EM à poids fixés* est mieux adaptée aux problèmes qui consistent à optimiser l'entrée de l'algorithme EM afin de minimiser MW_2^2. Pour illustrer ce phénomène, nous considérons un flot de particules qui se déplacent de sorte à approcher une GMM cible, c'est-à-dire en minimisant l'énergie $\mathcal{E}(X) := \mathrm{MW}_2^2(F_X^T(\theta_0, X), \nu)$, où ν est une GMM cible et F_X^T désigne l'algorithme EM appliqué à X pendant T itérations avec l'initialisation θ_0. Dans la Fig. 0.23, nous observons que le flot converge vers un minimum local insatisfaisant lorsque les GMM source et cible ont des poids non-uniformes (et différents), tandis que le cas avec des poids uniformes conduit à une

convergence vers la GMM cible. Sur la base de ces observations, nous recommandons l'utilisation d'EM avec TO-GMM pour des GMM à poids uniformes, en recourant à une variante de l'EM qui n'actualise pas les poids.

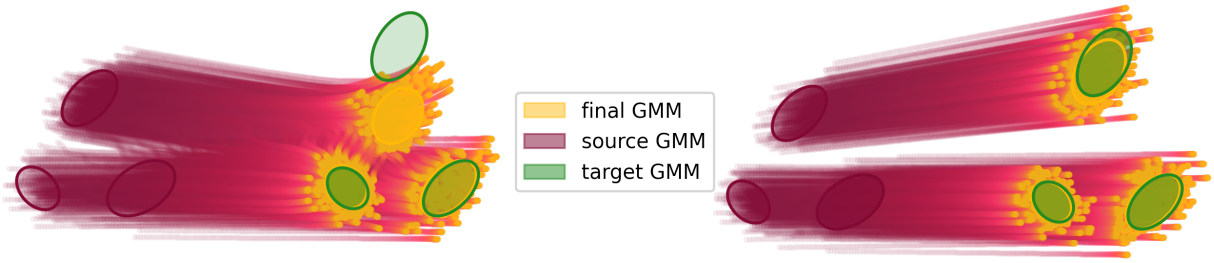
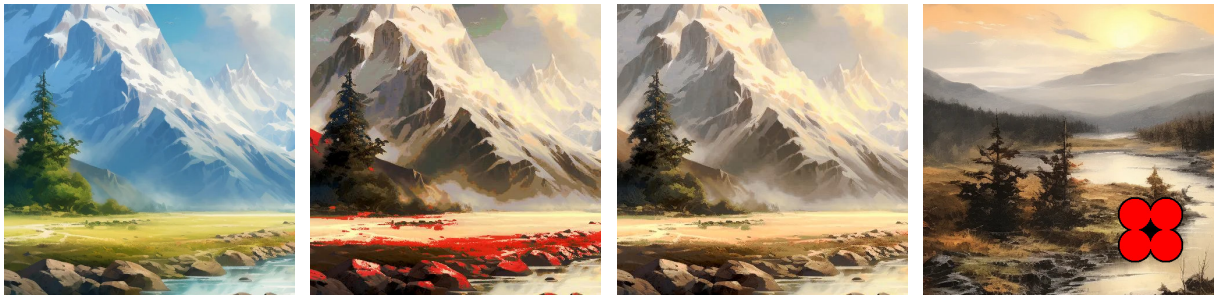


Figure 0.23: Illustration du flot de particules minimisant $\mathcal{E}$ en utilisant AD pour deux GMM différentes, avec des poids non-uniformes (à gauche) et uniformes (à droite).

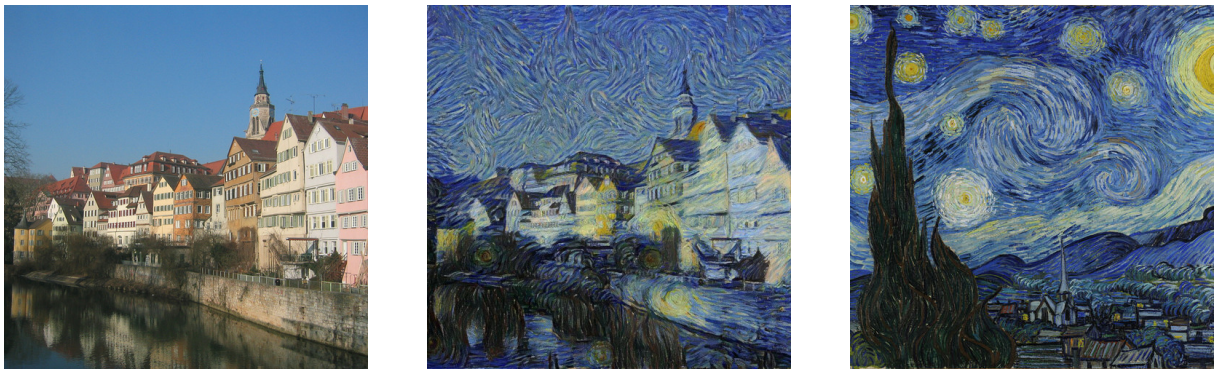
TO-GMM déséquilibré. Comme extension de TO-GMM et comme approche potentielle pour contourner les problèmes liés à l'optimisation des poids, nous introduisons une nouvelle variante de MW_2^2 qui pénalise les contraintes marginales au lieu de les imposer. Cette discrépance *Mixture Wasserstein* déséquilibrée s'inspire directement de le TO déséquilibré [LMS18] et permet de comparer des GMMs de masses différentes, tout en offrant une robustesse accrue face aux composantes aberrantes des GMMs. Pour illustrer cette idée, dans la Fig. 0.24, nous présentons un exemple jouet de transfert de couleur consistant à déplacer les distributions de couleur d'une image source en utilisant TO-GMM et son extension déséquilibrée. L'aberration présente dans l'image cible est correctement ignorée par le TO-GMM déséquilibré, tandis que le TO-GMM classique est contraint de la prendre en compte, ce qui induit une fuite de rouge dans le résultat.



(a) Image source (b) Résultat (équilibré) (c) R. (déséquilibré) (d) Cible corrompue

Figure 0.24: Transfert de couleur équilibré et déséquilibré.

Applications à plus grande échelle. Le chapitre contient de nombreuses illustrations, tels que le calcul de barycentres, d'autres flots associés à MW_2^2 avec différentes méthodes de gradient, ainsi qu'une étude quantitative de la convergence de l'algorithme EM et des méthodes d'approximation du gradient. Nous explorons également des applications à plus grande échelle, en introduisant un *GAN* régularisé par EM-MW_2^2 [Goo+14; ACB17]. De plus, nous présentons une application de transfert de style dans l'esprit de [GEB15], qui transforme une image source de sorte à avoir le style d'une image cible en faisant correspondre ses caractéristiques extraites dans un réseau VGG-19 pré-entraîné [SZ15] avec les caractéristiques de l'image cible, au moyen de EM-MW_2^2. Dans la Fig. 0.25, nous illustrons cette méthode en transférant le style du tableau *La Nuit Étoilée* de Vincent van Gogh vers une photographie de la ville de Tübingen, en Allemagne. Enfin, inspirés par [GLR18; LDD23], nous introduisons une nouvelle méthode de synthèse de textures basée sur EM-MW_2^2 dans des espaces de patchs à plusieurs échelles, voir la Fig. 0.26 pour quelques résultats. Étant donné la simplicité de la méthode et le fait qu'elle ne nécessite ni apprentissage ni réseaux de neurones, les résultats sont surprenamment bons.



(a) Image source (b) Image résultat (c) Style cible

Figure 0.25: Exemple de transfert de style avec EM-MW$_2^2$.

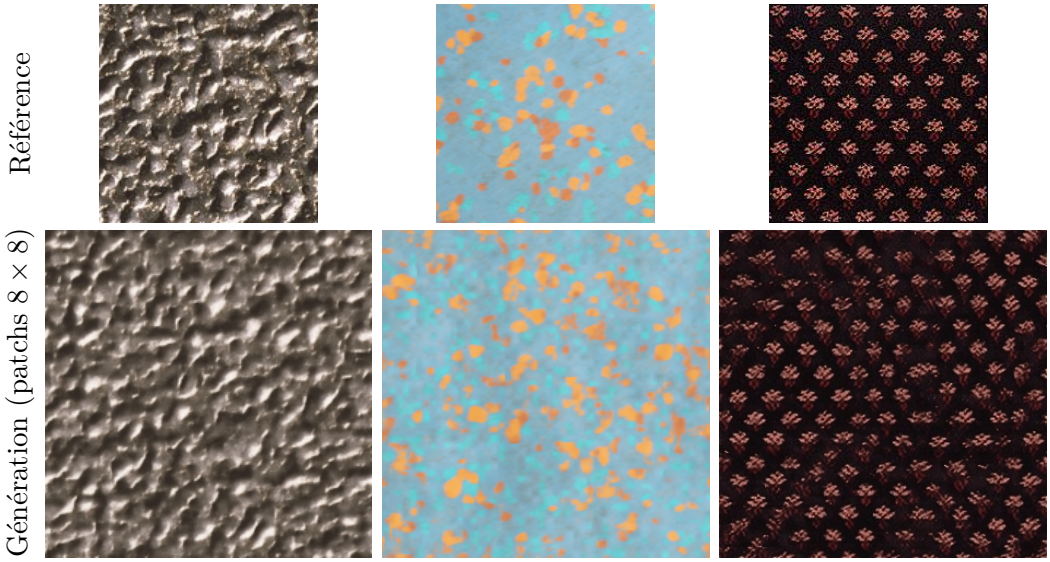


Figure 0.26: Synthèse de textures multi-échelles avec $K = 4$ composantes.

0.4.2 Partie B: Variantes de Plans et d'Applications de Transport Optimal

0.4.2.1 Chapitre B.I: Applications de Transport Approchées et Contraintes

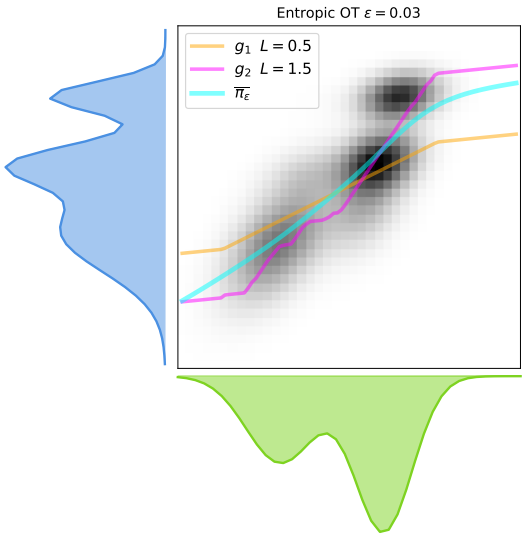
De nombreuses tâches de modélisation générative et de recalage peuvent être formulées comme la recherche d'une fonction f qui transporte une mesure source μ vers une mesure cible ν aussi fidèlement que possible, dans un certain sens. Comme dans [BBR06; ACB17; DZS18], ce critère peut être choisi comme la minimisation d'une discrépance de TO. Dans le Chapitre B.I, nous étudions le problème consistant à minimiser $\mathcal{T}_c(g\#\mu, \nu)$ dans une classe de fonctions G de candidats g, qui est un problème théorique naturel lorsqu'on étudie les méthodes de TO pour la modélisation générative. Ce problème correspond à une version relâchée du problème de Monge, qui pénalise la contrainte $g\#\mu = \nu$ au lieu de l'imposer. En pratique, il n'est pas souhaitable d'avoir exactement $g\#\mu = \nu$ (cela correspond à échantillonner des données connues depuis ν et non à générer de nouveaux exemples), et dans de nombreux cas, on souhaite que g soit régulière afin de garantir robustesse et interprétabilité. Ces contraintes interviennent dans la classe de fonctions G, qui est typiquement une classe de réseaux de neurones. Le cas particulier où G est l'ensemble des gradients L-Lipschitz de fonctions ℓ-fortement convexes et $c(x, y) = \|x - y\|_2^2$ a été étudié dans [PdC20], où les potentiels optimaux étaient appelés *Smooth Strongly Convex Nearest Brenier Potentials*: en un sens, ce problème peut être vu comme une version relâchée du problème de Monge, où la contrainte $g\#\mu = \nu$ est pénalisée, mais la propriété d'être le gradient d'une fonction convexe est imposée. Un atout de cette approche est que l'application apprise peut être évaluée en tout point de l'espace source, même en dehors du support de μ, ce qui rend la stratégie adaptée au *transfer learning* et à la généralisation hors distribution.

Existence. La première question naturelle que nous abordons dans le Chapitre B.I est celle de l'existence de minimiseurs, ce qui s'est avéré surprenamment difficile. Nous montrons que l'existence est garantie sous plusieurs hypothèses techniques, que nous formulons de manière approximative ci-dessous :

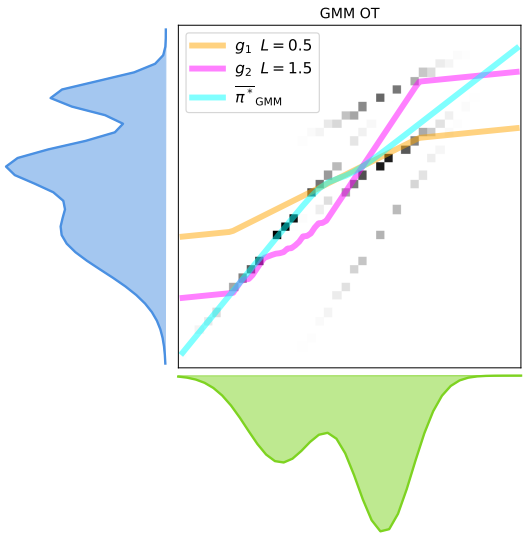
> **Théorème.** Soit $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ semi-continue inférieurement et $(\mathcal{X}, d_{\mathcal{X}})$ un espace polonais localement compact. Soient $\mu \in \mathcal{P}(\mathcal{X})$ et $\nu \in \mathcal{P}(\mathbb{R}^d)$ tels qu'il existe $g \in G$ avec $\mathcal{T}_c(g\#\mu, \nu) < +\infty$. Supposons en outre que c est coercive et que G est un sous-ensemble des fonctions L-Lipschitz de $\mathcal{X}$ vers $\mathbb{R}^d$ pour un certain $L > 0$, et que G est stable par limites uniformes sur les compacts. Alors il existe un minimiseur $g^* \in G$ du problème $\operatorname{argmin}_{g \in G} \mathcal{T}_c(g\#\mu, \nu)$.

En particulier, nous montrons que si G est une classe de réseaux de neurones avec des poids confinés à un ensemble compact et des fonctions d'activation Lipschitziennes, alors l'existence est garantie. De même, nous montrons que la classe des gradients L-Lipschitz de fonctions ℓ-convexes satisfait ces hypothèses. Une extension de ce résultat à des classes de fonctions qui vérifient ces conditions uniquement sur chaque partie d'une partition fixe de $\mathcal{X}$ est également fournie, ce qui permet à nos résultats de généraliser pleinement ceux de [PdC20].

Un problème d'approximation de plan. Une autre motivation de cette étude est le problème consistant à trouver une application qui approche au mieux un plan de transport π entre $\mu \in \mathcal{P}(\mathbb{R}^k)$ et $\nu \in \mathcal{P}(\mathbb{R}^d)$. Par exemple, étant donné uniquement un plan GMM (*Gaussian Mixture Model*) π entre des GMMs μ et ν [DD20], il est pratique pour les applications d'approximer ce plan par une application g. Une manière naturelle de le faire est de résoudre le problème $\min_{g \in G} \mathcal{T}_C((I, g)\#\mu, \pi)$ avec C un coût sur $(\mathbb{R}^k \times \mathbb{R}^d)^2$, qui est un cas particulier du problème d'application contrainte étudié dans le Chapitre B.I. Nous illustrons cette idée pour approximer un plan régularisé par entropie [Cut13] et un plan GMM [DD20] dans la Fig. 0.27.



(a) Approximation d'un plan entropique [Cut13].



(b) Approximation d'un plan GMM [DD20].

Figure 0.27: Illustration de solutions du problème d'approximation de plan pour deux plans différents entre mélanges de gaussiennes. Nous comparons les plans avec des solutions $L = 1/2$ et $L = 3/2$-Lipschitz, ainsi qu'à la projection barycentrique des plans donnés.

Peut-être contre-intuitivement, pour de nombreux coûts au sol C couramment utilisés en TO, le coût $\mathcal{T}_C((I, g)\#\mu, \pi)$ ne dépend en fait pas du couplage π, mais uniquement de ses marginales μ et ν. Cette observation décevante nous a conduit à un résultat d'un grand intérêt applicatif, fournissant une condition suffisante générale sur C pour que le coût $\mathcal{T}_C((I, g)\#\mu, \pi)$

soit indépendant de π. En particulier, si C est choisi comme une puissance d'une norme p, comme le coût euclidien au carré usuel, alors le coût $\mathcal{T}_C((I,g)\#\mu,\pi)$ est indépendant de π, rendant ainsi le problème d'approximation de plan peu pertinent en pratique.

Cas L^2. Pour le cas du coût euclidien au carré et $\mathcal{X}=\mathbb{R}^d$, un candidat naturel d'application g est une approximation de la projection barycentrique d'un plan de TO π^* entre μ et ν. Nous considérons la question suivante d'équivalence de problèmes :

$$\operatorname*{argmin}_{g\in G} \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x)-y\|_2^2 \mathrm{d}\pi(x,y) \overset{?}{=} \operatorname*{argmin}_{g\in G} \|g-\overline{\pi^*}\|^2_{L^2(\mu)}, \tag{0.14}$$

où π^* est un plan de TO fixé entre μ et ν, et $\overline{\pi^*}$ est sa projection barycentrique (aussi appelée espérance conditionnelle) : $\overline{\pi^*}(x)=\mathbb{E}_{(X,Y)\sim\pi}[Y|X=x]$. [PdC20] a montré que lorsque $d=1$, que la classe de fonctions est celle des gradients L-Lipschitz de fonctions ℓ-fortement convexes, et que μ est absolument continue ou bien discrète, alors les deux problèmes sont équivalents. Nous étendons ce résultat au cas où G est un ensemble quelconque de fonctions croissantes telles que $g\#\mu\in\mathcal{P}_2(\mathbb{R})$, sans hypothèses sur μ. Nous présentons également un contre-exemple à la question d'équivalence de problèmes en dimension $d=2$ pour le cas favorable où G est l'ensemble des fonctions *monotones* continues (ce qui est une propriété nécessaire des gradients Lipschitz de fonctions convexes).

Questions d'optimisation pratiques. Pour étudier la convergence des méthodes pratiques résolvant le problème d'approximation d'application, nous nous appuyons sur les résultats de régularité du TO discret du Chapitre A.II et fournissons des détails sur la différentielle de Clarke du coût de Kantorovich discret par rapport à la matrice de coût :

Proposition. Considérons des poids $a\in\triangle_n$, $b\in\triangle_m$, et le coût de Kantorovich discret :

$$W(a,b,\cdot):=\left\{\begin{array}{ccc}\mathbb{R}^{n\times m} & \longrightarrow & \mathbb{R}\\ M & \longmapsto & \min\limits_{\pi\in\Pi(a,b)} \pi\cdot M\end{array}\right. .$$

La fonction $W(a,b,\cdot)$ est semi-algébrique, Lipschitzienne, et son sous-gradient de Clarke est semi-algébrique et s'écrit pour $M\in\mathbb{R}^{n\times m}$:

$$\partial_C W(a,b,\cdot)(M)=\operatorname*{argmin}_{\pi\in\Pi(a,b)} \pi\cdot M.$$

Grâce à ce résultat et aux avancées récentes de [BLP23], nous avons pu utiliser des techniques similaires à celles du Chapitre A.III pour étudier le SGD sur la fonction de perte mini-batch suivante :

$$F(\theta):=\int \mathcal{T}_c(\delta_{h(\theta,X^{(n)})},\delta_{Y^{(m)}})\mathrm{d}\mu^{\otimes n}(X^{(n)})\mathrm{d}\nu^{\otimes m}(Y^{(m)}),$$

afin d'entraîner un réseau de neurones $h(\theta,\cdot)$ avec activations semi-algébriques et Lipschitziennes, et avec des paramètres θ dans un ensemble compact $\Theta\subset\mathbb{R}^p$. En omettant les détails techniques, nous montrons que les itérés du SGD à pas décroissant (θ_t) convergent presque sûrement, sous-séquentiellement, vers un point critique de Clarke de F. Nous proposons également une méthode simple pour rechercher des applications g dans un espace de Hilbert à noyau reproduisant, et nos résultats de régularité, associés à la théorie classique de la descente de gradient non convexe (GD) de [EN97], permettent de montrer que les itérés GD à pas décroissant (θ_t) convergent sous-séquentiellement vers un point critique de Clarke de l'énergie associée.

0.4.2.2 Chapitre B.II: Plans de Transport Sliced

L'efficacité computationnelle de la distance *Sliced Wasserstein* se fait au prix de ne pas obtenir un plan ni une application de transport entre les mesures. Certaines propositions pour définir des plans *Sliced* ont été suggérées dans le cas discret par [Rab+12; Mah+23; Liu+24], toutefois

de nombreuses questions théoriques et pratiques restent ouvertes. Dans le Chapitre B.II, nous introduisons différentes formalismes des idées présentées dans [Mah+23; Liu+24], avec une étude théorique approfondie.

Intuition sur les plans de transport *Sliced*. Les deux objets sur lesquels ce chapitre se concentre sont les discrépances (Min-)*Pivot Sliced* et *Expected Sliced*. Pour les introduire, la première étape consiste à voir comment définir un plan de transport entre deux mesures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ à partir de leurs projections $P_\theta \# \mu_i$ sur un axe $\theta \in \mathbb{S}^{d-1}$. Dans le cas discret, il serait naturel d'opérer simplement le même appariement entre les points dans $\mathbb{R}^d$ que celui effectué sur la droite, comme illustré dans la Fig. 0.28.

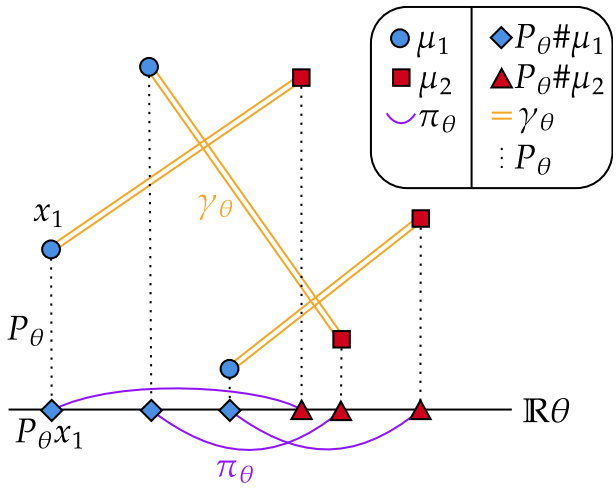


Figure 0.28: Étant donné des mesures discrètes avec des poids uniformes supportées sur des points ayant des projections distinctes sur la droite $\mathbb{R}\theta$, le plan de TO unidimensionnel entre les projections peut être relevé à un plan entre les deux mesures sans ambiguïté.

Introduite par [Mah+23], la technique de relèvement de la Fig. 0.28 est mal définie lorsque certains points ont la même projection, ou pour des mesures génériques dans $\mathcal{P}_2(\mathbb{R}^d)$. L'idée de la discrépance *Pivot Sliced* est de résoudre cette ambiguïté en sélectionnant des plans ayant un coût minimal en utilisant une mesure pivot (comme dans [Mah+23; NP23]). À l'inverse, [Liu+24] propose de contraindre un couplage indépendant pour résoudre l'ambiguïté. Avant de passer aux définitions mathématiques formelles, nous illustrons les deux approches dans la Fig. 0.29.

Définition formelle de la discrépance *Pivot Sliced*. Formellement, la discrépance *Pivot Sliced* entre $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ est définie comme suit :

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) := \min_{\rho \in \mathcal{P}_2(\mathbb{R}^{3d}):\ \rho_{0,1} \in \Pi^*(\mu_\theta, \mu_1),\ \rho_{0,2} \in \Pi^*(\mu_\theta, \mu_2)} \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2),$$

où pour $\rho \in \mathcal{P}_2(\mathbb{R}^{3d})$ et $i \in \{1, 2\}$, $\rho_{0,i} \in \mathcal{P}_2(\mathbb{R}^{2d})$ désigne la bi-marginale $P_{0,i} \# \rho$ avec $P_{0,i} := (y, x_1, x_2) \longmapsto (y, x_i)$. La mesure $\mu_\theta \in \mathcal{P}_2(\mathbb{R}^d)$ est la mesure milieu pour 2-Wasserstein entre $Q_\theta \# \mu_1$ et $Q_\theta \# \mu_2$ avec $Q_\theta := x \longmapsto (\theta^\top x)\theta$. Enfin, $\Pi^*(\mu_\theta, \mu_i)$ désigne l'ensemble des plans de transport optimaux entre μ_θ et μ_i pour le coût euclidien au carré. Nous montrons que PS_θ vérifie tous les axiomes pour être une distance à l'exception de l'inégalité triangulaire, et qu'elle majore W_2. Après avoir démontré de nouveaux résultats sur la distance *ν-based Wasserstein* [NP23], nous montrons que PS_θ est semi-continue inférieurement (mais pas continue) par rapport à la convergence faible des mesures. Nous montrons que PS_θ est égale à une autre discrépance CW_θ que nous appelons distance de Wasserstein contrainte, définie en ajoutant une contrainte le long de la droite $\mathbb{R}\theta$ dans le problème de Kantorovich.

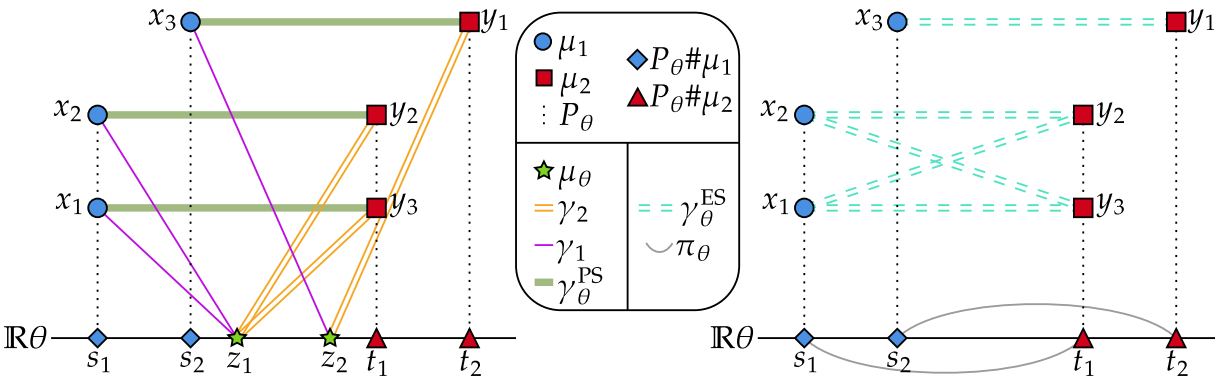


Figure 0.29: Comparaison des discrépances *Pivot Sliced* (gauche) et *Expected Sliced* (droite) pour deux mesures μ_1, μ_2 avec des poids uniformes supportées sur trois points ayant des projections non distinctes sur la droite $\mathbb{R}\theta$. Les projections sont $P_\theta\#\mu_1 = \frac{2}{3}\delta_{s_1} + \frac{1}{3}\delta_{s_2}$ et $P_\theta\#\mu_2 = \frac{2}{3}\delta_{t_1} + \frac{1}{3}\delta_{t_2}$. ***Pivot Sliced :*** d'abord, on calcule la mesure intermédiaire, ici $\mu_\theta = \frac{2}{3}\delta_{(s_1+t_1)/2} + \frac{1}{3}\delta_{(s_2+t_2)/2}$, puis pour $i \in \{1, 2\}$ on prend γ_i un plan optimal entre μ_θ et μ_i (dans ce cas, les deux sont uniques). On détermine maintenant un plan $\gamma_\theta^{\text{PS}} \in \Pi(\mu_1, \mu_2)$ qui minimise le coût de transport $\int \|x - y\|_2^2 \mathrm{d}\gamma_\theta^{\text{PS}}(x, y)$ parmi les plans cohérents avec γ_1 et γ_2 : ici, cela signifie que x_3 doit être apparié à y_1 car z_2 est apparié uniquement à x_3 dans γ_1 et z_2 est uniquement apparié à y_1 dans γ_2. Pour z_1, il y a ambiguïté, puisqu'il est assigné à la fois à x_1 et x_2 dans γ_1 et à la fois à y_2 et y_3 dans γ_2 : ici, nous prenons l'appariement qui associe x_1 à y_3 et x_2 à y_2, car cela minimise le coût de transport. ***Expected Sliced :*** on calcule le plan de TO 1D π_θ entre les projections $P_\theta\#\mu_1$ et $P_\theta\#\mu_2$, puis on le relève : d'abord, s_1 est assigné à t_1 par π_θ, donc on associe x_1 et x_2 à y_2 et y_3 via le couplage produit (indépendant). Ensuite, s_2 est apparié à t_2, donc on relie x_3 à y_1 (aucune ambiguïté cette fois).

Théorème. Pour toutes $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ et $\theta \in \mathbb{S}^{d-1}$, nous avons :

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) = \mathrm{CW}_\theta^2(\mu_1, \mu_2) := \min_{\substack{\omega \in \Pi(\mu_1, \mu_2) \\ (P_\theta, P_\theta)\#\omega = \pi_\theta}} \int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\omega(x_1, x_2),$$

où $P_\theta := x \longmapsto \theta^\top x$ et π_θ est l'unique plan de TO entre $P_\theta\#\mu_1$ et $P_\theta\#\mu_2$.

Formulation de Monge de la discrépance *Pivot Sliced* dans le cas discret. Dans le cas où $\mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ et $\mu_2 = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, il est naturel de rechercher une formulation de Monge pour le problème définissant la discrépance *Pivot Sliced*. La question est la suivante : dans ce cadre, un plan optimal pour PS_θ est-il nécessairement une permutation ? En montrant une version contrainte du théorème de Birkhoff-von Neumann [Bir46] (qui stipule que les points extrêmes du polytope des matrices doublement stochastiques sont les matrices de permutation), nous avons pu répondre positivement à cette question, comme détaillé ci-dessous.

Théorème. Pour tous $(x_1, \cdots, x_n) \in \mathbb{R}^{n\times d}$, $(y_1, \cdots, y_n) \in \mathbb{R}^{n\times d}$ et $\theta \in \mathbb{S}^{d-1}$, nous avons :

$$\mathrm{PS}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{(\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2,$$

où $\mathfrak{S}_\theta(X, Y) \subset \mathfrak{S}_n^2$ est l'ensemble des paires de permutations (σ, τ) telles que :

$$P_\theta x_{\sigma(1)} \leq \cdots \leq P_\theta x_{\sigma(n)}, \quad P_\theta y_{\tau(1)} \leq \cdots \leq P_\theta y_{\tau(n)}.$$

Discrépance Min-*Pivot Sliced*. Suivant les idées de [Mah+23], nous introduisons la discrépance Min-*Pivot Sliced* min PS comme le minimum de PS_θ sur tous les $\theta \in \mathbb{S}^{d-1}$. Nous montrons

que cette discrépance hérite des propriétés de PS_θ et qu'elle est égale à W_2 dans certains cas discrets favorables. Nous observons numériquement que l'inégalité triangulaire ne tient pas non plus pour min PS sur un exemple bien choisi.

Discrépance *Expected Sliced*. Nous avons pu généraliser la technique de relèvement de [Liu+24] illustrée dans la Fig. 0.29 à des mesures génériques, définissant ainsi une discrépance LS_θ sur $\mathcal{P}_2(\mathbb{R}^d)$ et des plans relevés $\gamma_\theta[\mu_1, \mu_2]$ pour tous $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. Cette discrépance LS_θ majore W_2 et vérifie tous les axiomes d'une distance, à l'exception du fait qu'il existe $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ tel que $\mathrm{LS}_\theta(\mu, \mu) > 0$. Le cas discret semble mieux adapté : si μ a un support discret dénombrable, alors $\mathrm{LS}_\theta(\mu, \mu) = 0$ pour presque tout $\theta \in \mathbb{S}^{d-1}$. La discrépance *Expected Sliced* est alors définie comme le coût de la *moyenne* pour $\mathfrak{s} \in \mathcal{P}(\mathbb{S}^{d-1})$ (une mesure de probabilité sur l'hypersphère) des plans relevés $\gamma_\theta[\mu_1, \mu_2]$, et s'écrit :

$$\mathrm{ES}^2_{\mathfrak{s}}(\mu_1, \mu_2) := \int_{\mathbb{S}^{d-1}} \mathrm{LS}^2_\theta(\mu_1, \mu_2) \mathrm{d}\mathfrak{s}(\theta) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^{2d}} \|x - y\|_2^2 \mathrm{d}\gamma_\theta[\mu_1, \mu_2](x, y) \mathrm{d}\mathfrak{s}(\theta).$$

Nous montrons que lorsque $\mathfrak{s}$ est absolument continue par rapport à la mesure uniforme sur $\mathbb{S}^{d-1}$, alors $\mathrm{ES}_{\mathfrak{s}}$ est une distance sur l'ensemble des mesures à support dénombrable. Cependant, nous montrons également que pour *toute* $\mathfrak{s} \in \mathcal{P}(\mathbb{S}^{d-1})$, choisir μ comme la mesure uniforme sur la boule unité de $\mathbb{R}^d$ donne $\mathrm{ES}_{\mathfrak{s}}(\mu, \mu) > 0$, et donc $\mathrm{ES}_{\mathfrak{s}}$ n'est jamais une distance sur $\mathcal{P}_2(\mathbb{R}^d)$.

0.4.2.3 Chapitre B.III : Sliced Gromov-Wasserstein

Dans le Chapitre B.III, nous commençons par rappeler l'algorithme Frank-Wolfe (FW) [FW+56] pour résoudre le problème notoirement difficile de Gromov-Wasserstein (GW) [Mém11]. Cette approche a été popularisée par [Vay+20] pour une variante de GW. Nous présentons ensuite une méthode d'optimisation alternée pour GW qui résout une borne inférieure, et expliquons comment chaque itération peut être vue comme un problème de TO. Nous étudions ensuite l'heuristique consistant à remplacer les problèmes de transport internes par des problèmes de transport *Sliced* en utilisant les plans *Pivot Sliced* introduits dans le Chapitre B.II. Motivée par des avantages computationnels, cette heuristique s'avère toutefois insatisfaisante en pratique.

0.4.3 Partie C : Barycentres de Transport Optimal

En tant que métrique sur l'espace des mesures de probabilité, la distance 2-Wasserstein fournit une manière naturelle de définir des barycentres au sens de Fréchet, comme introduit par [CE10; AC11] : un barycentre 2-Wasserstein de mesures $(\nu_1, \cdots, \nu_K) \in \mathcal{P}_2(\mathbb{R}^d)^K$ pour des poids $(\lambda_1, \cdots, \lambda_K) \in \triangle_K$ est toute mesure minimisant les distances au carré par rapport aux autres mesures :

$$\operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\mu, \nu_k).$$

Numériquement, le calcul d'un barycentre de mesures discrètes est complexe [Kro+19; AB22], et l'approche la plus utilisée consiste à minimiser par rapport à un support de taille fixe, donnant lieu à la fonctionnelle lagrangienne suivante :

$$X \in \mathbb{R}^{n \times d} \longmapsto \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\gamma_X, \nu_k),$$

où $\gamma_X := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Une approche itérative pour cette minimisation, consistant à utiliser les projections barycentriques des plans de TO, a été proposée dans [CD14], et demeure l'une des plus populaires pour les problèmes nécessitant l'optimisation du support (grâce à l'implémentation dans `POT` [Fla+21]).

Les généralisations de la notion de barycentres 2-Wasserstein sont surprenamment rares dans la littérature : le cas W_1 a été étudié récemment dans [CCE24], et le cas W_p^p encore plus récemment dans [BFR25], avec une extension aux coûts $c(x, y) = h(x - y)$ pour h lisse et fortement convexe dans [BFR24]. Une autre généralisation a été proposée dans [DGS21], où

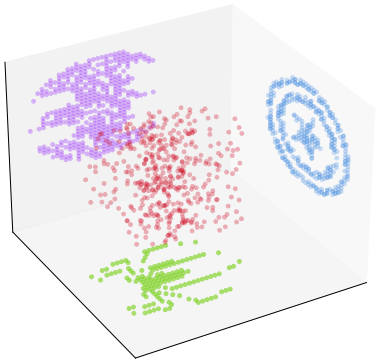
le barycentre est recherché dans $\mathcal{P}_2(\mathbb{R}^d)$ entre des mesures $\nu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$, en utilisant un coût $c_k(x,y) := \|P_k x - y_k\|_2^2$ avec $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$ des applications linéaires fixées. Cela donne l'expression variationnelle suivante, appelée problème de Barycentre Wasserstein Généralisé (GWB) :

$$\operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^{K} \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k).$$

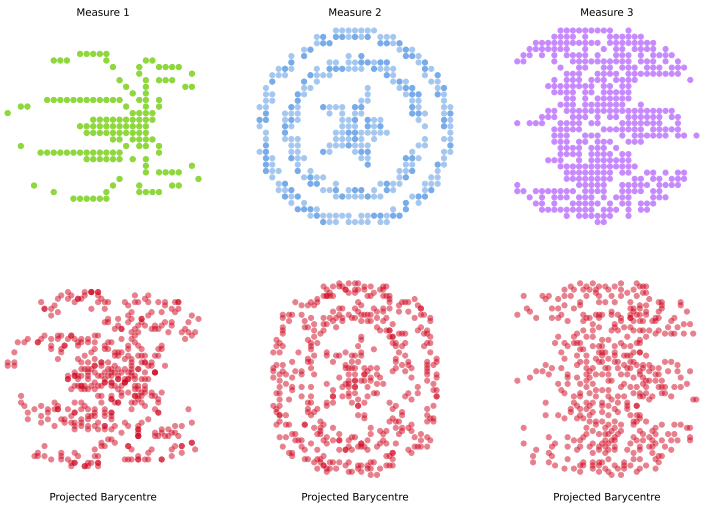
Dans le Chapitre C.I, nous étudions des méthodes numériques pour résoudre le problème GWB dans le cas discret, et introduisons une extension du problème qui optimise également sur les applications linéaires P_k. Dans le Chapitre C.II, nous proposons une notion très générale de barycentre entre des mesures situées dans différents espaces métriques. Nous proposons un algorithme à point fixe qui généralise la méthode introduite par [Álv+16] pour les barycentres 2-Wasserstein entre mesures absolument continues, et fournissons des algorithmes numériques ainsi que des liens avec la méthode populaire de "*free-support*" de [CD14].

0.4.3.1 Chapitre C.I : Barycentres de Wasserstein Généralisés (Aveugles)

Dans le Chapitre C.I, nous étudions le Barycentre de Wasserstein Généralisé (GWB), en proposant trois méthodes numériques supplémentaires pour sa résolution : GD, SGD et Descente Bloc-Coordonnée (BCD). En utilisant les résultats de régularité du TO du Chapitre B.I, nous avons pu appliquer un théorème de [EN97] pour montrer que les itérés de GD convergent sous-séquentiellement vers des points critiques de Clarke. De même, pour SGD, en utilisant à nouveau le Chapitre B.I, nous exploitons [Dav+20] pour montrer que les trajectoires de SGD convergent sous-séquentiellement vers des points critiques de Clarke presque sûrement, en supposant les itérées bornées. En pratique, la méthode BCD converge beaucoup plus rapidement vers un optimum local (qui peut ne pas être global), mais aucune garantie théorique n'a pu être établie. Nous illustrons le problème GWB dans la Fig. 0.30.



(a) Représentation 3D de la solution GWB.



(b) Projections de la solution GWB.

Figure 0.30: Résultats d'un GWB trouvant un barycentre dans $\mathbb{R}^3$ dont les projections correspondent à trois mesures discrètes données dans $\mathbb{R}^2$. Les mesures cibles sont représentées en bleu, vert et violet, tandis que le barycentre est représenté en rouge. À gauche, on observe le barycentre dans $\mathbb{R}^3$ avec les plongements des mesures bidimensionnelles. À droite, pour chacune des trois mesures cibles ν_k, nous comparons la mesure cible avec la projection $P_k \# \mu$ du barycentre μ dans $\mathbb{R}^3$.

Nous étendons également le problème GWB au cas où les applications linéaires P_k sont également optimisées. Après avoir montré l'existence de minimiseurs pour ce problème mal posé, nous adaptons nos algorithmes et montrons que les méthodes GD et SGD convergent dans le même sens que pour la résolution du GWB. Même si le problème est bien plus difficile à résoudre

numériquement, quelques astuces numériques simples permettent d'obtenir de bons résultats en pratique.

0.4.3.2 Chapitre C.II : Calcul des Barycentres de Transport Optimal

Généralisation des barycentres à tout coût. Dans le Chapitre C.II, nous définissons une notion de barycentre entre mesures $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$ où, pour $k \in [\![1, K]\!]$, l'espace $(\mathcal{Y}_k, d_{\mathcal{Y}_k})$ est un espace métrique compact. Un barycentre est recherché dans l'espace des mesures sur un espace métrique compact $(\mathcal{X}, d_{\mathcal{X}})$ en utilisant des coûts continus $c_k : \mathcal{X} \times \mathcal{Y}_k \longrightarrow \mathbb{R}_+$, et le problème de barycentre correspond à une moyenne de Fréchet généralisée :

$$\operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \sum_{k=1}^{K} \mathcal{T}_{c_k}(\mu, \nu_k) =: V(\mu).$$

Algorithme à point fixe. Pour formuler une méthode itérative pour ce problème, nous travaillons sous l'hypothèse que la moyenne de Fréchet généralisée entre les points $(y_k) \in \Pi_k \mathcal{Y}_k$ est bien définie (nous appelons cela le "barycentre au sol"), ce que nous formulons comme l'hypothèse que le problème suivant admet une solution unique :

$$B(y_1, \cdots, y_K) := \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k).$$

Notre algorithme à point fixe consiste en une large généralisation de la méthode proposée par [Álv+16] pour le barycentre 2-Wasserstein entre mesures absolument continues $(\nu_k) \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)^K$. À l'étape t, leur algorithme part de la mesure courante $\mu_t \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ et calcule les applications de TO T_k de μ_t vers ν_k. Cela fournit une linéarisation locale de l'espace 2-Wasserstein au sens de [MDC20], et un barycentre L^2 est calculé dans cet espace linéarisé, à savoir $S := \sum_{k=1}^{K} \lambda_k T_k$. L'étape suivante est alors $\mu_{t+1} := S \# \mu_t$, ce qui, en termes de variables aléatoires, peut s'écrire $X_{t+1} := \sum_k \lambda_k T_k(X_t)$ pour $X_t \sim \mu_t$. Nous résumons cette intuition de linéarisation dans la Fig. 0.31.

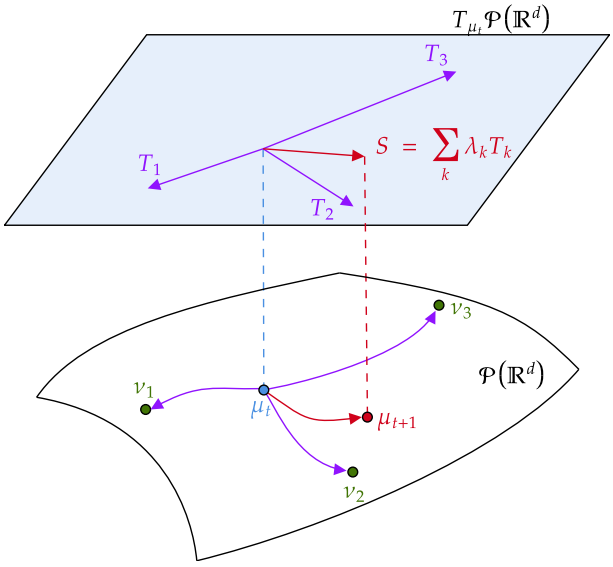


Figure 0.31: Visualisation de l'interprétation par linéarisation de l'algorithme à point fixe de [Álv+16] pour les barycentres 2-Wasserstein entre mesures absolument continues. L'espace $L^2(\mu_t)$ est vu informellement comme l'espace tangent de $\mathcal{P}_2(\mathbb{R}^d)$ en μ_t.

Dans notre cadre, il se peut que nous ne disposions pas d'applications de TO pour les coûts c_k, nous combinons donc des plans de TO plutôt que des applications de TO. À cette fin, nous introduisons un opérateur d'itération à valeurs multiples $G : \mathcal{P}(\mathcal{X}) \rightrightarrows \mathcal{P}(\mathcal{X})$ défini par :

$$\forall \mu \in \mathcal{P}(\mathcal{X}),\ G(\mu) := \left\{ \mathrm{Loi}[B(Y_1, \cdots, Y_K)] : \forall k \in [\![1, K]\!],\ \mathrm{Loi}[(X, Y_k)] \in \Pi^*_{c_k}(\mu, \nu_k) \right\},$$

où $\Pi^*_{c_k}(\mu,\nu_k)$ désigne l'ensemble des plans de TO entre μ et ν_k pour le coût c_k. Remarquons que si chaque $\Pi^*_{c_k}(\mu,\nu_k)$ est réduit à un unique plan induit par une application $(I,T_k)\#\mu$, alors $G(\mu)$ se réduit à la mesure unique $\mathrm{Loi}_{X\sim\mu}[B(T_1(X),\cdots,T_K(X))]$, ce qui correspond à prendre le barycentre au sol B des applications de TO T_k. À l'aide de cet opérateur multi-valué, nous définissons notre algorithme à point fixe par

$$\mu_0 \in \mathcal{P}(\mathcal{X}),\ \forall t\in\mathbb{N},\ \mu_{t+1}\in G(\mu_t),$$

qui est une généralisation de [Álv+16], puisque nous travaillons sans hypothèses de régularité sur les mesures et avec des coûts généraux c_k. En particulier, notre cadre permet de définir des barycentres entre mesures discrètes, ce qui est d'un intérêt primordial en pratique.

Convergence. Nous montrons que les résultats de convergence de [Álv+16] restent valables dans notre cadre, en exploitant notamment des résultats techniques sur la régularité de l'opérateur G. Premièrement, nous montrons que l'énergie des itérés est décroissante :

> **Proposition.** Soient $\mu\in\mathcal{P}(\mathcal{X})$ et $\overline{\mu}\in G(\mu)$. Il existe une fonction $\delta:\mathcal{X}^2\longrightarrow\mathbb{R}_+$ telle que $V(\mu)\geq V(\overline{\mu})+\mathcal{T}_\delta(\mu,\overline{\mu})$. Si μ^* est un barycentre, alors $G(\mu^*)=\{\mu^*\}$.

La fonction δ plus haut est explicite et possède plusieurs propriétés utiles que nous ne détaillons pas ici, mais notons que dans le cas W_2^2, il s'agit simplement de $\|x-y\|_2^2$. Grâce à ce résultat et à des résultats techniques de régularité sur G, nous obtenons le résultat de convergence suivant, qui étend celui de [Álv+16] :

> **Théorème.** Pour toute $\mu_0\in\mathcal{P}(\mathcal{X})$, soit (μ_t) telle que $\mu_{t+1}\in G(\mu_t)$. Alors (μ_t) possède des sous-suites convergentes, et toute sous-suite convergeant faiblement converge nécessairement vers $\mu\in\mathcal{P}(\mathcal{X})$ tel que $\mu\in G(\mu)$.

Illustrations numériques. Nous fournissons une implémentation de cet algorithme, en construisant en particulier un couplage adapté $(Y_1,\cdots,Y_K)$ à l'aide d'une généralisation de la *North-West Corner Rule*, qui est suffisamment parcimonieux pour éviter des problèmes de mémoire. Nous étudions également une simplification heuristique de G qui contraint la taille du support et le poids du barycentre à rester constants, ce qui peut être vu comme une généralisation directe de la méthode "*free-support*" de [CD14]. Nous montrons également que notre théorie s'étend aux barycentres pour des coûts entropiques, et détaillons l'application de nos méthodes au calcul des barycentres entre GMMs au sens de [DD20]. Dans la Fig. 0.32, nous illustrons une application numérique à un barycentre pour les coûts $c_k(x,y)=\|P_k(x)-y\|_2^2$, où P_k sont des projections (non-linéaires) sur des cercles, ce qui peut être vu comme une généralisation non-linéaire du problème GWB [DGS21] étudié dans le Chapitre C.I.

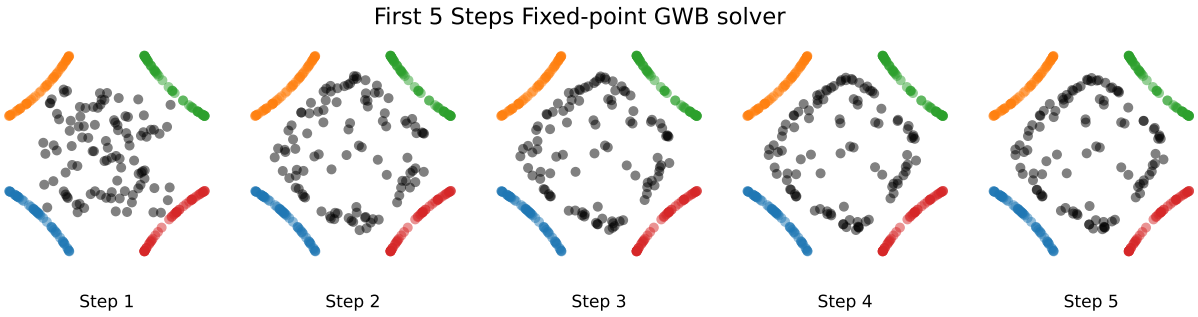


Figure 0.32: Premières 5 itérations de l'algorithme à points fixes pour les coûts $c_k(x,y)=\|P_k(x)-y\|_2^2$, où P_k sont des projections sur quatre cercles différents sur lesquels les ν_k sont supportées (représentés en couleur).

0.4.4 Partie D : Quelques Contributions aux Méthodes à Noyaux

Dans la dernière partie de cette thèse, nous présentons certaines contributions à un domaine de recherche relativement indépendant, celui des méthodes à noyaux. L'origine de ces projets provient de la tentative de résoudre le problème d'application contrainte du Chapitre B.I en utilisant des méthodes à noyaux. Malheureusement, nous montrons dans le Chapitre D.II que notre idée instinctive pour aborder ce problème avec des noyaux ne peut aboutir, car aucun noyau ne satisfait les propriétés requises. Pour introduire le domaine des Espaces de Hilbert à Noyau Reproductif (*Reproducing Kernel Hilbert Space* RKHS), nous commençons la Partie D par un rappel de la théorie des RKHS dans le Chapitre D.I, qui constitue une transcription de présentations introductives données par l'auteur au laboratoire MAP5. Enfin, dans le Chapitre D.III, nous présentons une construction explicite de noyaux universels sur des espaces métriques compacts, qui a suscité un certain intérêt dans le projet du Chapitre C.II, mais qui s'est avéré inutile pour ce dernier et a donné lieu à une direction de recherche distincte. Nous introduisons également une notion d'universalité approchée, et montrons que deux discrétisations du noyau universel proposé vérifient cette propriété, ce qui est utile en pratique compte tenu de leur calculabilité.

0.4.5 Résumé des Contributions en Code Source Ouvert

Au-delà des considérations théoriques de cette thèse, nous fournissons également de nombreux algorithmes avec des implémentations en code source ouvert en Python. La majeure partie de ce code a été contribué à la bibliothèque Python Optimal Transport (`POT`) [Fla+21], dont l'auteur est mainteneur.

0.4.5.1 Code Public et Reproductible

Tout le code des algorithmes et des expériences présentés dans le Chapitre C.I est disponible publiquement sur le dépôt Github https://github.com/eloitanguy/bgwb. De même, pour le Chapitre C.II, toutes les implémentations sont accessibles à l'adresse https://github.com/eloitanguy/ot_bar. Cette dernière a reçu le titre "label reproductible Argent" lors de la soumission d'une communication sur le Chapitre C.II au congrès GRETSI 2025. Une implémentation publique pour le Chapitre A.IV est en cours, tout comme pour le Chapitre B.II.

0.4.5.2 Contributions à la Bibliothèque POT

Tout au long de cette thèse, l'auteur a contribué à la bibliothèque `POT` [Fla+21] à travers plusieurs *pull requests* (PRs) implémentant des algorithmes liés à ce travail.

- Barycentres de Wasserstein Généralisés ([DGS21], Chapitre C.I): PR 372
- *Smooth Strongly Convex Nearest Brenier Potentials* ([PdC20], Chapitre B.I): PR 526
- Transport Optimal pour modèles de mélanges gaussiens ([DD20], Chapitres C.II and A.IV): PR 649
- Solveurs à point fixe pour barycentres de TO (Chapitre C.II): PR 715

Part A

Optimal Transport Discrepancies as Losses

Chapter A.I studies the solutions of a problem recovering a measure whose images by linear maps must equate the images of a fixed discrete measure. This chapter is based on the paper:

[TFD24b] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Reconstructing discrete measures from projections.
Consequences on the empirical Sliced Wasserstein Distance".
Comptes Rendus. Mathématique 362 (Jun. 2024), pp. 1121-1129.

Chapter A.II focuses on the energy defined by the Sliced Wasserstein distance between two discrete probability measures, and studies its regularity and optimisation properties as a function of the support of one of the measures. This chapter is based on the paper:

[TFD24a] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Properties of Discrete Sliced Wasserstein Losses".
Mathematics of Computation (Jun. 2024).

Chapter A.III studies the convergence of Stochastic Gradient Descent (SGD) trajectories for training generative neural networks with the Sliced Wasserstein loss. This chapter is based on the paper:

[Tan23] Eloi Tanguy.
"Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses".
Transactions on Machine Learning Research (Oct. 2023).

Chapter A.IV introduces methods to differentiate the Expectation-Maximisation algorithm with respect to the input data, and studies its use for Gaussian Mixture Model Optimal Transport. This chapter is based on the paper:

[Boï+25] Samuel Boïté*, Eloi Tanguy*, Julie Delon, Agnès Desolneux and Rémi Flamary.
"Differentiable Expectation-Maximisation
and Applications to Gaussian Mixture Model Optimal Transport".
arxiv preprint 2509.02109 (Sept. 2025). (*: equal contribution)

A.I

Reconstructing Discrete Measures from Projections

A.I.1 Introduction . . . 45
A.I.2 Solutions of the Reconstruction Problem . . . 46
A.I.2.1 Computing Linear Push-Forwards of Discrete Measures 47
A.I.2.2 Restraining the support of solutions of RP 47
A.I.2.3 Conditions for unicity of solutions of RP 48
A.I.2.4 Details on the critical case $\sum_i d_i = d$ 50
A.I.3 Consequence for the empirical Sliced Wasserstein Distance . . . 51
A.I.4 Conclusion: Discussion on SW as a Loss in Machine Learning . . . 52

Abstract

This chapter deals with the reconstruction of a discrete measure γ_Z on $\mathbb{R}^d$ from the knowledge of its pushforward measures $P_i\#\gamma_Z$ by linear applications $P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$ (for instance projections onto subspaces). The measure γ_Z being fixed, assuming that the rows of the matrices P_i are independent realizations of laws which do not give mass to hyperplanes, we show that if $\sum_i d_i > d$, this reconstruction problem has almost certainly a unique solution. This holds for any number of points in γ_Z. A direct consequence of this result is an almost-sure separability property on the empirical Sliced Wasserstein distance. This chapter is based on the paper:

[TFD24b] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Reconstructing discrete measures from projections.
Consequences on the empirical Sliced Wasserstein Distance".
Comptes Rendus. Mathématique 362 (Jun. 2024), pp. 1121-1129.

A.I.1 Introduction

In this chapter, we are interested in the following question: for a given discrete probability measure γ_Z on $\mathbb{R}^d$, and r linear transformations $P_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$, can we characterize the set of probability measures on $\mathbb{R}^d$ with exactly the same images as γ_Z through all of the maps P_i? Formally, this set writes

$$\mathcal{S} = \left\{ \gamma \in \mathcal{P}(\mathbb{R}^d) \mid \forall i \in [\![1, r]\!],\ P_i\#\gamma = P_i\#\gamma_Z \right\}, \tag{RP}$$

where $P_i\#\gamma$ denotes the push-forward of γ by P_i, *i.e.* the measure on $\mathbb{R}^{d_i}$ such that for any Borelian $A \subset \mathbb{R}^{d_i}$, $(P_i\#\gamma)(A) = \gamma(P_i^{-1}(A))$, and $\mathcal{P}(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$. The set $\mathcal{S}$ is nonempty since it contains at least γ_Z. A natural underlying question is to know when we get uniqueness, *i.e.* when $\mathcal{S} = \{\gamma_Z\}$. Indeed, in this case γ_Z can be exactly reconstructed from the knowledge of all the $P_i\#\gamma_Z$, which is why we refer to this problem as a reconstruction problem.

This reconstruction problem appears in many applied fields where a multidimensional measure is known only through a finite set of images or projections. This is the case for instance in medical or geophysical imaging problems such as tomography [GG99]. It is also strongly related to the separability properties of the empirical version of the Sliced Wasserstein distance [Rab+12; Bon+15a], which is frequently used in machine learning applications [Kar+18; DZS18; Wu+19].

In our reconstruction problem, it is clear that if one of the P_i is injective (which implies $d \leq d_i$), then $\mathcal{S} = \{\gamma_Z\}$, which is why we focus here on the cases where none of the P_i is injective. We will also assume in this chapter that the P_i are surjective and that all the d_i are strictly smaller than d, since we can always replace $\mathbb{R}^{d_i}$ by the smaller subspace $\mathrm{Im}(P_i)$. To the best of our knowledge, this problem has not been widely discussed in the literature, perhaps because of its apparent simplicity. A close and more discussed question is the one of the existence of probability measures γ with marginal constraints [KR19; DKS12; Kel64]. Existence results for such couplings are known for some families of measures [Joe93], or measures exhibiting some specific correlation structures [Cua92]. However, in the general case, even if marginal constraints are compatible with each other, the existence of solutions is not always ensured [GKZ19].

Our study case is different, since the constraints are all obtained as push-forwards of an unknown γ_Z, and the central question is not existence but uniqueness of solutions. It is well known that a measure is uniquely determined by its projections on *all* lines of $\mathbb{R}^d$ (Cramer-Wold theorem [CW36]), and more generally by its projections on a set of subspaces as soon as they cover the whole space together [Rén52]. The problem for a non-discrete target measure has been studied by Chafai [Cha21] which considers a similar reconstruction problem from the viewpoint of characteristic functions, using the equivalence $\forall i \in [\![1, r]\!],\ P_i\#\mu = P_i\#\nu = 0 \iff \forall i \in [\![1, r]\!],\ \forall x \in \mathbb{R}^d,\ \phi_\mu(P_i x) = \phi_\nu(P_i x)$, where ϕ_μ and ϕ_ν denote the respective characteristic functions of μ and ν. In general, the knowledge of the characteristic function on a finite union of strict subspaces is insufficient to determine a measure [Cha21]. For a finite number of directions and in the case of a discrete measure γ_Z, simple linear algebra shows that if the number r of projections is large enough, we get $\mathcal{S} = \{\gamma_Z\}$. When the P_i are projections on different hyperplanes for instance, Heppes showed in 1956 [Hep56] that a discrete distribution of at most n points $\gamma_Z = \frac{1}{n}\sum_{\ell=1}^n \delta_{z_\ell}$ is uniquely characterized by its projections $P_i\#\gamma_Z$ if the number r of these projections is larger than $n+1$, and that simple counter-examples could be exhibited with only $r = n$ hyperplanes. More recent works [FJG17] show that uniqueness can be ensured with less projections as soon as the set of points is known to belong to a specific quadratic manifold. These results are deterministic, they hold for every set of points and hyperplanes with the appropriate cardinality. In this chapter, we add some stochasticity to the problem, and assume that the lines of the matrices P_i [1] are i.i.d. following a law $\mathbb{P}$ which does not give mass to hyperplanes. Under this assumption, we show that if $\sum_{i=1}^r d_i > d$, then $\mathbb{P}$-almost surely $\mathcal{S} = \{\gamma_Z\}$. This result is very different from the ones already present in the literature: it holds only a.s., but this permits a considerably weaker condition on the P_i, and the condition for the reconstruction surprisingly does not depend on the number of points.

A.I.2 Solutions of the Reconstruction Problem

In this section, we characterize the set $\mathcal{S}$ of solutions defined in Eq. (RP) depending on the set of linear maps P_i. We write $\gamma_Z = \sum_{\ell=1}^n b_\ell \delta_{z_\ell}$ with $Z = (z_1, \cdots, z_n) \in (\mathbb{R}^d)^n$, and assume that all points are distinct ($k \neq j \implies z_k \neq z_j$). We also always assume that $n > 1$. The weights $(b_\ell) \in (\mathbb{R}_+^*)^n$ sum to one and are each nonzero.

As we shall see, given a discrete measure γ_Z with n points in dimension d, the Reconstruction Problem Eq. (RP) has a unique solution $\mathcal{S} = \{\gamma_Z\}$ almost-surely when drawing the linear maps P_i randomly, and when the dimensions strictly exceed d, *i.e.* when $D := d_1 + \cdots + d_r > d$.

[1] With a slight abuse of notation, we use the same notation here for the linear maps and their associated matrices.

A.I.2.1 Computing Linear Push-Forwards of Discrete Measures

Characterizing $\mathcal{S}$ requires the following technical Lemma, which provides a geometrical viewpoint of the push-forward operation.

Lemma A.I.1 (Linear push-forward formula)**.** Let $P \in \mathcal{M}_{h,d}(\mathbb{R})$ of rank $h \leq d$ and $B \subset \mathbb{R}^h$.
Then $P^{-1}(B) = P^\top (PP^\top)^{-1} B + \mathrm{Ker} P$.

Fig. A.I.1 shows a visualization of the set $P^\top (PP^\top)^{-1} B + \mathrm{Ker} P$, first where B is comprised of two points of $\mathbb{R}^2$ and $\mathrm{Ker} P$ is a horizontal plane in 3D, and second with B a measurable set of $\mathbb{R}^2$. This illustrates the ill-posedness of the problem when the dimension of the projections and number of projections is too small. In this case with $r = 1$, $d = 3$ and $d_1 = 2$, the condition $P^{-1}(A) = P^{-1}(B)$ leaves a degree of freedom, which we can visualize as the vertical axis here.

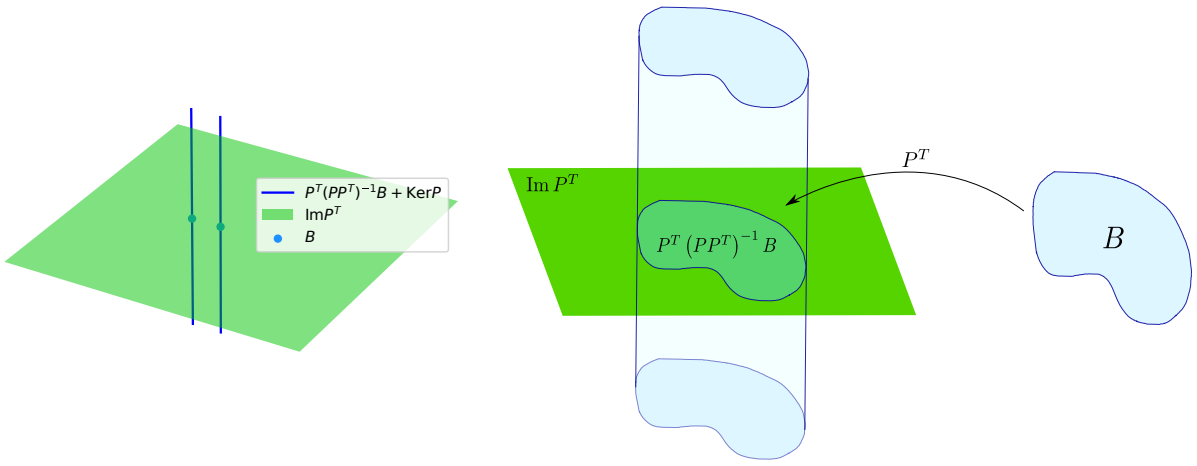


Figure A.I.1: Illustrations of the linear push-forward formula $P^{-1}(B)$ for a 3D to 2D projection P, (left) when B is a set of two points and (right) for a more general set B.

Proof. If $a \in P^\top (PP^\top)^{-1} B + \mathrm{Ker} P$, then by writing $a = P^\top (PP^\top)^{-1} b + x$ with $b \in B$ and $x \in \mathrm{Ker} P$, we have $Pa = b \in B$, thus $a \in P^{-1}(B)$.

For the opposite inclusion, consider $a \in P^{-1}(B)$. Since P is of full rank h, we have the decomposition $\mathbb{R}^d = \mathrm{Im} P^\top \overset{\perp}{\oplus} \mathrm{Ker} P$, with $Q := P^\top (PP^\top)^{-1} P$ the orthogonal projection on $\mathrm{Im} P^\top$.

Thus we can write $a = Qa + (I - Q)a = P^\top (PP^\top)^{-1} Pa + (I - Q)a$. Since $Pa \in B$, we conclude that $a \in P^\top (PP^\top)^{-1} B + \mathrm{Ker} P$. □

A.I.2.2 Restraining the support of solutions of RP

The following theorem states that the support of any solution of Eq. (RP) is constrained to a set S obtained as the intersection of all sets $Z + \mathrm{Ker} P_i$. Without loss of generality, we will assume that each P_i is of full rank d_i.

Theorem A.I.1 (Support of solutions of Eq. (RP))**.**
If γ is a solution of Eq. (RP), then $\gamma(S) = 1$ with

$$S := \bigcap_{i=1}^{r} (Z + \mathrm{Ker} P_i) = \bigcup_{(\ell_1, \cdots, \ell_r) \in [\![1,n]\!]^r} \; \bigcap_{i=1}^{r} (z_{\ell_i} + \mathrm{Ker} P_i)\,. \tag{A.I.1}$$

Proof. Using the same notations as in the proof of Lemma A.I.1, we write $Q_i := P_i^\top (P_i P_i^\top)^{-1} P_i$

the orthogonal projection on $\mathrm{Im}P_i^\top$ and we recall the decomposition $\mathbb{R}^d = \mathrm{Im}P_i^\top \overset{\perp}{\oplus} \mathrm{Ker}P_i$. Thus, for any borelian set A of $\mathbb{R}^d$, $A \subset Q_i A + \mathrm{Ker}P_i$. Then $\gamma(A) \leq \gamma(Q_i A + \mathrm{Ker}P_i) = P_i\#\gamma(P_i A)$, where the last equality is a direct consequence of Lemma A.I.1.

Now, assume that $\gamma \in \mathcal{S}$ and define $S := \bigcap_{i=1}^{r} K_i$ with $K_i = Z + \mathrm{Ker}P_i$. For each i, applying the previous inequality to K_i^c yields $\gamma(K_i^c) \leq P_i\#\gamma(P_i K_i^c)$. Since γ is a solution, we have $P_i\#\gamma = P_i\#\gamma_Z$. By construction, $K_i = \{x,\ P_i x \in P_i Z\}$ thus $K_i^c = \{x,\ P_i x \notin P_i Z\}$. Since $P_i\#\gamma$ is supported by $P_i Z$, it follows that $P_i\#\gamma(P_i K_i^c) = 0$. Finally, $\gamma(S^c) = \gamma(\bigcup_{i=1}^{r} K_i^c) \leq \sum_{i=1}^{r} \gamma(K_i^c) = 0$ and thus $\gamma(S) = 1$.

□

Fig. A.I.2 illustrates the previous result, with $r = 2$ projections onto lines in $\mathbb{R}^2$, with $n = 3$ points $Z = (z_1, z_2, z_3)$. The support of any solution is confined to the intersections between any two lines of the form $z_\ell + \mathrm{Ker}P_i$. Here this corresponds to the intersecting points between an orange and a red line, allowing for 9 possible points, including the original 3. In this case any weighting of the 9 Dirac masses that respect the marginal constraints will give a solution: there exists an infinity of possible solutions.

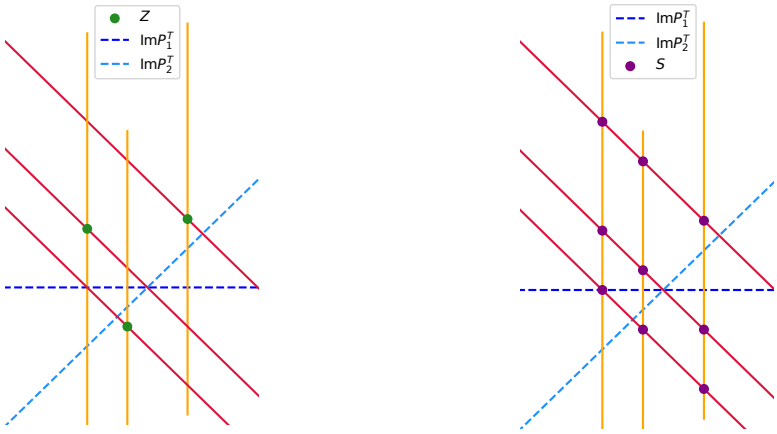


Figure A.I.2: Illustration of the possible points for the support of a solution. On the left, Z is the original measure points, and on the right, S is the set of possible points for the support of a solution.

A.I.2.3 Conditions for unicity of solutions of RP

Leveraging the previous support restriction and elementary random affine geometry, we can further restrict the condition on the set of solutions $\mathcal{S}$. Theorem A.I.2 below shows that if the random linear maps P_i cover the original space $\mathbb{R}^d$ with redundancy (i.e. the sum of their target space dimensions strictly exceeds d), then almost surely, the reconstruction problem has a unique solution γ_Z. We formalize this random setting by the following assumption.

Assumption A.I.1 ($\mathcal{A}_\mathbb{P}$).

$$\forall i \in [\![1, r]\!],\ P_i = \begin{pmatrix} — & \left(u_i^{(1)}\right)^\top & — \\ — & \vdots & — \\ — & \left(u_i^{(d_i)}\right)^\top & — \end{pmatrix} \quad \text{where } u_i^{(j)} \sim \mathbb{P} \text{ i.i.d,}$$

where $\mathbb{P}$ is a probability distribution over $\mathbb{R}^d$ s.t. for any hyperplane $H \subset \mathbb{R}^d$, $\mathbb{P}(H) = 0$.

The condition on the probabilities is verified in particular if $\mathbb{P}$ is absolutely continuous w.r.t. the Lebesgue measure of $\mathbb{R}^d$, or w.r.t. $\mathfrak{o}$, the uniform measure over $\mathbb{S}^d$ (the unit sphere of $\mathbb{R}^d$).

These two examples are the most common for practical reconstruction problems, which is why we formulate $\mathcal{A}_{\mathbb{P}}$ in this manner.

The next theorems use assumption $\mathcal{A}_{\mathbb{P}}$ but still hold true under milder hypotheses, where the lines $\left(u_i^{(j)}\right)^\top$ of the matrices P_i are assumed independent with (possibly different) probability laws giving no mass to hyperplanes.

Theorem A.I.2 (Almost-sure unicity in Eq. (RP))**.** Let γ_Z be a fixed discrete probability measure. Assume that the matrices P_i follow assumption $\mathcal{A}_{\mathbb{P}}$, and that $D := \sum_{i=1}^{r} d_i > d$. Then $\mathbb{P}$-almost surely, $S = Z$ and $\mathcal{S} = \{\gamma_Z\}$.

The idea behind the proof of Theorem A.I.2 is that S is the union of sets of the form $\bigcap_{i=1}^{r} (z_{\ell_i} + \mathrm{Ker} P_i)$, which can be rewritten as intersections of more than d affine subspaces in dimension d, thus are $\mathbb{P}$-almost surely either singletons or empty.

Proof. — *Step 1*: $S = Z$

Let $\boldsymbol{\ell} := (\ell_1, \cdots, \ell_r) \in [\![1, n]\!]^r$ and $S_{\boldsymbol{\ell}} := \bigcap_{i=1}^{r} (z_{\ell_i} + \mathrm{Ker} P_i)$. We want to show $S_{\boldsymbol{\ell}} \subset Z$.

First, observe that $x \in S_{\boldsymbol{\ell}} \iff \forall i \in [\![1, r]\!],\ \forall j \in [\![1, d_i]\!],\quad (u_i^{(j)})^\top x = (u_i^{(j)})^\top z_{\ell_i}$.

We write $D = \sum_{i=1}^{r} d_i$.

For the sake of simplicity, we rewrite the k^{th} vector $u_i^{(j)}$ as v_k, where $k \in [\![1, D]\!]$, and in the same way we write $(w_k)_{k=1\cdots D}$ the vectors $(z_{\ell_1}, \cdots, z_{\ell_1}, z_{\ell_2}, \cdots, z_{\ell_2}, \cdots, z_{\ell_r}, \cdots, z_{\ell_r})$ with each z_{ℓ_i} repeated d_i times. With these notations, we get

$$x \in S_{\boldsymbol{\ell}} \iff v_k^\top x = v_k^\top w_k,\ \ \forall k \in [\![1, D]\!]. \tag{A.I.2}$$

Let us call (LS) the linear system on the right of Eq. (A.I.2). (LS) has D equations and d unknowns, with $D > d$, it is therefore overdetermined. When all w_k are equal, *i.e.* when $\boldsymbol{\ell} := (\ell, \cdots, \ell)$, clearly $x = z_\ell$ is a solution, which shows that $z_\ell \in S$ and thus $Z \subset S$.

If $\mathcal{A}_{\mathbb{P}}$ is satisfied, the matrix $U^{(d)} = (v_1, \cdots, v_d)^\top$ is almost surely of full rank and the linear system $v_k^\top x = v_k^\top w_k$ for $k \in [\![1, d]\!]$ almost surely has a unique solution x^*. The $(d+1)^{\text{th}}$ equality of (LS) is $v_{d+1}^\top (x^* - w_{d+1}) = 0$, which happens iif $x^* = w_{d+1}$, or $x^* \neq w_{d+1}$ and $v_{d+1} \in (x^* - w_{d+1})^\perp$. In the first case, the solution x^* belongs to Z since w_{d+1} is one of the z_{ℓ_i}. If $x^* \neq w_{d+1}$, since all the $\{v_k\}$ are i.i.d. of law $\mathbb{P}$, conditionally to $U^{(d)}$ the probability that v_{d+1} is orthogonal to $(x^* - w_{d+1})$ is null and (LS) has almost surely no solution. We conclude that $S = Z$ almost surely.

— *Step 2*: The set of solutions of Eq. (RP) is $\{\gamma_Z\}$ a.s.

We have proven that $S = Z$ a.s., and thus that any solution $\gamma \in \mathcal{S}$ is supported by Z a.s.. Let us write $\gamma = \sum_{\ell=1}^{n} a_\ell \delta_{z_\ell}$ and $\gamma_Z = \sum_{\ell=1}^{n} b_\ell \delta_{z_\ell}$. It follows in particular that $\sum_{\ell=1}^{n} (a_\ell - b_\ell) \delta_{P_1 z_\ell} = 0$, and since for $k \neq \ell$, $z_k \neq z_\ell$, hence $\mathbb{P}(P_1 z_\ell = P_1 z_k) \leq \mathbb{P}(u_1^{(1)} \in (z_\ell - z_k)^\perp) = 0$, thus $\forall \ell$, $a_\ell = b_\ell$ a.s.. □

The previous Theorem A.I.2 only holds almost-surely, however "improbable" counter examples do exist with excessive symmetry. Below we present a counter-example adapted from [FJG17]. Let $d := 2$, $r := n > d$ and $\forall i \in [\![1, r]\!]$, $d_i := 1$.

Consider $z_\ell := \left(\cos\left(\frac{(2\ell+1)\pi}{n}\right), \sin\left(\frac{(2\ell+1)\pi}{n}\right)\right)^\top$, $P_\ell := \left(\cos\left(\frac{(2\ell+1)\pi}{2n}\right), \sin\left(\frac{(2\ell+1)\pi}{2n}\right)\right)$.

As can be seen below (Fig. A.I.3), for $n = 3$, this corresponds to placing the (z_ℓ) on every other vertex of a regular $2n$-gon, and defining the P_ℓ such that $\mathrm{Im} P_\ell^\top$ is the ℓ-th bisector of the $2n$-gon.

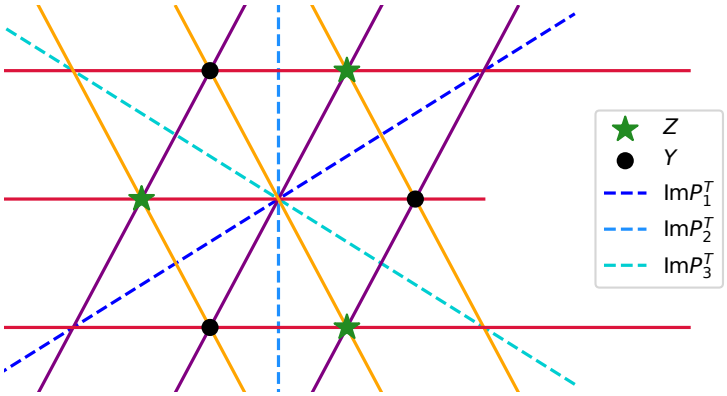


Figure A.I.3: Illustration of a pathological super-critical case without unicity for specific projections P_i. In this case, Y and Z are distinct solutions with the same projections.

The points of S are the points of the form $\bigcap_{i=1}^{3}(z_{\ell_i}+\mathrm{Ker}P_i)$, or visually the intersection points of a yellow line, a red line and a purple line. We can see that the remaining vertices of the polygon constitute another valid measure γ_Y whose push-forwards $P_i\#\gamma_Y$ are all the same as those of the original measure.

As mentioned in [FJG17], for a fixed list of hyperplanes, there always exists two sets of points with the same projections on all of these hyperplanes. [FJG17, Theorem I.2] indicates that a necessary condition to ensure uniqueness in this case is $r > n$. In our Theorem A.I.2, the points are fixed and uniqueness of the reconstruction holds almost surely when the P_i follow assumption $\mathcal{A}_{\mathbb{P}}$ and as soon as $D > d$, whatever the number n of points in the discrete measure.

A.I.2.4 Details on the critical case $\sum_i d_i = d$

In the theorem below, we show that the example in Fig. A.I.2 is representative of the critical case.

Theorem A.I.3 (Number of admissible points in the critical case)**.** Let γ_Z be a fixed discrete probability measure. Assume that the matrices P_i follow assumption $\mathcal{A}_{\mathbb{P}}$, and that $D := \sum_{i=1}^{r} d_i = d$. Then the cardinality of S is exactly n^r, $\mathbb{P}$-a.s..

Proof. Using the notation $\boldsymbol{\ell} := (\ell_1, \cdots, \ell_r) \in [\![1,n]\!]^r$, we know that $S = \bigcup_{\boldsymbol{\ell}\in[\![1,n]\!]^r} S_{\boldsymbol{\ell}}$ where $S_{\boldsymbol{\ell}} = \bigcap_{i=1}^{r}(z_{\ell_i}+\mathrm{Ker}P_i)$. Following the proof of Theorem A.I.2, in the case $D = d$, we see that assumption $\mathcal{A}_{\mathbb{P}}$ implies that $S_{\boldsymbol{\ell}}$ is almost surely a singleton $\{x_{\boldsymbol{\ell}}\}$. It follows that S is almost surely the union of at most n^r singletons. Let us show that if $\boldsymbol{\ell} \neq \boldsymbol{\ell}'$ then $S_{\boldsymbol{\ell}} \cap S_{\boldsymbol{\ell}'} = \varnothing$ a.s.. Indeed, if x belongs to $S_{\boldsymbol{\ell}} \cap S_{\boldsymbol{\ell}'}$ then x is solution of a linear system of $2d$ equations:

$$\forall i \in [\![1,r]\!],\ \forall j \in [\![1,d_i]\!],\quad \begin{cases} (u_i^{(j)})^\top x = (u_i^{(j)})^\top z_{\ell_i} \\ (u_i^{(j)})^\top x = (u_i^{(j)})^\top z_{\ell'_i} \end{cases},$$

which implies $\forall i \in [\![1,r]\!],\ \forall j \in [\![1,d_i]\!],\quad \ell_i = \ell'_i,\ \text{ or }\ \ell_i \neq \ell'_i \text{ and } u_{\ell_i}^{(j)} \in (z_{\ell_i} - z_{\ell'_i})^\perp$.

Now, under $\mathcal{A}_{\mathbb{P}}$, if $\ell_i \neq \ell'_i$, then $\mathbb{P}(u_{\ell_i}^{(j)} \in (z_{\ell_i} - z_{\ell'_i})^\perp) = 0$, and thus $\boldsymbol{\ell} = \boldsymbol{\ell}'$ a.s.. □

Let us clarify what the set of solutions $\mathcal{S}$ looks like in this critical case $D = d$. Let γ be a solution of Eq. (RP) and denote $S = (x_{\boldsymbol{\ell}})_{\boldsymbol{\ell}\in[\![1,n]\!]^r}$. By Theorem A.I.3, γ is of the form $\gamma = \sum_{\boldsymbol{\ell}\in[\![1,n]\!]^r} a_{\boldsymbol{\ell}}\delta_{x_{\boldsymbol{\ell}}}$. that by construction, $P_i x_{\mathbf{l}} = P_i z_{l_i}$. Now, since γ is a solution, we have for

$i \in [\![1,r]\!]$, $P_i\#\gamma = P_i\#\gamma_Z$, thus $\sum_{\boldsymbol{\ell}\in[\![1,n]\!]^r} a_{\boldsymbol{\ell}}\delta_{P_i z_{\ell_i}} = \sum_{k=1}^n b_k \delta_{P_i z_k}$. Since the $(P_i z_\ell)_\ell$ are all distinct a.s., this entails for all $k \in [\![1,n]\!]$: $\sum_{\boldsymbol{\ell}_{-i}\in[\![1,n]\!]^{r-1}} a_{\ell_1,\cdots,\ell_{i-1},k,\ell_{i+1},\cdots,\ell_r} = b_k$, where $\boldsymbol{\ell}_{-i}$ indicates that we index this $(r-1)$-tuple on $[\![1,n]\!]\setminus\{i\}$. We can re-write this condition as $a \in \Pi_n^r(b)$, the set of n-dimensional r-tensors on $\mathbb{R}_+$ ($\mathbb{R}_+^{n^r}$) with all r marginals equal to b. Conversely, if γ is of the form $\gamma = \sum_{\boldsymbol{\ell}\in[\![1,n]\!]^r} a_{\boldsymbol{\ell}}\delta_{x_{\boldsymbol{\ell}}}$ with $a \in \Pi_n^r(b)$, then we have by construction $\forall i \in [\![1,r]\!]$, $P_i\#\gamma = P_i\#\gamma_Z$ and thus γ is a solution.

In the particular case where γ_Z is uniform, if we restrain γ to be also a uniform measure, the problem in this critical case has a combinatorial amount of solutions. Without this restriction, the problem has an infinite amount of solutions, as is discussed in the particular case of Fig. A.I.2.

A.I.3 Consequence for the empirical Sliced Wasserstein Distance

The Sliced Wasserstein (SW) distance between probability measures is frequently used in applied fields such as image processing or machine learning, as an efficient alternative to the Wasserstein distance. It was introduced in [Rab+12] to generate barycentres between images of textures, and it is commonly used nowadays as a loss [Kar+18; DZS18; Wu+19] to train generative networks. This distance writes:

$$\forall \alpha,\beta \in \mathcal{P}_2(\mathbb{R}^d), \quad \mathrm{SW}^2(\alpha,\beta) = \int_{\theta\in\mathbb{S}^d} \mathrm{W}_2^2(P_\theta\#\alpha, P_\theta\#\beta)\mathrm{d}\sigma(\theta),$$

where σ is the uniform distribution over the unit sphere $\mathbb{S}^d$ of $\mathbb{R}^d$, and P_θ denotes the linear projection on the line of direction θ. In its empirical (Monte-Carlo) approximation, used for numerical applications, it becomes:

$$\forall \alpha,\beta \in \mathcal{P}_2(\mathbb{R}^d), \quad \widehat{\mathrm{SW}}_r^2(\alpha,\beta) := \frac{1}{r}\sum_{i=1}^r \mathrm{W}_2^2(P_{\theta_i}\#\alpha, P_{\theta_i}\#\beta). \tag{A.I.3}$$

The main advantage of SW over the usual Wasserstein distance is computational: for two d-dimensional uniform discrete measures with n samples, the empirical estimation with r projections Eq. (A.I.3) can be computed in $\mathcal{O}(rdn + rn\log(n))$ ([Nad21], §2.6), leveraging the fact that the 1D projected Wasserstein distances can be computed by a sorting algorithm. In the same setting, the Wasserstein distance can only be computed in super-quadratic complexity with respect to n [PC19b].

Since W_2 is a distance on $\mathcal{P}_2(\mathbb{R}^d)$ (the space of probability measures over $\mathbb{R}^d$ admitting a finite second-order moment), $\widehat{\mathrm{SW}}_r$ is non-negative, homogeneous and satisfies the triangle inequality. However, $\widehat{\mathrm{SW}}_r$ is only a pseudo-distance since it does not satisfy the separation property: whatever the r directions chosen, it is always possible to find two different distributions α and β such that $\widehat{\mathrm{SW}}_r(\alpha,\beta) = 0$. Now, our previous reconstruction results show that when the r directions are drawn from σ and β is a fixed discrete measure, then $\forall \alpha \in \mathcal{P}_2(\mathbb{R}^d)$, $\widehat{\mathrm{SW}}_p^2(\alpha,\beta) = 0 \implies \alpha = \beta$ almost surely provided that $r > d$. Indeed, $\widehat{\mathrm{SW}}_r(\alpha,\beta) = 0$ if and only if α belongs to the set $\mathcal{S}$ (for $\gamma_Z = \beta$). On the contrary, when the number of projections is too small, the set of discrete measures at distance 0 from a given one is infinite, as stated in the theorem below.

Theorem A.I.4. Let $\gamma_Z := \sum_{\ell=1}^n b_\ell \delta_{z_\ell}$, where the (z_ℓ) are fixed and distinct. Assume $\theta_1,\cdots,\theta_r \sim \sigma^{\otimes r}$.

- if $r \leq d$, there exists σ-a.s. an infinity of measures $\gamma \neq \gamma_Z \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\widehat{\mathrm{SW}}_r(\gamma,\gamma_Z) = 0$.

- if $r > d$, we have ϖ-almost surely $\{\gamma_Z\} = \underset{\gamma \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \widehat{\mathrm{SW}}_r(\gamma, \gamma_Z)$.

In the limit case $r = d$, the distance can be grown by scaling the points of γ_Z further away from the origin. In the case $r < d$, the supports of solution measures can be infinitely far from the support of γ_Z, as illustrated in Fig. A.I.1. To conclude, if $r \leq d$, the information $\widehat{\mathrm{SW}}_r(\gamma, \gamma_Z) \approx 0$ yields no valuable information regarding the closeness of γ and γ_Z. If $r > d$, the information $\widehat{\mathrm{SW}}_r(\gamma, \gamma_Z) = 0$ yields $\gamma = \gamma_Z$ almost-surely with random projections. Due to possible local optima of $Y \longmapsto \widehat{\mathrm{SW}}_r(\gamma_Y, \gamma_Z)$, knowing only $\widehat{\mathrm{SW}}_r(\gamma, \gamma_Z) \approx 0$ may be too weak to conclude that the measures are close even in the favourable case $r > d$, although these considerations are beyond the scope of this work.

A.I.4 Conclusion: Discussion on SW as a Loss in Machine Learning

In Sliced-Wasserstein-based Machine Learning, the question of global optima is paramount since in practice, one must default to a surrogate of SW: be it through stochastic gradient descent (drawing a batch of θ_i at each iteration), or directly through the estimation $\widehat{\mathrm{SW}}_r$. To be precise, generative models such as [DZS18] minimize $\theta \longmapsto \mathrm{SW}(T_\theta \# \mu_0, \mu)$ - or a surrogate thereof - where μ_0 is a low-dimensional input distribution (often chosen as realizations of Gaussian noise), where μ is the target distribution (the discrete dataset), and where T_θ is the model of parameters θ. In this case, the dimension d of the data, which for images can easily exceed one million, can be very large. Our results suggest that performing optimisation with less than d projections is unsound in this case, since it leads to solutions that can be arbitrarily far away from the true data distribution. However, these results do not take into account the intrinsic dimension of the target measure γ_Z, which in the case of images is probably supported on a d'-dimensional manifold with $d' \ll d$. This question has been addressed very recently in [TCD23], which was unknown to us at the time of working on the paper [TFD24b] on which this chapter is based, and provides insightful reconstruction results when γ_Z is known to belong to a d'-dimensional manifold.

Furthermore, it is important to underline that having the guarantee that the global optima are the desired original measure is insufficient in practice. Indeed, the landscape $Y \longmapsto \widehat{\mathrm{SW}}_r(\mu_Y, \mu_Z)$ can present numerous local optima, which can limit the usefulness of SW as a loss function. For practical considerations, this study on global optima could be complemented by an analysis of the aforementioned landscape, which we leave for future work. Namely, one may find a study of the optimization properties of $Y \longmapsto \widehat{\mathrm{SW}}_r(\mu_Y, \mu_Z)$ and $Y \longmapsto \mathrm{SW}(\mu_Y, \mu_Z)$ in Chapter A.II, and a related extension to the study of Stochastic Gradient Descent for SW generative networks in Chapter A.III.

A.II

Properties of Discrete Sliced Wasserstein Losses

A.II.1 Introduction . . . 54
A.II.2 Sliced and Empirical Sliced Wasserstein Energies and their Regularities . . . 57
A.II.2.1 The discrete SW energies $\mathcal{E}$ and $\mathcal{E}_p$ 57
A.II.2.2 Regularity properties of $\mathcal{E}_p$ and $\mathcal{E}$ 58
A.II.2.3 Cell structure of $\mathcal{E}_p$ 60
A.II.2.4 Consequences of the cell structure on the regularity of $\mathcal{E}_p$ and $\mathcal{E}$ 61
A.II.2.5 Convergence of $\mathcal{E}_p$ to $\mathcal{E}$ 62
A.II.2.6 Illustration in a simplified case 64
A.II.3 Properties of the Optimisation Landscapes of $\mathcal{E}$ and $\mathcal{E}_p$ . . . 65
A.II.3.1 Optimising $\mathcal{E}$ 65
A.II.3.2 Optimising $\mathcal{E}_p$ 67
A.II.4 Stochastic Gradient Descent on $\mathcal{E}$ and $\mathcal{E}_p$ . . . 69
A.II.4.1 Theoretical framework 72
A.II.4.2 Convergence of piecewise affine interpolated SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$ 72
A.II.4.3 Convergence of Noised SGD Schemes on $\mathcal{E}$ and $\mathcal{E}_p$ 75
A.II.4.4 Discussion on result generalisation 78
A.II.4.5 A Result for Decreasing Learning Rates 80
A.II.5 Numerical Experiments . . . 80
A.II.5.1 Empirical study of Block Coordinate Descent on $\mathcal{E}_p$ 81
A.II.5.2 Empirical study of SGD on $\mathcal{E}$ and $\mathcal{E}_p$ 82
A.II.6 Conclusion and Outlook . . . 87
A.II.7 Appendix . . . 88
A.II.7.1 Proof of the Central Limit Theorem for Discrete SW 88
A.II.7.2 Computing $\mathcal{E}$, W_2^2 and $\mathcal{E}_p$ in a simple case 89
A.II.7.3 Discrete Wasserstein stability 91
A.II.7.4 Proof of Theorem A.II.7 and convergence rate 93
A.II.7.5 Closed-form expression for Block-Coordinate Descent 99

Abstract

The Sliced Wasserstein (SW) distance has become a popular alternative to the Wasserstein distance for comparing probability measures. Widespread applications include image processing, domain adaptation and generative modelling, where it is common to optimise some parameters in order to minimise SW, which serves as a loss function between discrete probability measures (since measures admitting densities are numerically unattainable). All these optimisation problems bear the same sub-problem, which is minimising the Sliced Wasserstein energy. In this chapter we study the properties of $\mathcal{E} : Y \longmapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, i.e. the SW distance between two uniform discrete measures with the same amount of points as a function of the support $Y \in \mathbb{R}^{n\times d}$ of one of the measures. We investigate the regularity and optimisation

properties of this energy, as well as its Monte-Carlo approximation $\mathcal{E}_p$ (estimating the expectation in SW using only p samples) and show convergence results on the critical points of $\mathcal{E}_p$ to those of $\mathcal{E}$, as well as an almost-sure uniform convergence and a uniform Central Limit result on the process $\mathcal{E}_p$. Finally, we show that in a certain sense, Stochastic Gradient Descent methods minimising $\mathcal{E}$ and $\mathcal{E}_p$ converge towards (Clarke) critical points of these energies. This chapter is based on the paper:

[TFD24a] Eloi Tanguy, Rémi Flamary and Julie Delon.
"Properties of Discrete Sliced Wasserstein Losses".
Mathematics of Computation (Jun. 2024).

A.II.1 Introduction

Optimal Transport (OT) has grown in popularity as a way of lifting a notion of cost between points in a space onto a way of comparing measures on said space. In particular, endowing $\mathbb{R}^d$ with a p-norm yields the Wasserstein distance, which metrises the convergence in law on the space of Radon measures with a finite moment of order p.

The most studied object that arises from this theory is perhaps the 2-Wasserstein distance, which is defined as follows (see [PC19b; San15; Vil09] for a complete practical and theoretical presentation):

$$\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{W}_2^2(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_1 - x_2\|^2 \mathrm{d}\pi(x_1,x_2), \tag{A.II.1}$$

where $\Pi(\mu,\nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ of first marginal μ and second marginal ν. We denote $\mathcal{P}_2(\mathbb{R}^d)$ as the set of probability measures on $\mathbb{R}^d$ admitting a second-order moment.

The 1 and 2-Wasserstein distances are commonly used for generation tasks, formulated as probability density fitting problems. One defines a statistical model μ_θ, a probability measure which is designed to approach a target data distribution μ. A typical way of solving this problem is to minimise in θ the distance between μ_θ and μ: one may choose any probability discrepancies (Kullback-Leibler, Csiszár divergences, f-divergences or Maximum Mean Discrepancy), or alternatively the Wasserstein Distance. In the case of Generative Adversarial Networks, the so-called "Wasserstein GAN" [ACB17; Gul+17] draws its formulation from the dual expression of the 1-Wasserstein distance.

Unfortunately, computing the Wasserstein distance is prohibitively costly in practice. The discrete formulation of the Wasserstein distance (the Kantorovich linear problem) is typically solved approximately using standard linear programming tools. These methods suffer from a super-cubic worst-case complexity with respect to the number of samples from the two measures. Furthermore, given n samples from each measure μ and ν, the convergence of the estimated distance $\mathrm{W}_2(\hat{\mu}_n, \hat{\nu}_n)$ is only in $\mathcal{O}(n^{-1/d})$ towards the true distance, thus OT suffers from the curse of dimensionality, as is known since Dudley, 1969 [Dud69].

Several efforts have been made in recent years to make Optimal Transport more accessible computationally. In particular, many surrogates for W_2 have been proposed, perhaps the most notable of which is the Sinkhorn Divergence (see [PC19b; Cut13; GPC18]). The Sinkhorn Divergence adds entropic regularisation to OT, yielding a strongly convex algorithm which can be solved efficiently.

Another alternative is the Sliced Wasserstein (SW) Distance, which leverages the simplicity of computing the Wasserstein distance between one-dimensional measures. Indeed, given

$$\gamma_X := \frac{1}{n}\sum_{k=1}^n \delta_{x_k},\ \gamma_Y := \frac{1}{n}\sum_{k=1}^n \delta_{y_k} \text{ with } x_1, \cdots, x_n, y_1, \cdots, y_n \in \mathbb{R},$$

the 2-Wasserstein distance between these two measures can be computed by sorting their supports:

$$\mathrm{W}_2^2(\gamma_X, \gamma_Y) = \frac{1}{n}\sum_{k=1}^n (x_{\sigma(k)} - y_{\tau(k)})^2, \tag{A.II.2}$$

where σ is a permutation sorting $(x_1, \cdots, x_n)$, and τ is a permutation sorting $(y_1, \cdots, y_n)$.

The idea of the Sliced Wasserstein Distance [Rab+12] is to compute the 1D Wasserstein distances between projections of input measures. We write $P_\theta : \mathbb{R}^d \longrightarrow \mathbb{R}$ the map $x \longmapsto \theta^\top x$, and $\mathfrak{o}$ the uniform measure over the euclidean unit sphere of $\mathbb{R}^d$, $\mathbb{S}^{d-1}$. Denoting $\#$ the push-forward operation[1], the Sliced Wasserstein distance between two measures μ and ν is defined as

$$\mathrm{SW}_2^2(\mu, \nu) := \int_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \mu, P_\theta \# \nu) \mathrm{d}\mathfrak{o}(\theta). \tag{A.II.3}$$

Similarly, for $q \geq 1$, the q-Sliced Wasserstein distance SW_q^q (to the power q) is obtained by replacing W_2^2 in the previous equation by the q Wasserstein distance (to the power q) W_q^q.

SW has enjoyed a substantial amount of theoretical study, albeit not as extensively as for the original Wasserstein distance. For measures supported on a fixed compact of $\mathbb{R}^d$, [Bon13, Chapter 5] has shown that the Wasserstein and Sliced Wasserstein distances are equivalent. The same work also developed a theory of gradient flows for SW, which justifies some generative methods. Further discussion on this equivalence has been performed by Bayraktar and Guo [BG21]. Nadjahi et al. [Nad+19] showed that SW metrises the convergence in law (without restrictions of the measure supports), and further concluded guarantees for SW-based generative models.

Continuous measures being out of the reach of practical computation, it is necessary to perform sample estimation and replace them with discrete empirical estimates. Thankfully, as shown in [Nad+20b], the *sample complexity* (i.e. the rate of convergence of the estimates w.r.t. the number of samples) for sliced distances such as SW is in $1/\sqrt{n}$, which in particular avoids the curse of dimensionality from which the Wasserstein Distance suffers. This fuels interest for the study of $Y \mapsto \mathrm{SW}(\gamma_Y, \gamma)$, which is to say the variation of SW w.r.t. the discrete support of one of the measures. It is currently unknown whether this functional presents strict local optima, for instance.

Originally, SW was introduced as a more computable alternative to the Wasserstein distance, notably for texture mixing using a barycentric formulation [Rab+12; Bon+15a]. Other uses of SW have been suggested, notably in statistics as a probability discrepancy. For instance, Nadjahi et al. [Nad+20a] proposed an approximate bayesian computation method, where the estimation of the posterior parameters is done by selecting those under which the SW distance between observed and synthetic data is below a fixed threshold. Other widespread uses of SW in image processing include colour transfer [AGD19] and colour harmonisation [Bon+15b].

Nowadays, SW is commonly used as a training or validation loss in generative Machine Learning. Karras et al. [Kar+18] propose to use SW to compare GAN results, by comparing images via multi-scale patched descriptors. Some generative models (including GANs and auto-encoders), leverage the computational advantages of SW in order to learn a target distribution. This is done under the implicit generative modelling framework, where a network T_u of parameters u is learned such as to minimise $\mathrm{SW}(T_u \# \mu_0, \mu)$, where μ_0 is a low-dimensional input distribution (often chosen as Gaussian or uniform noise), and where μ is the target distribution. Deshpande et al. [DZS18] and Wu et al. [Wu+19] train GANs and auto-encoders within this framework; Liutkus et al. [Liu+19] perform generation by minimising a regularised SW problem, which they solve by gradient flow using an SDE formulation. SW can be used to synthesise images by minimising the SW distance between features of the optimised image and a target image, as done by [Hei+21] for textures with neural features, and by [TPG16] with wavelet features (amongst other methods).

In practice, the integration over the unit sphere in SW is intractable, and one must resort to a Monte-Carlo approximation, taking the average between p projections instead of the expectation, usually during iterations of a Stochastic Gradient Descent [Kol+19b]. This implies that for a finite number of iterations, a fixed number of projections p, potentially very small compared to what is needed to explore the hypersphere, is explored in practice. The question of this finite number of final projection directions is made even more important by the fact that practitioners usually optimise the expectation of the SW distance on large mini-batches [DZS18] that also

[1] The push-forward of a measure μ on $\mathbb{R}^d$ by an application $T : \mathbb{R}^d \to \mathbb{R}^k$ is defined as a measure $T\#\mu$ on $\mathbb{R}^k$ such that for all Borel sets $B \in \mathcal{B}(\mathbb{R}^k), T\#\mu(B) = \mu(T^{-1}(B))$.

limits the total number of effective projections p. The estimation error of this approximation has not been extensively studied, and it is common in practice to assume that this empirical version presents the same properties as the true SW distance.

An important question is the conditions under which these approximations for SW are valid. In practice, sliced-Wasserstein Generative Models compute SW in the data space or in the data encoding space ([Kol+19b; DZS18]), which yields high values for the dimension d, in particular for images. Note that the necessity behind having a large number of projections p was already hinted at in [Kol+19b, Section 3.3]. Another untreated question is the complexity of optimising this approximation of SW, and how this optimisation landscape compares to the true SW landscape.

Bonneel et al. [Bon+15a] studied the uses of SW for barycentre computation, and in particular proved that the empirical SW distance is $\mathcal{C}^1$ on a certain open set, with respect to the measure positions. They remarked that in practice, numerical resolutions for discretised SW distances converged towards (eventual) local optima, however the convergence and local optima have not been studied theoretically.

In this chapter, we propose to study $\mathcal{E} : Y \mapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, where γ_Y and γ_Z are two uniform discrete measures supported by n points, denoted by Y and Z. Our main objective is to provide optimisation properties for the landscapes of $\mathcal{E}$ and its Monte-Carlo counterpart $\mathcal{E}_p$, obtained by replacing the expectation by an average over p projections. In Section A.II.2, we prove several regularity properties for both energies, such as semi-concavity, and we show that the convergence of the Monte-Carlo estimation is uniform (on every compact) w.r.t. the measure locations. Section A.II.3 focuses on the respective landscapes of $\mathcal{E}$ and $\mathcal{E}_p$, and shows that the critical points of $\mathcal{E}$ satisfy a fixed-point equation, and how the critical points of $\mathcal{E}_p$ relate to this fixed-point equation when the number of projections p increases (with convergence rates). Mérigot et al. follow a similar methodology in [MSS21], where they study optimisation properties for $Y \longmapsto \mathrm{W}_2(\gamma_Y, \mu)$, with μ a continuous measure. The main difficulty they face arises from the non-convexity of the map, and this difficulty is also central in our work. The last two sections of our chapter tackle numerical considerations. To begin with, since $\mathcal{E}$ and $\mathcal{E}_p$ are usually minimised in the literature using Stochastic Gradient Descent (SGD), we provide in Section A.II.4 the first complete convergence study of SGD for $\mathcal{E}$ and $\mathcal{E}_p$, relying on the recent works [BHS22] and [Dav+20]. Finally, Section A.II.5 challenges our theoretical results with extensive numerical experiments, quantifying the impact of the dimension and several other parameters on the convergence.

Notations

- d is the dimension, n is the number of points
- p is the number of projections $(\theta_1, \cdots, \theta_p)$
- $\|\cdot\|_2$: Euclidean norm of $\mathbb{R}^n$
- Matrices $X \in \mathbb{R}^{n\times d}$ are written $X = (x_1, \cdots, x_n)^\top$ with the $x_i \in \mathbb{R}^d$
- $\|Y\|_{\infty,2}$ for $Y \in \mathbb{R}^{n\times d}$ denotes $\max_{i\in[\![1,n]\!]} \|Y_{i,\cdot}\|_2 = \max_{i\in[\![1,n]\!]} \|y_i\|_2$
- $M \cdot N$: inner product $\mathrm{Trace}(M^\top N)$ for matrices
- W_2: 2-Wasserstein Distance Eq. (A.II.1)
- σ: Uniform measure on the unit sphere $\mathbb{S}^{d-1}$ of $\mathbb{R}^d$
- P_θ: for $\theta \in \mathbb{S}^{d-1}$, $P_\theta = x \longmapsto \theta^\top x$
- SW_2: Sliced 2-Wasserstein distance Eq. (A.II.3)
- SW_q: Sliced q-Wasserstein distance
- γ_X: for $X \in \mathbb{R}^{n\times d}$: discrete measure $\frac{1}{n}\sum_i \delta_{X_{i,\cdot}} = \frac{1}{n}\sum_i \delta_{x_i}$

- $\mathcal{E}(Y)$: $\mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$ Eq. (A.II.4)
- $\mathcal{E}_p(Y)$: Monte-Carlo approximation of $\mathcal{E}(Y)$ with p projections Eq. (A.II.5)
- $\triangle_n$: n-simplex: $a \in (0, +\infty)^n$ such that $\sum_i a_i = 1$
- $\|M\|_F$: Frobenius norm: $\sqrt{\sum_{i,j} M_{i,j}^2}$
- $\mathbf{m}$ denotes p permutations $(\sigma_1, \cdots, \sigma_p)$ of $[\![1, n]\!]$, see Section A.II.2.3
- $\mathcal{C}_{\mathbf{m}}$: cell of configuration $\mathbf{m}$, see Section A.II.2.3

A.II.2 Sliced and Empirical Sliced Wasserstein Energies and their Regularities

A.II.2.1 The discrete SW energies $\mathcal{E}$ and $\mathcal{E}_p$

The Sliced Wasserstein distance has been widely studied as an alternative to the Wasserstein distance, in particular it is arguably simpler to compute in order to minimise measure discrepancies. In practice, one may not work with continuous measures, which are beyond the capabilities of numerical approximations, thus one must sometimes contend with discrete measures. To that end, we study in this chapter the SW distance between discrete measures, and in particular the associated energy landscape with respect to the support of one of the measures:

$$\mathcal{E} := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ Y & \longmapsto & \displaystyle\int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z) \mathrm{d}\sigma(\theta) \end{array} \right. , \tag{A.II.4}$$

where n denotes the number of points in the data matrices Y, Z, which we write as data entries stacked vertically: $Y = (y_1, \cdots, y_n)^\top$, with points in $\mathbb{R}^d$. The associated (uniform) discrete measure supported on $\{y_1, \cdots, y_n\}$ will be denoted $\gamma_Y := \frac{1}{n}\sum_k \delta_{y_k}$.

For instance, this framework encompasses SW-based implicit generative models ([DZS18], [Wu+19]), which optimise parameters ρ by minimising $\mathrm{SW}(T_\rho \# \mu_0, \mu)$, where μ_0 is comprised of samples of a simple distribution, and μ corresponds to data samples which we would like to generate. In this setting, one would need to minimise *through* $\mathcal{E}$. The use of discrete measures is also backed theoretically by the study of the *sample complexity* of SW [Nad+20b], which is to say the rate of decrease of the approximation error between $\mathrm{SW}(\mu, \nu)$ and its discretised counterpart $\mathrm{SW}(\hat{\mu}_n, \hat{\nu}_n)$.

In practical and realistic settings, the only numerically accessible workaround to optimise through $\mathcal{E}$ is a form of discretisation of the set of directions. The first and most common method, due to its efficiency and simplicity, is to minimize $\mathcal{E}$ through stochastic gradient descent (SGD): at each time set t, p random directions $(\theta_1^{(t)}, \cdots, \theta_p^{(t)})$ are drawn, and a gradient descent step is performed by approximating $\mathcal{E}$ by a discrete sum on these p random directions. This method is optimisation-centric, since it does not concern itself with computing the final SW distance and focuses on optimising the parameters. A second possible discretisation method consists in fixing the p directions $(\theta_1, \cdots, \theta_p)$ once for all and replacing $\mathcal{E}$ in the minimization by its Monte-Carlo estimator [2]

$$\mathcal{E}_p := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ Y & \longmapsto & \displaystyle\frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(P_{\theta_i} \# \gamma_Y, P_{\theta_i} \# \gamma_Z) \end{array} \right. . \tag{A.II.5}$$

It is important to note that both methods are intuitively tied, since in both cases there is a finite amount of sampled directions. If the SGD method lasts T iterations with p projections every time, it amounts to a specific way of optimising $\mathcal{E}_{pT}$. For this reason, studying $\mathcal{E}_p$ theoretically is not only interesting in itself as an approximation of $\mathcal{E}$, but also yields a better insight on the SGD strategy.

[2] In this notation the projection axes $\theta_1, \cdots, \theta_p \in \mathbb{S}^{d-1}$ are written implicitly, the complete notation being $\mathcal{E}_p(Y; (\theta_i)_{i\in[\![1,p]\!]})$ when required.

The study of $\mathcal{E}$ is also tied with the study of the SW barycentres, which solve the optimisation problem

$$\mathrm{Bar}(\lambda_j, \gamma_{Z^{(j)}})_{j\in[\![1,J]\!]} = \operatorname*{argmin}_{Y\in\mathbb{R}^{n\times d}} \sum_{j=1}^{J} \lambda_j \mathcal{E}(Y, Z^{(j)}) =: \mathcal{E}_{\mathrm{bar}}(Y), \tag{A.II.6}$$

where the notation $\mathcal{E}(Y, Z^{(j)})$ reflects the dependency on Z in the definition of $\mathcal{E}$ Eq. (A.II.4). The regularity and convergence results will immediately be applicable to the barycentre energy Eq. (A.II.6). While the optimisation results on $\mathcal{E}$ and $\mathcal{E}_p$ will not generalise naturally due to the sum, the SGD convergence results shall still hold.

As a Monte-Carlo estimator, the law of large numbers yields the point-wise convergence of $\mathcal{E}_p$ to $\mathcal{E}$ if the $(\theta_i)_{i\in\mathbb{N}}$ are i.i.d. of law $\mathfrak{o}$:

$$\mathcal{E}_p(Y; (\theta_i)_{i\in[\![1,p]\!]}) \xrightarrow[p\to+\infty]{\text{a.s.}} \mathcal{E}(Y). \tag{A.II.7}$$

For this reason, it is often assumed that numerically, $\mathcal{E}_p$ and $\mathcal{E}$ will behave similarly, which is perhaps why research has been scarce on the landscape of $\mathcal{E}_p$, the focus remaining on the theoretical properties of the true or mini-batch Sliced Wasserstein Distance [Nad+19; Nad+20a]. But as discussed in the introduction, practitioners often optimize the SW distance using SGD with a finite number of projection directions [Kol+19b; DZS18], and the landscape of $\mathcal{E}_p$ is of paramount importance. This section and the next one are dedicated to studying the relations and differences between $\mathcal{E}_p$ and $\mathcal{E}$.

Remark A.II.1. Some of our results can in fact be extended to q-SW instead of 2-SW, especially regularity results Lemma A.II.1, Proposition A.II.1 and Theorem A.II.1, as well as the statistical estimation results Theorem A.II.3 and Theorem A.II.4. However, as soon as we need the *cell structure* of $\mathcal{E}_p$ (Section A.II.2.3), we leverage the simplicity of the quadratic case $q = 2$.

A.II.2.2 Regularity properties of $\mathcal{E}_p$ and $\mathcal{E}$

In order to study the regularity of our energies, we first focus on the regularity of w_θ, the 2-Wasserstein distance between two discrete measures projected on the line $\mathbb{R}\theta$:

$$w_\theta := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R} \\ Y & \longmapsto & \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z) \end{array} \right. . \tag{A.II.8}$$

With this notation, observe that $\mathcal{E}$ and $\mathcal{E}_p$ can be written

$$\mathcal{E}(Y) = \mathbb{E}_{\theta\sim\mathfrak{o}}\left[w_\theta(Y)\right] \quad \text{and} \quad \mathcal{E}_p(Y) = \mathbb{E}_{\theta\sim\mathfrak{o}_p}\left[w_\theta(Y)\right], \tag{A.II.9}$$

where $\mathfrak{o}_p := \dfrac{1}{p}\sum_{i=1}^{p} \delta_{\theta_i}$ for p fixed directions $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$.

We now provide an important regularity result about the uniformly locally Lipschitz property of the functions $(w_\theta)_\theta$, which will yield easily that our energies $\mathcal{E}$ and $\mathcal{E}_p$ are also locally Lipschitz, a central property in the convergence study of particular SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$ (see Section A.II.4.2). To show this result on (w_θ), we need the following Lemma A.II.1, whose proof is provided in Section A.II.7.3. This result shows that the Wasserstein cost is regular in some sense with respect to the measure weights and the cost matrix, which will be helpful when studying the regularity of the functions w_θ.

Lemma A.II.1 (Stability of the Wasserstein cost)**.** Let $\alpha, \overline{\alpha}, \in \triangle_n$, $\beta, \overline{\beta} \in \triangle_m$ and $C, \overline{C} \in \mathbb{R}_+^{n\times m}$. Denote by $\mathrm{W}(\alpha, \beta; C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C$ the cost of the discrete Kantorovich problem of cost matrix C between the weights α, β. We have the following Lipschitz-like

inequalities, assuming $\alpha, \overline{\alpha}, \beta, \overline{\beta} > 0$ entry-wise:

$$\left|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};\overline{C})\right| \leq \|C-\overline{C}\|_\infty + \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1), \tag{A.II.10}$$

$$\left|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};\overline{C})\right| \leq \|C-\overline{C}\|_F + \|C\|_F\left(\|\alpha-\overline{\alpha}\|_2 + \|\beta-\overline{\beta}\|_2\right). \tag{A.II.11}$$

Remark A.II.2. Using Eq. (A.II.11) twice with $(C,\overline{C})$ and $(\overline{C},C)$ yields a symmetric second term with a factor $\min(\|C\|_\infty, \|\overline{C}\|_\infty)$ instead of $\|C\|_\infty$, and likewise for $\|\cdot\|_F$ with Eq. (A.II.11).

Remark A.II.3. The result of Lemma A.II.1 assumes *positive* weights, but in the case of the q-Wasserstein cost $C_{i,j} = \|x_i - y_j\|_2^q$ with $q \geq 1$, we can remove this assumption by a continuity argument, since the q-Wasserstein distance metrises the weak convergence of measures (see [San15, Theorem 5.10 or 5.11], applied to the simple case of discrete measures for which convergence of moments is immediate).

The following regularity property on (w_θ) uses the norm $\|X\|_{\infty,2} = \max_{k\in[\![1,n]\!]} \|x_k\|_2$ on $\mathbb{R}^{n\times d}$. We also denote $D := n \times d$ for convenience.

Proposition A.II.1. The $(w_\theta)_{\theta\in\mathbb{S}^{d-1}}$ are uniformly locally Lipschitz.. More precisely, in a neighbourhood $X \in \mathbb{R}^D$ or radius $r > 0$, writing $\kappa_r(X) := 2n(r + \|X\|_{\infty,2} + \|Z\|_{\infty,2})$, each w_θ is $\kappa_r(X)$ Lipschitz, which is to say

$$\forall X \in \mathbb{R}^D,\ \forall Y, Y' \in B_{\|\cdot\|_{\infty,2}}(X,r),\ \forall \theta \in \mathbb{S}^{d-1},\ |w_\theta(Y) - w_\theta(Y')| \leq \kappa_r(X)\|Y - Y'\|_{\infty,2}.$$

Proof. Let $X \in \mathbb{R}^D$, $Y, Y' \in B_{\|\cdot\|_{\infty,2}}(X,r)$, and $\theta \in \mathbb{S}^{d-1}$. By Lemma A.II.1 Equation Eq. (A.II.11), we have $|w_\theta(Y) - w_\theta(Y')| \leq \|C - C'\|_F$, where for $(k,l) \in [\![1,n]\!]^2$, $C_{k,l} := (\theta^\top y_k - \theta^\top z_l)^2$, likewise for C'. Then:

$$\begin{aligned}
[C-C']_{k,l} &= \left(\theta^\top(y_k - y'_k)\right)\left(\theta^\top(y_k + y'_k - 2z_l)\right) \\
&\leq \|y_k - y'_k\|_2 \|y_k + y'_k - 2z_l\|_2 \\
&= \|y_k - y'_k\|_2 \|y_k - x_k + y'_k - x_k + 2z_l + 2x_k\|_2 \\
&\leq \|y_k - y'_k\|_2 \left(2r + 2\|Z\|_{\infty,2} + 2\|X\|_{\infty,2}\right).
\end{aligned}$$

Finally, $\|C - C'\|_F = \sqrt{\sum_{k,l\in[\![1,n]\!]} [C-C']_{k,l}^2} \leq 2n(r + \|X\|_{\infty,2} + \|Z\|_{\infty,2})\|Y - Y'\|_{\infty,2}$. □

As a consequence, we deduce immediately that $\mathcal{E}_p$ and $\mathcal{E}$ are locally Lipschitz.

Theorem A.II.1. $\mathcal{E}$ and $\mathcal{E}_p$ are locally Lipschitz.

Proof. Let $X \in \mathbb{R}^D$, $r > 0$ and $\mu \in \{\sigma, \sigma_p\}$. By Proposition A.II.1, for any $Y, Y' \in B_{\|\cdot\|_{2,\infty}}(X,r)$,

$$\left|\mathbb{E}_{\theta\sim\mu}\left[w_\theta(Y)\right] - \mathbb{E}_{\theta\sim\mu}\left[w_\theta(Y')\right]\right| \leq \mathbb{E}_{\theta\sim\mu}\left[\left|w_\theta(Y) - w_\theta(Y')\right|\right] \leq \kappa_r(X)\|Y - Y'\|_{\infty,2}.$$

□

As a locally Lipschitz function, $\mathcal{E}$ is differentiable almost everywhere. The expression of its gradient is quite simple and corresponds to the simple differentiation of w_θ in the integral, as was shown in [Bon+15a]. We remind here their result for the sake of completeness, and because

the derivative will be useful on several occasions in this chapter. We define $\mathcal{U}$ the open set of matrices with distinct lines

$$\mathcal{U} = \left\{(x_1, \cdots, x_n)^\top \in \mathbb{R}^{n\times d} \mid \forall i \neq j,\ [\![1,n]\!]^2,\ x_i \neq x_j\right\}. \tag{A.II.12}$$

Theorem A.II.2. Regularity of $\mathcal{E}$ [Bon+15a, Theorem 1] $\mathcal{E}$ is continuous on $\mathbb{R}^{n\times d}$, and of class $\mathcal{C}^1$ on $\mathcal{U}$. There exists $\kappa \geq 1$ such that $\nabla\mathcal{E}$ is κ-Lipschitz on $\mathcal{U}$. For $Y \in \mathcal{U}$, one has the expression:

$$\frac{\partial \mathcal{E}}{\partial y_k}(Y) = \frac{2}{n}\int_{\mathbb{S}^{d-1}} \theta\theta^\top (y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)}) \mathrm{d}\varpi(\theta), \tag{A.II.13}$$

where for $\theta \in \mathbb{S}^{d-1}, X \in \mathcal{U}$, $\tau_X^\theta \in \mathfrak{S}_n$ is any permutations s.t. $\theta^\top x_{\tau_X^\theta(1)} \leq \cdots \leq \theta^\top x_{\tau_X^\theta(n)}$.

Proving this theorem requires to be cautious. Firstly, differentiating directly under the integral using standard calculus theorems is impossible, since the integrand is only differentiable on a set $\mathcal{U}_\theta$ which depends on the integration variable θ. Fortunately, these irregularities are smoothed out as θ rotates, yielding differentiability almost-everywhere. Secondly, the problematic term τ_Y^θ can be dealt with for $Y \in \mathcal{U}$ by remarking that for any Y' ε-close to Y, we have $\tau_Y^\theta = \tau_{Y'}^\theta$ for every θ in a certain subset of $\mathbb{S}^{d-1}$ which is of ϖ-measure exceeding $1 - C\varepsilon$. Regarding the multiplicative constant, Theorem 1 in Bonneel et. al omits the $1/n$ factor (we believe that this is a typing error).

A.II.2.3 Cell structure of $\mathcal{E}_p$

In order to further study the optimisation properties of $\mathcal{E}_p$ and $\mathcal{E}$, we need to exhibit more explicitly the structure of the landscape of $\mathcal{E}_p$. The semi-concavity of $\mathcal{E}_p$ and $\mathcal{E}$ will follow, as well as the fact that $\mathcal{E}$ is semi-algebraic (see Proposition A.II.5). We can compute $\mathcal{E}_p$ by leveraging the formula for 1D Wasserstein distances:

$$\forall Y \in \mathbb{R}^{n\times d}, \quad \mathcal{E}_p(Y) = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}\left(\theta_i^\top\left(y_k - z_{\tau_Z^{\theta_i}\circ(\tau_Y^{\theta_i})^{-1}(k)}\right)\right)^2. \tag{A.II.14}$$

For now we consider Z and the (θ_i) fixed, and we write $\mathbf{m}(Y) := (\mathbf{m}_i(Y))_{i\in[\![1,p]\!]}$ where $\mathbf{m}_i(Y) = \tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1}$. Writing $\mathfrak{S}_n$ the set of permutations of $\{1, \cdots, n\}$, $\mathbf{m}_i$ is the element σ of $\mathfrak{S}_n$ which solves the (Monge) quadratic optimal transport between the points $(\theta_i^\top y_1, \cdots, \theta_i^\top y_n)$ and $(\theta_i^\top z_1, \cdots, \theta_i^\top z_n)$. The matching configuration $\mathbf{m}(Y)$ depends implicitly on the fixed directions (θ_i).

Note that the permutations τ_Y^θ and τ_Z^θ are not always uniquely defined: for any $\theta \in \mathbb{S}^{d-1}$, there exists $Y \in \mathcal{U}$ such that τ_Y^θ is not uniquely defined (take Y such that $\theta \in (y_1 - y_2)^\perp$ for instance). However, for a given set of directions (θ_i), these permutations are uniquely defined almost everywhere on $\mathbb{R}^{n\times d}$.

A set of interest is $\mathcal{C}_\mathbf{m} = \{Y \in \mathcal{U} \mid \mathbf{m}(Y)$ is uniquely defined and equal to $\mathbf{m}\}$, the cell of points Y of configuration $\mathbf{m}$. Writing $\mathbf{m} = (\mathbf{m}_1, \cdots, \mathbf{m}_p)$, and using the optimality of each $\mathbf{m}_i$, note that each cell $\mathcal{C}_\mathbf{m}$ can be also written as

$$\begin{aligned}\mathcal{C}_\mathbf{m} = \Bigg\{ & Y \in \mathbb{R}^{n\times d} : \forall i \in [\![1,p]\!],\ \forall \sigma \in \mathfrak{S}_n \setminus \{\mathbf{m}_i\}, \\ & \sum_{k=1}^{n} z_{\mathbf{m}_i(k)}^\top \theta_i\theta_i^\top y_k > \sum_{k=1}^{n} z_{\sigma(k)}^\top \theta_i\theta_i^\top y_k\Bigg\}.\end{aligned} \tag{A.II.15}$$

Thus, each $\mathcal{C}_\mathbf{m}$ is an open polyhedral cone, obtained as the intersection of $p(n! - 1)$ half-open planes. Moreover, the union of these cells $\bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \mathcal{C}_\mathbf{m}$ is a strict subset of $\mathcal{U}$ (as a consequence of the non uniqueness of the permutations for some Y), but is dense in $\mathbb{R}^{n\times d}$. $\mathbb{R}^{n\times d}$. $Y \in \mathcal{U}$ such

that τ_Y^θ is not uniquely defined (take Y such that $\theta \in (y_1 - y_2)^\perp$ for instance). These cells are of particular interest since by definition, $\mathcal{E}_p$ is quadratic on each $\mathcal{C}_\mathbf{m}$, and can be written

$$\forall Y \in \mathcal{C}_\mathbf{m},\ \mathcal{E}_p(Y) = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}\left(\theta_i^\top(y_k - z_{\mathbf{m}_i(k)})\right)^2 =: q_\mathbf{m}(Y). \tag{A.II.16}$$

Furthermore, the sorting interpretation of the 1D Wasserstein distance allows us to re-write $\mathcal{E}_p(Y)$ as a minimum of quadratics,

$$\forall Y \in \mathbb{R}^{n\times d},\ \mathcal{E}_p(Y) = \min_{\mathbf{m}\in\mathfrak{S}_n^p} q_\mathbf{m}(Y) = q_{\mathbf{m}(Y)}(Y). \tag{A.II.17}$$

Remark A.II.4. To each $Y = (y_1, \cdots, y_n)^\top$ (seen as a $n \times d$ matrix), we can associate the column vector $\mathrm{vec}(y) := (y_1^\top, \cdots, y_n^\top)^\top$, which is now a vector of $\mathbb{R}^D = \mathbb{R}^{n\times d}$ without any abuse of notation. We re-write the quadratic from equation Eq. (A.II.16) in standard form: $q_\mathbf{m}(\mathrm{vec}(y)) = \frac{1}{2}\mathrm{vec}(y)^\top B\mathrm{vec}(y) - a_\mathbf{m}^\top\mathrm{vec}(y) + b$, where:

$$\begin{aligned} B &:= \frac{2}{n}\begin{pmatrix} A & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & A \end{pmatrix};\ A := \frac{1}{p}\sum_{i=1}^{p}\theta_i\theta_i^\top; \\ a_\mathbf{m} &:= \frac{2}{pn}\begin{pmatrix} \sum_{i=1}^{p}\theta_i\theta_i^\top z_{\mathbf{m}_i(1)} \\ \vdots \\ \sum_{i=1}^{p}\theta_i\theta_i^\top z_{\mathbf{m}_i(n)} \end{pmatrix};\ b := \frac{1}{n}\sum_{k=1}^{n} z_k^\top A z_k. \end{aligned} \tag{A.II.18}$$

Note in particular that only the linear component depends on $\mathbf{m}$.

Finding the minimum of each quadratic $q_\mathbf{m}$ can be done in closed form, thanks to the computations of Remark A.II.4. This computational accessibility will be leveraged during our discussions on minimising $Y \longmapsto \mathcal{E}_p(Y)$ (Section A.II.3.2.4), wherein we shall present the Block Coordinate Descent method (Algorithm A.II.1), which computes iteratively minima of quadratics in closed form.

A.II.2.4 Consequences of the cell structure on the regularity of $\mathcal{E}_p$ and $\mathcal{E}$

The cell decomposition presented in Section A.II.2.3 permits to show several additional regularity results.

Proposition A.II.2. $\mathcal{E}_p$ is quadratic on each cell $\mathcal{C}_\mathbf{m}$, thus of class $\mathcal{C}^\infty$ on $\bigcup_{\mathbf{m}\in\mathfrak{S}_n^p}\mathcal{C}_\mathbf{m}$, hence $\mathcal{C}^\infty$ a.e..

The formulation as an infimum of quadratics also allows us to prove that $\mathcal{E}_p$ is semi-concave, which is an extremely useful property for optimisation.

Proposition A.II.3. $\mathcal{E}_p$ is $\frac{1}{n}$-semi-concave, i.e. $\mathcal{E}_p - \frac{1}{n}\|\cdot\|_2^2$ is concave.

Proof. Using the notations from Remark A.II.4, $\mathcal{E}_p(\mathrm{vec}(y)) = \frac{1}{2}\mathrm{vec}(y)^\top B\mathrm{vec}(y) + \min_{\mathbf{m}\in\mathfrak{S}_n^p} a_\mathbf{m}^\top\mathrm{vec}(y) + b$. Now, $\mathrm{vec}(y) \longmapsto \min_{\mathbf{m}\in\mathfrak{S}_n^p} a_\mathbf{m}^\top\mathrm{vec}(y) + b$ is concave, as an infimum of affine functions. Furthermore

$$\frac{1}{2}\mathrm{vec}(y)^\top B\mathrm{vec}(y) - \frac{1}{n}\|\mathrm{vec}(y)\|_2^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^\top(A - I)y_k,$$

and since $A \preceq I_d$, the equation above defines a concave function of $\mathrm{vec}(y)$. □

The semi-concavity of $\mathcal{E}_p$ and point-wise convergence allows us to deduce the semi-concavity of $\mathcal{E}$:

Proposition A.II.4. $\mathcal{E}$ is $\frac{1}{n}$-semi-concave.

Proof. By Proposition A.II.3, $\forall p \in \mathbb{N}^*$, $\mathcal{E}_p$ is $\frac{1}{n}$-semi-concave. Let $p \in \mathbb{N}^*$, $Y, Y' \in \mathbb{R}^{n\times d}$ and $\lambda \in [0,1]$. We have

$$\begin{aligned}&\mathcal{E}_p((1-\lambda)Y + \lambda Y') - \frac{1}{n}\|(1-\lambda)Y + \lambda Y'\|_F^2 \\ &\geq (1-\lambda)\mathcal{E}_p(Y) + \lambda\mathcal{E}_p(Y') - \frac{1}{n}\left((1-\lambda)\|Y\|_F^2 + \lambda\|Y'\|_F^2\right).\end{aligned}$$

Taking the limit $p \longrightarrow +\infty$ in this inequality yields the $\frac{1}{n}$-semi-concavity of $\mathcal{E}$. □

The cell formulation also allows us to show that $\mathcal{E}_p$ is semi-algebraic, which means that it can be written using a finite number of polynomial expressions. This result induces strong optimisation results akin to semi-concavity for our purposes. We recall the definition of a semi-algebraic set ([Wak08, Definition 1]). $S \subset \mathbb{R}^D$ is *semi-algebraic* if it can be written $S = \cup_{n=1}^{M} \cap_{m=1}^{M} A_{n,m}$ where $(A_{n,m})$ is a finite family of sets such that $A_{n,m} = \{X \in \mathbb{R}^D \mid p_{n,m}(X) \geq 0\}$ or $A_{n,m} = \{X \in \mathbb{R}^D \mid p_{n,m}(X) = 0\}$, with $p_{n,m}$ being D-variate polynomials with real coefficients. A *semi-algebraic* function is a function whose graph is a semi-algebraic set.

Proposition A.II.5. $\mathcal{E}_p$ is semi-algebraic.

Proof. We shall prove that the set $G := \{(X, \mathcal{E}_p(X)) \mid X \in \mathcal{U}_\Theta\}$ is semi-algebraic, where $\mathcal{U}_\Theta := \bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \mathcal{C}_\mathbf{m}$. Observe that

$$G = \bigcup_{\mathbf{m}\in\mathfrak{S}_n^p} \left\{(X,y) \in \mathbb{R}^{D+1}, X \in \mathcal{C}_\mathbf{m} \text{ and } y = q_\mathbf{m}(X)\right\}.$$

The function $q_\mathbf{m}$ is quadratic, thus polynomial, and the cells $\mathcal{C}_\mathbf{m}$ are intersections of a finite number of half planes, so we conclude that G is semi-algebraic.

The closure of $\mathcal{U}_\Theta$ verifies $\overline{\mathcal{U}_\Theta} = \mathbb{R}^D$, furthermore, since $\mathcal{E}_p$ is continuous on $\mathbb{R}^D$ (by Theorem A.II.2), the closure of G is exactly the graph of $\mathcal{E}_p$. Now by [Wak08, Lemma 4], since G is semi-algebraic, then $\overline{G}$ is also semi-algebraic. As a conclusion, $\mathcal{E}_p$ is a semi-algebraic function. □

A.II.2.5 Convergence of $\mathcal{E}_p$ to $\mathcal{E}$

We have already seen that $\mathcal{E}_p(Y)$ converges to $\mathcal{E}(Y)$ almost surely when $p \to +\infty$. In practice, since we want to optimise through $\mathcal{E}_p$ as a surrogate for $\mathcal{E}$, we would wish for the strongest possible convergence. Below, we show almost-sure *uniform* convergence over any compact, which is substantially better than point-wise convergence. Note that this stronger mode of convergence is unfortunately still too weak to transport local optima properties.

Theorem A.II.3 (Uniform Convergence of $\mathcal{E}_p$)**.** Let $\mathcal{K} \subset \mathbb{R}^{n\times d}$ compact. We have $\mathbb{P}\left(\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\to+\infty]{} 0\right) = 1$, where for $f \in \mathcal{C}(\mathcal{K}, \mathbb{R})$, $\|f\|_{\ell^\infty(\mathcal{K})} := \sup_{x\in\mathcal{K}} |f(x)|$.

Proof. We shall temporarily write $\mathcal{E}_p(Y) = \mathcal{E}_p(Y;\Theta)$ to illustrate the dependency on the random variable $\Theta := (\theta_i)_{i\in\mathbb{N}^*}$ on a probabilistic space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(\mathbb{S}^{d-1})^{\mathbb{N}}$. By point-wise almost-sure convergence, for any fixed $Y \in \mathbb{R}^{n\times d}$, there exists a $\mathbb{P}$-null set $\mathcal{N}_Y$ such that for every $\omega \in \Omega \setminus \mathcal{N}_Y$, the deterministic real number $\mathcal{E}_p(Y;\Theta(\omega))$ converges to $\mathcal{E}(Y)$. Let $\mathcal{D} := \mathcal{K} \cap \mathbb{Q}^{n\times d}$,

which is dense in $\mathcal{K}$ and countable. Let $\mathcal{N} := \bigcup_{Y\in\mathcal{D}} \mathcal{N}_Y$: $\mathcal{N}$ is $\mathbb{P}$-null as a countable union of $\mathbb{P}$-null sets.

Fixing $\omega \in \Omega \setminus \mathcal{N}$, we have $\forall Y \in \mathcal{D}, \quad \mathcal{E}_p(Y; \Theta(\omega)) \xrightarrow[p\to+\infty]{} \mathcal{E}(Y)$, thus point-wise convergence on $\mathcal{D}$ of the (now) deterministic function $\mathcal{E}_p(\cdot; \Theta(\omega))$ to $\mathcal{E}$. Now, a consequence of Proposition A.II.1 is that the family of functions $(Y \mapsto \mathcal{E}_p(Y; \Theta'))_{\Theta'\in(\mathbb{S}^{d-1})^p}$ is equi-continuous on any compact (thus on $\mathcal{K}$). As a consequence, the point-wise convergence on $\mathcal{D}$ implies the uniform convergence of $\mathcal{E}_p(\cdot; \Theta(\omega))$ to $\mathcal{E}$ on $\overline{\mathcal{D}} = \mathcal{K}$ (a detailed presentation of this classic result can be found in [LHL12, Proposition 3.2]). This holds for any event $\omega \in \Omega \setminus \mathcal{N}$, with $\mathbb{P}(\Omega \setminus \mathcal{N}) = 1$, thus the uniform convergence is almost-sure. □

To complement this uniform almost-sure convergence, we prove a uniform Central Limit result on the error process $\sqrt{p}(\mathcal{E}_p - \mathcal{E})$ on a fixed compact set $\mathcal{K}$. This result provides insight on the law of the approximation error, uniformly with respect to the position $Y \in \mathcal{K}$.

Theorem A.II.4. Let $\mathcal{K} \subset \mathbb{R}^{n\times d}$ be a compact and non-empty set. On this domain, we have the following uniform Central Limit convergence in distribution of the approximation error of the random process $\mathcal{E}_p$:

$$\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathcal{K})} G, \tag{A.II.19}$$

where the convergence is in law in the sense of $\ell^\infty(\mathcal{K})$, the space of bounded functions $z : \mathcal{K} \longrightarrow \mathbb{R}$ equipped with the uniform norm. The limit process G is the centred Gaussian process on $\mathcal{K}$ of covariance

$$\mathbb{C}(G)[Y, Y'] = \int_{\mathbb{S}^{d-1}} w_\theta(Y) w_\theta(Y') \mathrm{d}\varpi(\theta) - \mathcal{E}(Y)\mathcal{E}(Y').$$

This result implies the convergence in law of the uniform error

$$\sqrt{p}\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} \|G\|_{\ell^\infty(\mathcal{K})}. \tag{A.II.20}$$

We provide the proof of Theorem A.II.4 in Section A.II.7.1, along with a brief presentation of the Donsker class arguments at hand.

Remark A.II.5. Our Central Limit result from Theorem A.II.4 allows one to build (uniform) confidence intervals for the approximation $\mathcal{E}_p \approx \mathcal{E}$ on any compact, but is of limited practical interest due to the complexity of estimating the Gaussian process G. Nevertheless, such confidence intervals provide additional theoretical insight on the Monte-Carlo approximation of the discrete SW distance.

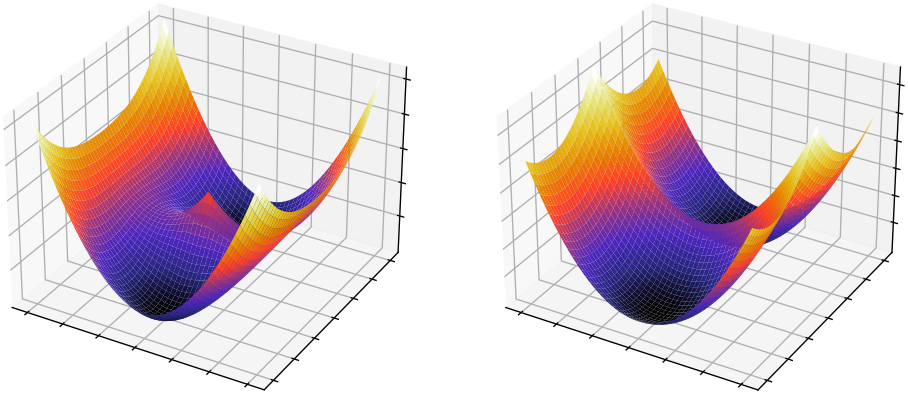
Our result Eq. (A.II.19) complements a result by Xi and Niles-Weed [XN22], which shows the following distributional convergence of a related process which is a function of θ:

$$\mathbb{H}_n := \{\sqrt{n}\left(\mathrm{W}_q^q(P_\theta\#\hat{\mu}_n, P_\theta\#\hat{\nu}_n) - \mathrm{W}_q^q(P_\theta\#\mu, P_\theta\#\nu)\right), \theta \in \mathbb{S}^{d-1}\} \xrightarrow[n\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathbb{S}^{d-1})} \mathbb{H},$$

where μ, ν are compactly supported probability measures, and $\hat{\mu}_n, \hat{\nu}_n$ are discrete empirical versions supported on n samples of respective laws μ, ν, and $\mathbb{H}$ is a centred Gaussian process on $\mathbb{S}^{d-1}$.

The distributional convergence in Eq. (A.II.19) also complements a (quantified) convergence in probability by Xu and Huang [XH22]. For $q \geq 1$ and $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$, let $M_q(\mu) := (\int \|x\|^q \mathrm{d}\mu)^{1/q}$ and $L := q\mathrm{W}_q^{q-1}(\mu, \nu)(M_q(\mu) + M_q(\nu))$. In [XH22, Proposition 4], they prove that for any $\varepsilon, \delta > 0$, if $p \geq \frac{2L^2}{(d-1)\varepsilon^2} \log(\frac{2}{\delta})$, then

$$\mathbb{P}(|\widehat{\mathrm{SW}}_{q,p}^q(\mu, \nu) - \mathrm{SW}_q^q(\mu, \nu)| \geq \varepsilon) \leq \delta, \tag{A.II.21}$$



(a) $Y \longmapsto \mathcal{E}(Y)$ (b) $Y \longmapsto \mathrm{W}_2^2(\gamma_Y, \gamma_Z)$

Figure A.II.1: Comparison between Sliced Wasserstein (a) and Wasserstein (b) landscapes for 2-point discrete measures $Y = (y, -y)^\top$ and $Z = (z_1, z_2)^\top$ with $z_1 = (0, -1)^\top$ and $z_2 = (0, 1)^\top$.

where $\widehat{\mathrm{SW}}_{q,p}^q(\mu, \nu)$ is the Monte-Carlo approximation of $\mathrm{SW}_q^q(\mu, \nu)$ with p projections. Their result is of a different nature, since it deals with general measures in $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ and does not study the process associated to moving the support of one of the measures. Point-wise, Eq. (A.II.21) from [XH22] is a more general result than Eq. (A.II.20), but the strength of our result comes from the study of the *process* $\mathcal{E}_p$, for which Eq. (A.II.21) is not informative in distribution, since it is a point-wise result. Furthermore, Eq. (A.II.21) is not tailored to our almost-sure uniform convergence case Theorem A.II.3.

A.II.2.6 Illustration in a simplified case

Let us illustrate $\mathcal{E}$ in a simple case, that was briefly presented in Bonneel et al. [Bon+15a], in order to grasp the difficulties at hand. This example is interesting for understanding the difficulty of performing computations with $\mathcal{E}$ and $\mathcal{E}_p$. Let $z_1 = (0, -1)^\top$ and $z_2 = (0, 1)^\top$. Instead of computing $\mathcal{E}(Y)$ for any $Y \in \mathbb{R}^{2\times 2}$, we simplify by assuming $Y = (y, -y)^\top = (y_1, y_2)^\top$. We will assume further $y \neq 0$ and write $y = (u, v)^\top$. The interested reader may seek the computations in Section A.II.7.2. With these notations, we can show that

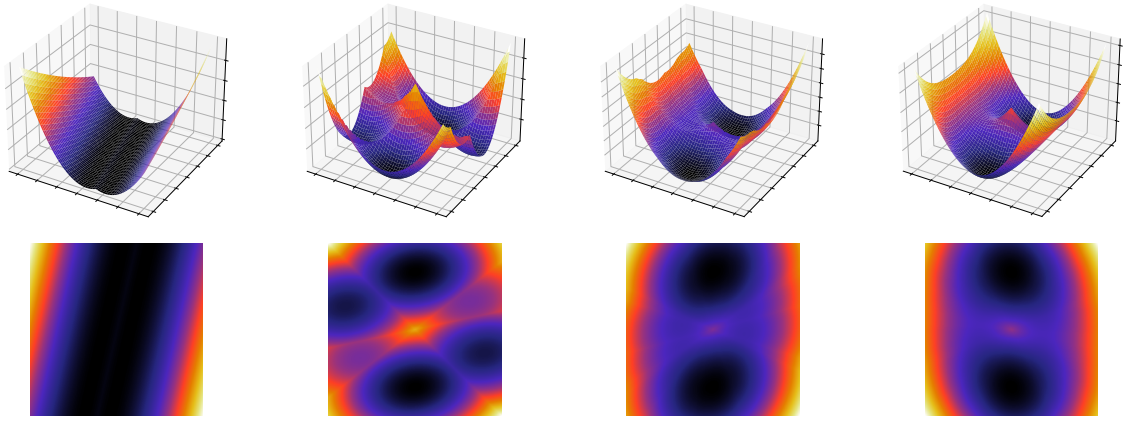
$$\mathcal{E}(Y) = \mathrm{SW}_2^2(\gamma_Y, \gamma_Z) = \frac{u^2 + v^2}{2} + \frac{1}{2} - \frac{2}{\pi}\left(|u| + |v| \operatorname{Arctan}\left|\frac{v}{u}\right|\right). \tag{A.II.22}$$

For W_2^2, one may show (see Section A.II.7.2 for the computations) that $\mathrm{W}_2^2(\gamma_Y, \gamma_Z) = u^2 + (|v| - 1)^2$ in this setting. We compare $\mathcal{E}$ and W_2^2 in Fig. A.II.1.

Notice differences in regularity. $\mathcal{E}$ is smooth on the open set $\mathcal{U}$ (defined in Eq. (A.II.12)) of the $Y \in \mathbb{R}^{n\times d}$ with distinct points (this is known in general, [Bon+15a]), but is not differentiable anywhere in $\mathcal{U}^c$. Here this is clear at $(0, 0)$. Furthermore, $\mathcal{E}$ presents two saddle points, $(\pm\frac{2}{\pi}, 0)$. In Section A.II.3.1.2, we shall study the critical points of $\mathcal{E}$ in full generality. Finally, W_2^2 presents the typical landscape of the minimum of two quadratics.

We now move to computing $\mathcal{E}_p$ in this setting. In the case $n = 2$, a significant simplification occurs since $\mathfrak{S}_2 = \{I, (2, 1)\}$, and we express a simple formula for the cells in the Appendix, see Section A.II.7.2. We illustrate the cell structure in Fig. A.II.2.

Notice that as p increases, the number of new strict local optima also increases, however their associated cells become very small, thus one may hope that the probability of ending up in a strict local optimum would decrease as p increases. Specifically, in the heatmap visualisation, one may notice 6 large cells for $p = 3$, and for $p = 10$, two large cells corresponding to the global optima, and 8 small cells which may present local optima. This observation suggests that as $p \longrightarrow +\infty$, the total size of cells containing local optima decreases, and thus the probability of a numerical scheme converging to a local optimum decreases as well. Moreover, it is clear for the



(a) $\mathcal{E}_p,\ p=1$ (b) $\mathcal{E}_p,\ p=3$ (c) $\mathcal{E}_p,\ p=10$ (d) $\mathcal{E}$

Figure A.II.2: The landscape $\mathcal{E}_p$ approaches $\mathcal{E}$ as p increases, but introduces numerous strict local optima. Notice that when p is too small ($p = 1 \leq d$ in particular), $\mathcal{E}_p$ even introduces other global optima.

landscape $\mathcal{E}_p$ with $p = 3$ that the critical points (points of differentiability with a null gradient) are exactly the minima of the cell quadratics. Remark that a cell may not contain the minimum of its quadratic, which is why we will refer to cells containing their minimum as "stable" (as is the case for all cells in $p = 3$ illustration, but seemingly not for $p = 10$).

As is suggested by Fig. A.II.2, even with a large number of projections p compared to the dimension d, the presence of strict local optima may prevent numerical solvers from converging to the global optimum $\gamma_Y = \gamma_Z$. This practical concern motivates the study of the landscapes $\mathcal{E}$ and $\mathcal{E}_p$, which is the topic of Section A.II.3.

A.II.3 Properties of the Optimisation Landscapes of $\mathcal{E}$ and $\mathcal{E}_p$

The goal of this section is to study the respective landscapes of $\mathcal{E}$ and $\mathcal{E}_p$, their critical points and the links between them.

A.II.3.1 Optimising $\mathcal{E}$

A.II.3.1.1 Global optima of $\mathcal{E}$

As its name suggests, the SW distance is indeed a distance on $\mathcal{P}_2(\mathbb{R}^d)$ (this result can be proven in the same manner for the q-SW distances, for $q \geq 1$).

Proposition A.II.6. [Bon13, Theorem 5.1.2] SW is a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

As a consequence, the global optima of $\mathcal{E}$ are exactly the points Y^* such that $\gamma_{Y^*} = \gamma_Z$, or said otherwise the points such that $(y_1^*, \cdots, y_n^*)$ is a permutation of $(z_1, \cdots, z_n)$.

A.II.3.1.2 Critical points of $\mathcal{E}$

A first step in studying the landscape $\mathcal{E}$ is to determine its critical points, which we define as the set of points Y where $\mathcal{E}$ is differentiable and $\nabla\mathcal{E}(Y) = 0$. Thanks to Theorem A.II.2, these critical points can be shown to satisfy a fixed point equation.

Corollary A.II.1 (Equation characterising the critical points of $\mathcal{E}$)**.** Let $Y \in \mathcal{U}$ (defined in Eq. (A.II.12)). For $(k,l) \in [\![1,n]\!]$, define $\Theta_{k,l}^{Y,Z} := \left\{\theta \in \mathbb{S}^{d-1} \mid \tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k) = l\right\} \subset \mathbb{S}^{d-1}$

and $S_{k,l}^{Y,Z} := d\int_{\Theta_{k,l}^{Y,Z}} \theta\theta^\top \mathrm{d}\sigma \in S_d^+(\mathbb{R})$. Y is a critical point of $\mathcal{E}$ iif Y satisfies

$$\forall k \in [\![1,n]\!],\ y_k = \sum_{l=1}^n S_{k,l}^{Y,Z} z_l. \tag{A.II.23}$$

Proof. Let $k \in [\![1,n]\!]$. We have $\mathbb{S}^{d-1} = \bigcup_{l=1}^n \Theta_{k,l}^{Y,Z}$, where the union is disjoint, therefore one may write

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial y_k}(Y) &= \frac{2}{n}\int_{\mathbb{S}^{d-1}} \theta\theta^\top (y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)}) \mathrm{d}\sigma(\theta) \\ &= \frac{2}{n}\sum_{l=1}^n \int_{\Theta_{k,l}} \theta\theta^\top (y_k - z_l)\mathrm{d}\sigma(\theta) = \frac{2}{dn} y_k - \frac{2}{dn}\sum_{l=1}^n S_{k,l}^{Y,Z} z_l,\end{aligned}$$

where we have used $\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\sigma(\theta) = I/d$ in the last equality. Equating the partial differential to 0 yields Eq. (A.II.23). □

Eq. (A.II.23) shows that the critical points can be written as combinations of the points (z_l), "weighted" by the normalised conditional covariance matrices $S_{k,l}^{Y,Z} = d\mathbb{E}_{\theta\sim\sigma}\left[\mathbb{1}(\theta \in \Theta_{k,l}^{Y,Z})\theta\theta^\top\right]$. With

$$\Psi := \begin{cases} \mathcal{U} & \longrightarrow & \mathbb{R}^{n\times d} \\ Y & \longmapsto & \begin{pmatrix} \sum_{l=1}^n z_l^\top S_{1,l}^{Y,Z} \\ \vdots \\ \sum_{l=1}^n z_l^\top S_{n,l}^{Y,Z} \end{pmatrix} \end{cases},$$

Eq. (A.II.23) writes as a fixed-point equation $Y = \Psi(Y)$.

Further notice that Ψ cannot be properly defined on $\mathcal{U}^c$, for instance if $n = 2$, and if $Y = (y, y)$, the two possible sorting choices $\tau_Y^\theta \in \{(1,2),(2,1)\}$ yield two different values for $\Psi(Y)$ (the first value is the second with the indices exchanged). We show below that Ψ is continuous on $\mathcal{U}$. Unfortunately, Ψ cannot be extended to the whole space $\mathbb{R}^{n\times d}$, since the restrictions $\Psi|_{C_\mathbf{m}}$ may have distinct limits at the borders of the cells.

Proposition A.II.7 (Regularity of Ψ)**.** Ψ is continuous on $\mathcal{U}$ (defined in Eq. (A.II.12)).

Proof. It is sufficient to prove the continuity of $G := Y \to S_{k,l}^{Y,Z}$ on $\mathcal{U}$, for k, l fixed. Let $Y \in \mathcal{U}$ and $\varepsilon > 0$. Define

$$\Theta_\varepsilon(Y) := \left\{\theta \in \mathbb{S}^{d-1} \mid \forall \delta Y \in B(0,\varepsilon),\ \left(\theta^\top y_{\tau_Y^\theta(k)} + \theta^\top \delta y_{\tau_{\delta Y}^\theta(k)}\right)_{k\in[\![1,n]\!]} \in \mathcal{U}_{n,1}\right\}, \tag{A.II.24}$$

with $\mathcal{U}_{n,1}$ the open set of lists $(x_1, \cdots, x_n) \in \mathbb{R}^n$ with distinct entries. By Bonneel et al. [Bon+15a, Appendix A, Lemma 2], $\forall \theta \in \Theta_\varepsilon(Y)$, $\forall \delta Y \in B(0,\varepsilon)$, $\tau_Y^\theta = \tau_{Y+\delta Y}^\theta$. Let ε small enough such that $\forall \delta Y \in B(0,\varepsilon)$, $Y + \delta Y \in \mathcal{U}$. Let $\delta Y \in B(0,\varepsilon)$. Separating the integral yields:

$$\begin{aligned} G(Y+\delta Y) &= \int_{\Theta_{k,l}^{Y+\delta Y,Z}} \theta\theta^\top \mathrm{d}\sigma(\theta) \\ &= \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)} \theta\theta^\top \mathrm{d}\sigma(\theta) + \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^\top \mathrm{d}\sigma(\theta). \end{aligned}$$

Using the fact that $\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y) = \Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)$, and denoting $\|\cdot\|_{\text{op}}$ the $\|\cdot\|_2$-induced operator norm on $\mathbb{R}^{d\times d}$, we get

$$\begin{aligned} G(Y+\delta Y) - G(Y) &= \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^\top \mathrm{d}\varpi(\theta) - \int_{\Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \theta\theta^\top \mathrm{d}\varpi(\theta), \\ \|G(Y+\delta Y) - G(Y)\|_{\text{op}} &\leq \int_{\Theta_{k,l}^{Y+\delta Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \left\|\theta\theta^\top\right\|_{\text{op}} \mathrm{d}\varpi + \int_{\Theta_{k,l}^{Y,Z} \bigcap \Theta_\varepsilon(Y)^c} \left\|\theta\theta^\top\right\|_{\text{op}} \mathrm{d}\varpi \\ &\leq 2\int_{\Theta_\varepsilon(Y)^c} 1 \mathrm{d}\varpi = 2\varpi(\Theta_\varepsilon(Y)^c). \end{aligned}$$

By Bonneel et al. [Bon+15a, Appendix A, Lemma 3], there exists a constant C such that $\varpi(\Theta_\varepsilon(Y)^c) \leq C\varepsilon$, which proves the continuity of G on $\mathcal{U}$. □

A.II.3.2 Optimising $\mathcal{E}_p$

A.II.3.2.1 Global optima of $\mathcal{E}_p$

We saw in Proposition A.II.6 that SW is a distance. Unfortunately, its discretised version $\widehat{\text{SW}}_p$ is only a pseudo-distance: non-negativity, symmetry and the triangular inequality still hold, but separation fails.

For generic measures, a measure-theoretic way of seeing this is through characteristic functions. Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$, the condition $\widehat{\text{SW}}_p(\mu,\nu) = 0$ is equivalent to $\forall i \in [\![1,p]\!]$, $\forall t \in \mathbb{R}$, $\phi_\mu(t\theta_i) = \phi_\nu(t\theta_i)$, where ϕ_μ (resp. ϕ_ν) is the characteristic function of μ (resp. ν). This condition only constrains the characteristic functions on p radial lines, and Bochner or Pólya-type criteria may be considered to find a characteristic function ϕ which equals ϕ_μ on these lines but differs on a non-null set.

The discrete case pertains more to our setting. As shown in Chapter A.I, for p large enough, almost-sure separation holds. This result can be proven by leveraging the geometrical consequences of the constrains $P_{\theta_i}\#\gamma_Y = P_{\theta_i}\#\gamma_Z$, and determining the a.s. solution set using random affine geometry.

Theorem A.II.5 (Theorem A.I.4)**.** Let $\gamma_Z := \sum_{l=1}^n b_l \delta_{z_l}$, where the (z_l) are fixed and distinct. Assuming $\theta_1, \cdots, \theta_p \sim \varpi^{\otimes p}$ and $n \geq 2$, we have

- if $p \leq d$, there exists ϖ-a.s. an infinity of measures $\gamma \neq \gamma_Z \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\widehat{\text{SW}}_p(\gamma, \gamma_Z) = 0$.
- if $p > d$, we have ϖ-almost surely $\{\gamma_Z\} = \underset{\gamma\in\mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}}\ \widehat{\text{SW}}_p(\gamma, \gamma_Z)$.

With a sufficient amount of projections, $\widehat{\text{SW}}_p(\gamma_Y, \gamma_Z) = 0 \Rightarrow \gamma_Y = \gamma_Z$ (a.s.), hence when minimising $\widehat{\text{SW}}_p(\gamma_Y, \gamma_Z)$ in Y, there is some hope of recovering γ_Z. Unfortunately, this does not guarantee that the (unique) solution will be attained numerically. This practical reality motivates the study of eventual local optima of $\mathcal{E}_p$.

The computation of the critical points of $\mathcal{E}_p$ can be done using the cell decomposition of Section A.II.2.3. We show that the critical points of $\mathcal{E}_p$ are exactly the local optima of $\mathcal{E}_p$, and correspond to "stable cells", which is to say cells that contain the minimum of their quadratic.

A.II.3.2.2 Critical points of $\mathcal{E}_p$ and cell stability

The objective of this section is to confirm theoretically some of the intuitions provided by the illustrations of Section A.II.2.6, namely that the critical points of $\mathcal{E}_p$ correspond to stable cells. Since the union of cells is exactly the differentiability set of $\mathcal{E}_p$, any critical point Y of $\mathcal{E}_p$ is necessarily within a cell $\mathcal{C}_\mathbf{m}$. Since $\mathcal{E}_p$ is quadratic on $\mathcal{C}_\mathbf{m}$, then a critical point Y is the minimum

of the cell's quadratic $q_{\mathbf{m}}$. As a consequence, the critical points of $\mathcal{E}_p$ are exactly the "stable cell optima", i.e. the $Y \in \mathcal{U}$ (see the definition Eq. (A.II.12)) such that $Y = \underset{Y' \in \mathbb{R}^{n\times d}}{\operatorname{argmin}}\ q_{\mathbf{m}(Y)}(Y')$.

The following theorem shows that there are no local optima of $\mathcal{E}_p$ outside of $\mathcal{U}$, and therefore that the set of local optima of $\mathcal{E}_p$, the set of critical points of $\mathcal{E}_p$ and the set of stable cell optima coincide. As previously, we define the set of critical points of $\mathcal{E}_p$ as the set of points Y where $\mathcal{E}_p$ is differentiable and $\nabla\mathcal{E}_p(Y) = 0$.

Theorem A.II.6 (The local optima of $\mathcal{E}_p$ are within cells)**.** Assume that $(\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p}$, then the following results hold $\mathfrak{o}$-almost surely. Let $Y \in \mathbb{R}^{n\times d}$ a local optimum of $\mathcal{E}_p$, then $\exists \mathbf{m} \in \mathfrak{S}_n^p$ such that $Y \in \mathcal{C}_{\mathbf{m}}$. As a consequence, we have the equality between the three sets:

- Local optima of $\mathcal{E}_p$;
- Critical points of $\mathcal{E}_p$;
- Stable cell optima: $\left\{ Y \in \mathcal{U} \mid Y = \underset{Y' \in \mathbb{R}^{n\times d}}{\operatorname{argmin}}\ q_{\mathbf{m}(Y)}(Y') \right\}$.

Proof. Let $Y \in \mathbb{R}^{n\times d}$ a local optimum of $\mathcal{E}_p$. Let $M := \{\mathbf{m} \in \mathfrak{S}_n^p \mid Y \in \overline{\mathcal{C}_{\mathbf{m}}}\}$.

Let $\mathbf{m} \in M$. Let us show that $\nabla q_{\mathbf{m}}(Y) = 0$ by contradiction: suppose $\nabla q_{\mathbf{m}}(Y) \neq 0$. For t positive and small enough,

$$\begin{aligned}\mathcal{E}_p(Y) &\leq \mathcal{E}_p\left(Y - t\frac{\nabla q_{\mathbf{m}}(Y)}{\|\nabla q_{\mathbf{m}}(Y)\|}\right) \leq q_{\mathbf{m}}\left(Y - t\frac{\nabla q_{\mathbf{m}}(Y)}{\|\nabla q_{\mathbf{m}}(Y)\|}\right) \\ &= q_{\mathbf{m}}(Y) - t\|\nabla q_{\mathbf{m}}(Y)\| + o(t) = \mathcal{E}_p(Y) - t\|\nabla q_{\mathbf{m}}(Y)\| + o(t).\end{aligned}$$

Therefore, for $t > 0$ sufficiently small, we have $\mathcal{E}_p(Y) < \mathcal{E}_p(Y)$, which is a contradiction. We now prove that $\#M = 1$. Using the notations of Remark A.II.4, for $\mathbf{m} \in M$, we have $\nabla q_{\mathbf{m}}(Y) = 0$, thus $B\overrightarrow{y} = a_{\mathbf{m}}$. For $(\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p}$, we have $\mathfrak{o}$-almost surely that B is invertible and that $\mathbf{m} \neq \mathbf{m}' \implies a_{\mathbf{m}} \neq a_{\mathbf{m}'}$, thus $\mathfrak{o}$-almost surely, $\#M = 1$, proving that in fact Y belongs to $\mathcal{C}_{\mathbf{m}}$ and not to its boundary.

□

A.II.3.2.3 Closeness of critical points of $\mathcal{E}_p$ and $\mathcal{E}$

In practice, all numerical optimisation methods converge towards a local optimum. Theorem A.II.6) of $\mathcal{E}_p$. One may wonder what is the link between the critical points of $\mathcal{E}_p$, which we reach in practice, and the critical points of $\mathcal{E}$, among which are the theoretical solutions we would like to reach.

The following theorem shows that at the limit $p \to +\infty$, any sequence of critical points of $\mathcal{E}_p$ become fixed points of Ψ Eq. (A.II.23) in probability, which is to say that they exhibit similar properties to the critical points of $\mathcal{E}$.

Theorem A.II.7 (Approximation of the fixed-point equation)**.**
For $p > d$, let Y_p any critical point of $\mathcal{E}_p$. Then we have the convergence in probability:

$$Y_p - \Psi(Y_p) \xrightarrow[p\longrightarrow+\infty]{\mathbb{P}} 0. \tag{A.II.25}$$

Specifically (see Corollary A.II.2), in order to reach a precision of ε, we have $\|Y_p - \Psi(Y_p)\|_{\infty,2} \leq \varepsilon$ with probability exceeding $1-\eta$ if $p \geq \mathcal{O}\left(d^3 n \log(1/\eta)/\varepsilon^3\right)$ and $p \geq \mathcal{O}\left(d^3 n^2 \log(1/\eta)/\varepsilon^2\right)$, omitting logarithmic multiplicative terms in d and n.

We provide the proof in Section A.II.7.4, where we also estimate more precisely the convergence rate. The idea behind this result stems from computing the minima of the quadratics.

Algorithm A.II.1: Minimising $\mathcal{E}_p$ with Block-Coordinate Descent

Data: Fixed axes $(\theta_1, \cdots, \theta_p) \in (\mathbb{S}^{d-1})^p$, projections $(z_k^\top \theta_i)_{k \in [\![1,n]\!],\, i \in [\![1,p]\!]}$.
Result: Positions $Y \in \mathbb{R}^{n \times d}$.

1 **Initialisation:** Draw $Y^{(0)} \in \mathbb{R}^{n \times d}$;
2 **for** $t \in [\![1, T_{\max}]\!]$ **do**
3 Update the OT maps by solving $\pi^{(t)} \in \underset{\pi \in \mathbb{U}^p}{\operatorname{argmin}}\ J(\pi, Y^{(t-1)})$;
4 Update the positions by solving $Y^{(t)} = \underset{Y \in \mathbb{R}^{n \times d}}{\operatorname{argmin}}\ J(\pi^{(t)}, Y)$;
5 **if** $\|Y^{(t)} - Y^{(t-1)}\|_{\infty,2} < \varepsilon$ **then**
6 Declare convergence and terminate.
7 **end**
8 **end**

Let $Y^* := \underset{Y}{\operatorname{argmin}}\ q_{\mathbf{m}}(Y)$, we have

$$y_k^* = A^{-1} \left(\frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top z_{\mathbf{m}_i(k)} \right) = \frac{A^{-1}}{p} \sum_{l \in [\![1,n]\!]} \sum_{\substack{i \in [\![1,p]\!] \\ \mathbf{m}_i(k) = l}} \theta_i \theta_i^\top z_l, \tag{A.II.26}$$

with $A = \frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top$ which approaches the covariance matrix of $\theta \sim \mathfrak{o}$, i.e. I/d. Likewise, $\frac{1}{p} \sum_{\substack{i \in [\![1,p]\!] \\ \mathbf{m}_i(k) = l}} \theta_i \theta_i^\top$ can be seen as an empirical conditional covariance, and it approaches $S_{k,l}^{Y,Z}/d$. We then apply matrix concentration inequalities to quantify the approximation error.

A.II.3.2.4 Critical points of $\mathcal{E}_p$ and Block Coordinate Descent

Leveraging on the cell structure of $\mathcal{E}_p$, we present an algorithm alternatively solving for the transport matrices and for the positions. Writing $\mathbb{U}$ the set of valid transport plans between two uniform measures with n points, we minimise the following energy (with $(\theta_1, \cdots, \theta_p)$ fixed)

$$J := \left\{ \begin{array}{ccc} \mathbb{U}^p \times \mathbb{R}^{n \times d} & \longrightarrow & \mathbb{R}_+ \\ (\pi^{(1)}, \cdots, \pi^{(p)}), Y & \longmapsto & \frac{1}{p} \sum_{i=1}^{p} \sum_{k=1}^{n} \sum_{l=1}^{n} (\theta_i^\top y_k - \theta_i^\top z_l)^2 \pi_{k,l}^{(i)} \end{array} \right. . \tag{A.II.27}$$

Observe that minimising J amounts to minimising $\mathcal{E}_p$.

The computation in Algorithm A.II.1, line 3 is done using standard 1D OT solvers [Fla+21], and the update on the positions at line 4 can be computed in closed form (we provide the closed-form expression in Section A.II.7.5 for the sake of reproducibility). BCD can be seen as a walk from cell to cell (see Section A.II.2.3), as illustrated in Fig. A.II.3. BCD moves from cell to cell and converges towards a stable cell optimum, and thus towards a local optimum of $\mathcal{E}_p$ (since these two sets are equal by Theorem A.II.6). This behaviour is further studied in the experimental section.

A.II.4 Stochastic Gradient Descent on $\mathcal{E}$ and $\mathcal{E}_p$

As seen in Section A.II.3.1.2, the optimisation properties of $\mathcal{E}$ and $\mathcal{E}_p$ indicate that optimising their landscapes might prove difficult in practice. In real-world applications, these landscapes (and especially $\mathcal{E}$, which is the most used) are minimised using Stochastic Gradient Descent. Perhaps unsurprisingly given the difficulties presented in Section A.II.3.1.2 and due to the non-differentiable and non-convex properties of the landscapes, there has been no attempt to prove

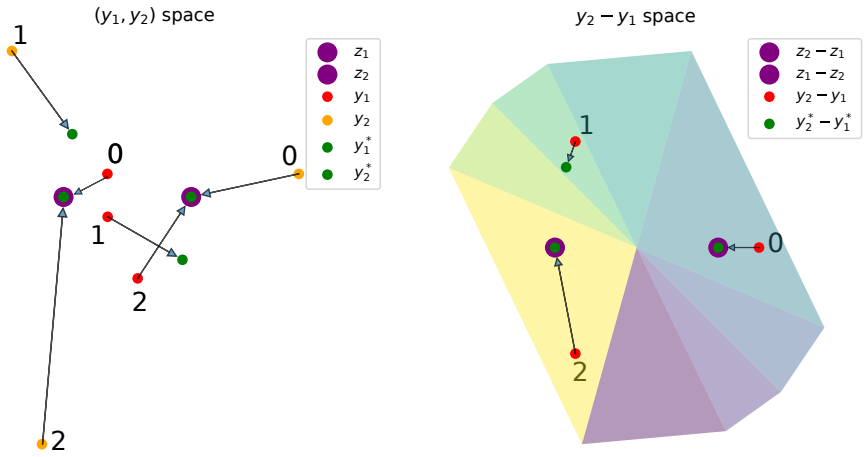


Figure A.II.3: Illustration of the cell structure for $p = 4$ in dimension 2 from a BCD viewpoint. On the left, we view different points $Y = (y_1, y_2)$ (in red and orange) and the minima of their respective quadratics: (y_1^*, y_2^*), which should be compared to the original points (z_1, z_2) in purple. On the right, we view the cell structure depending on the position of $y_2 - y_1 \in \mathbb{R}^2$, since the cell conditions only depend on this difference (see Eq. (A.II.46)). We can see that in this example all cells are stable, thus there are three strict local optima of $\mathcal{E}_p$ in addition to the global optimum. The (y_1, y_2) pair number 0 is sent to (z_2, z_1), while the pair "1" is sent to a local optimum, and the pair "2" is sent to (z_1, z_2).

the convergence of such SGD schemes in the literature (to our knowledge). This section aims to bridge this knowledge gap, using recent theoretical results on the convergence of constant-step SGD schemes due to Bianchi et al. [BHS22], and using results on decreasing-step SGD by Davis et al. [Dav+20]. Related works include Minibatch Wasserstein [Fat+21b] in particular Section 5 wherein they leverage another non-convex non-differentiable SGD convergence framework from Majewski et al. [MMM18] in order to derive convergence results for minibatch gradient descent on the Wasserstein and entropic Wasserstein distances.

There are other frameworks than Bianchi et al. [BHS22] or Davis et al. [Dav+20] that one may consider in order to prove non-smooth, non-convex convergence of SGD, in particular the work of Majewski et al. [MMM18]. Unfortunately, this work focuses on the case of *tame* functions, which is to say either *Clarke regular* functions ([Cla90]), or *stratifiable functions* ([Bol+07]). It is not known whether $\mathcal{E}$ is Clarke regular (the graph in Fig. A.II.1 could intuitively point to the contrary, due to the local shape in $-\|x\|$, and it is known that $-\|\cdot\|$ is not Clarke regular). Likewise, it is not known whether $\mathcal{E}$ is stratifiable. Thankfully, our regularity results from Section A.II.2 will allow us to show that $\mathcal{E}$ is *path differentiable*, which is another (more general) regularity class which is enough to apply the results from Bianchi et al. [BHS22] and Davis et al. [Dav+20].

Let us also mention the very recent work [LM25] (contemporary to ours), which studies the convergence of stochastic gradient schemes with decreasing steps and applied directly on probability measures instead of point clouds. Working with absolutely continuous measures μ and ν, the authors consider a scheme of the form

$$\mu^{(t+1)} = \left((1 - \alpha^{(t)})I + \alpha^{(t)} T_{\theta^{(t)}, \mu^{(t)}, \nu}\right) \# \mu^{(t)}, \tag{A.II.28}$$

where the $\theta^{(t)}$ are i.i.d. drawn from the uniform measure on the sphere, and where $T_{\theta^{(t)}, \mu^{(t)}, \nu}(x) := x + \left(\tau^{(t)}((\theta^{(t)})^\top x) - (\theta^{(t)})^\top x\right) \theta^{(t)}$, with $\tau^{(t)}$ the one-dimensional optimal transport map between the projected measures $P_{\theta^{(t)}} \# \mu^{(t)}$ and $P_{\theta^{(t)}} \# \nu$ (this map is uniquely defined on the support of $P_{\theta^{(t)}} \# \mu^{(t)}$ because $\mu^{(t)}$ and ν are absolutely continuous). It is quite easy to see that this scheme implements a stochastic gradient descent on $\mu \to \frac{1}{2}\mathrm{SW}_2^2(\mu, \nu)$. Building on proof techniques of [Bac+25], they show that if the sequence of learning rates $(\alpha^{(t)})_k$ is decreasing with $\sum \alpha^{(t)} = +\infty$ and $\sum (\alpha^{(k)})^2 < +\infty$, under some assumptions the sequence of measures $(\mu^{(t)})$

converges to ν for W_2. Unfortunately, extending these methods to discrete measures is not straightforward. Indeed, the authors of [Bac+25] claim that although they believe their results might hold for discrete measures, the difficulties of generalising their proofs to the discrete case were too important to achieve satisfying proofs in this case.

Before presenting our core results and the necessary theoretical framework from Bianchi et al. [BHS22], we provide in Algorithm A.II.2 the description of the SGD scheme used to minimise either $\mathcal{E}$ or $\mathcal{E}_p$, i.e. for projections drawn with $\mu \in \{\varpi, \varpi_p\}$ respectively. Starting with random initial points $Y^{(0)} \sim \nu$, at each step t, we draw a random projection $\theta^{(t+1)} \sim \varpi$ and compute an SGD iteration of step $\alpha^{(t)}$ in the direction of the gradient of $Y \longmapsto w_{\theta^{(t+1)}}(Y)$. This scheme uses optionally an additive noise term controlled by a parameter a (that can be set to 0). In Section A.II.4.2 to Section A.II.4.3, we shall study constant-step SGD schemes, and in Section A.II.4.5, we will focus on decreasing-step SGD.

Algorithm A.II.2: Minimising $\mathcal{E}$ or $\mathcal{E}_p$ with Stochastic Gradient Descent

Data: Learning rate sequence $(\alpha^{(t)})_{t\in\mathbb{N}}$, noise level $a \geq 0$, convergence threshold $\beta > 0$, and probability distribution μ on $\mathbb{S}^{d-1}$.
Result: Positions $Y \in \mathbb{R}^{n\times d}$, assignment $\tau \in \mathfrak{S}_n$.

Initialisation: Draw $Y^{(0)} \in \mathbb{R}^{n\times d}$;
for $t \in [\![0, T_{\max} - 1]\!]$ **do**
  Draw $\theta^{(t+1)} \sim \mu$ and $\varepsilon^{(t+1)} \sim \mathcal{N}(0, I_{nd})$.
  SGD update:
  $$Y^{(t+1)} = Y^{(t)} - \alpha^{(t)} \left[\frac{\partial}{\partial Y}\mathrm{W}_2^2(P_{\theta^{(t+1)}}\#\gamma_Y, P_{\theta^{(t+1)}}\#\gamma_Z)\right]_{Y=Y^{(t)}} + \alpha^{(t)} a\varepsilon^{(t+1)}$$
  if $\|Y^{(t+1)} - Y^{(t)}\|_{\infty,2} < \beta$ **then**
    Declare convergence and terminate.
  end
end
return $Y^{(t_{\text{final}})}$ *and the assignment* τ *of* $\mathrm{W}_2^2(P_{\theta^{(t_{\text{final}})}}\#\gamma_{Y^{(t_{\text{final}})}}, P_{\theta^{(t_{\text{final}})}}\#\gamma_Z)$.

Overview of Main Results

In [BHS22], Bianchi et al. establish conditions under which a constant-step SGD converges (in a certain sense), for a non-convex, locally Lipschitz cost function. Observe that both $\mathcal{E}$ and $\mathcal{E}_p$ are indeed locally Lipschitz, as shown in Theorem A.II.1. In Section A.II.4.2 and Section A.II.4.3, we verify the required conditions for $\mathcal{E}$ and $\mathcal{E}_p$ (with p fixed projections), and prove results which can be broadly summarised as follows:

Theorem (Theorem A.II.8: Convergence of the interpolated SGD (without noise) for $\mathcal{E}$ and $\mathcal{E}_p$)**.** Given a sequence of SGD schemes $(Y_\alpha^{(t)})$ for $\mathcal{E}$ (resp. $\mathcal{E}_p$) of steps α, their associated piecewise affine interpolated schemes (Y_α) converge, in a weak sense as $\alpha \longrightarrow 0$, to the set of solutions of the differential inclusion equation $\dot{Y}(s) \in -\partial_C\mathcal{E}(Y(s))$ (resp. $\mathcal{E}_p$), where $\partial_C\mathcal{E}$ denotes the Clarke differential of $\mathcal{E}$.

If we instead consider a *noised* SGD scheme (with noise magnitude $\alpha \times a$, $a > 0$), we have a stronger convergence result:

Theorem (Theorem A.II.9: Convergence of the noised SGD for $\mathcal{E}$ and $\mathcal{E}_p$)**.** Given a sequence of noised SGD schemes $(Y_\alpha^{(t)})$ for $\mathcal{E}$ (resp. $\mathcal{E}_p$) of steps α, they converge, in a weak sense as $\alpha \longrightarrow 0$, to the set of (Clarke) critical points of $\mathcal{E}$ (resp. $\mathcal{E}_p$).

These results rely on the notion of Clarke differentiability, which generalises differentiability to non smooth functions as soon as these functions are locally Lipschitz (*i.e.* Lipschitz in a

neighbourhood of each point). More precisely, for such a function $f : \mathbb{R}^d \to \mathbb{R}$, its Clarke sub-differential at x is defined as the convex hull of the limits of gradients of f

$$\partial_C f(x) := \operatorname{conv}\left\{ v \in \mathbb{R}^d : \exists (x_i) \in (\mathcal{D}_f)^{\mathbb{N}} : x_i \xrightarrow[i\longrightarrow+\infty]{} x \text{ and } \nabla f(x_i) \xrightarrow[i\longrightarrow+\infty]{} v \right\},$$

where $\mathcal{D}_f$ denotes the set of differentiability of f, whose complementary is of Lebesgue measure 0 by Rademacher's theorem, since f is locally Lipschitz. This notion of differentiability coincides with the classical one for differentiable functions, and with the usual sub-differential for convex functions. Clarke *critical points* of f are points x such that $0 \in \partial_C f(x)$.

In the case of decreasing-step SGD, in Section A.II.4.5 we leverage the results of [Dav+20] to prove the following result, under typical assumptions on the decreasing steps $(\alpha^{(t)})$.

> **Theorem** (Theorem A.II.10: Convergence of decreasing-step noised SGD for $\mathcal{E}$ and $\mathcal{E}_p$)**.** Consider $(Y^{(t)})$ a trajectory of noised decreasing-step SGD for $\mu \in \{\mathfrak{o}, \mathfrak{o}_p\}$ respectively, assume that it is almost-surely bounded. Then the sequence $F(Y^{(t)})$ is almost-surely convergent for $F \in \{\mathcal{E}, \mathcal{E}_p\}$ respectively, and almost-surely, any subsequential limit of $(Y^{(t)})$ belongs to the set of Clarke critical points of F.

A.II.4.1 Theoretical framework

In the following, we briefly present the theoretical framework of Bianchi et al. [BHS22]. They consider a function $f : \mathbb{R}^D \times \Theta \longrightarrow \mathbb{R}$, locally Lipschitz continuous in the first variable (for each θ), and μ a probability measure on $\Theta \subset \mathbb{R}^d$. Since f is locally Lipschitz in the first variable, the gradient $\nabla f(\cdot, \theta)$ of $f(\cdot, \theta)$ (w.r.t. the first variable) can be defined almost everywhere on $\mathbb{R}^D$, and any function $\varphi : \mathbb{R}^D \times \Theta \longrightarrow \mathbb{R}^D$ such that $\lambda \otimes \mu$ a.e., $\varphi = \nabla f$ is called an almost-everywhere gradient of f (see [BHS22, Definition 1]). Let $F := Y \longrightarrow \int_\Theta f(Y, \theta) \mathrm{d}\mu(\theta)$. A SGD scheme of step $\alpha > 0$ for F is a sequence $(Y^{(t)})$ of the form:

$$Y^{(t+1)} = Y^{(t)} - \alpha \varphi(Y^{(t)}, \theta^{(t+1)}), \quad \left(Y^{(0)}, (\theta^{(t)})_{t\in\mathbb{N}}\right) \sim \nu \otimes \mu^{\otimes \mathbb{N}}, \tag{A.II.29}$$

where ν is the distribution of the initial position $Y^{(0)}$, which we shall assume to be absolutely continuous w.r.t. the Lebesgue measure.

Within this framework, we can define an SGD scheme for $\mathcal{E}$ and $\mathcal{E}_p$. The function w_θ (Equation Eq. (A.II.8)) plays the role of f. We know from Proposition A.II.1 that w_θ is locally Lipschitz (uniformly in θ), hence differentiable almost everywhere, and that at these points of differentiability, using [Bon+15a, Appendix A, "proof of differentiability"], the derivative of w_θ in Y is

$$\varphi(Y, \theta) := \left[\frac{2}{n} \theta\theta^\top \left(y_k - z_{\tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k)} \right) \right]_{k \in [\![1,n]\!]}, \tag{A.II.30}$$

which corresponds to the definition of an almost-everywhere gradient as proposed by [BHS22]. Moreover, φ can be extended everywhere by choosing the sorting permutations arbitrarily when there is ambiguity.

Within this framework, given a step $\alpha > 0$, and an initial position $Y^{(0)} \sim \nu$, the fixed-step SGD iterations Eq. (A.II.29) can be applied to $F = \mathcal{E}$ by choosing $\mu = \mathfrak{o}$ or to $F = \mathcal{E}_p$ by choosing $\mu = \mathfrak{o}_p := \frac{1}{p}\sum_i^p \delta_{\theta_i}$. We assume that $\operatorname{Span}(\theta_i)_{i\in[\![1,p]\!]} = \mathbb{R}^d$, which is satisfied $\mathfrak{o}$-almost surely if $(\theta_i)_{i\in[\![1,p]\!]} \sim \mathfrak{o}^{\otimes p}$, since $p > d$.

A.II.4.2 Convergence of piecewise affine interpolated SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$

The *piecewise-affine interpolated SGD scheme* associated to a discrete SGD scheme $(Y_\alpha^{(t)})$ of step α is defined as:

$$Y_\alpha(s) = Y_\alpha^{(t)} + \left(\frac{s}{\alpha} - t \right) (Y_\alpha^{(t+1)} - Y_\alpha^{(t)}), \quad \forall s \in [t\alpha, (t+1)\alpha[, \quad \forall t \in \mathbb{N}.$$

We consider the space of absolutely continuous curves from $\mathbb{R}_+$ to $\mathbb{R}^D$, denoted $\mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^D)$, and endow it with the metric of uniform convergence on all segments:

$$d_c(Y, Y') := \sum_{k \in \mathbb{N}^*} \frac{1}{2^k} \min\left(1, \max_{s \in [0,k]} \|Y(s) - Y'(s)\|_{\infty,2}\right). \tag{A.II.31}$$

We will show that when the step decreases, the interpolated processes approach the set of solutions of a differential inclusion equation. To that end, we define the set of absolutely continuous curves that start within a given compact $\mathcal{K}$ of $\mathbb{R}^D$ and are a.e. solutions of the differential inclusion:

$$S_{-\partial_C F}(\mathcal{K}) := \left\{ Y \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^D) \mid \underline{\forall} s \in \mathbb{R}_+,\ \dot{Y}(s) \in -\partial_C F(Y(s));\ Y(0) \in \mathcal{K} \right\}, \tag{A.II.32}$$

where $\underline{\forall}$ denotes "for almost every". Bianchi et al. [BHS22] present three conditions under which they prove the convergence (in a certain weak sense) of interpolated SGD schemes on F. For the sake of self-containedness, we reproduce them here and verify them successively. Recall that for our two respective applications, $f(Y, \theta) = w_\theta(Y)$, $\mu \in \{\mathfrak{w},\ \mathfrak{w}_p\}$ and $F \in \{\mathcal{E},\ \mathcal{E}_p\}$.

Assumption A.II.1.

i) There exists $\kappa : \mathbb{R}^D \times \mathbb{S}^{d-1} \longrightarrow \mathbb{R}_+$ measurable such that each $\kappa(X, \cdot)$ is μ-integrable, and:

$$\exists \varepsilon > 0,\ \forall Y, Y' \in B(X, \varepsilon),\ \forall \theta \in \mathbb{S}^{d-1},\ |f(Y, \theta) - f(Y', \theta)| \leq \kappa(X, \theta)\|Y - Y'\|.$$

ii) There exists $X \in \mathbb{R}^D$ such that $f(X, \cdot)$ is μ-integrable.

Since f is the same in both cases, we can satisfy Assumption A.II.1 for both schemes simultaneously. The (quantified) uniformly locally Lipschitz property of the w_θ (Proposition A.II.1) allows us to verify Assumption A.II.1, by letting $r := 1$ and $\kappa(X, \theta) := \kappa_1(X)$. Assumption A.II.1 ii) is immediate since for *all* $Y \in \mathbb{R}^D$, $\theta \longmapsto w_\theta(Y)$ is continuous, therefore $\mathfrak{w} - L^1$ and $\mathfrak{w}_p - L^1$.

Assumption A.II.2. The function κ of Assumption A.II.1 verifies:

i) There exists $c \geq 0$ such that $\forall X \in \mathbb{R}^D$, $\displaystyle\int_{\mathbb{S}^{d-1}} \kappa(X, \theta) \mathrm{d}\mu(\theta) \leq c(1 + \|X\|)$.

ii) For every $\mathcal{K}$ compact of $\mathbb{R}^D$, $\displaystyle\sup_{X \in \mathcal{K}} \int_{\mathbb{S}^{d-1}} \kappa(X, \theta)^2 \mathrm{d}\mu(\theta) < +\infty$.

The choice $\kappa(X, \theta) := \kappa_1(X)$ (independent on θ, and as defined in Proposition A.II.1) satisfies Assumption A.II.2. We now consider the Markov kernel associated to the SGD schemes, denoting the Borel sets $\mathcal{B}(\mathbb{R}^D)$:

$$P_\alpha : \left\{ \begin{array}{ccc} \mathbb{R}^D \times \mathcal{B}(\mathbb{R}^D) & \longrightarrow & [0, 1] \\ Y, B & \longmapsto & \displaystyle\int_{\mathbb{S}^{d-1}} \mathbb{1}_B(Y - \alpha\varphi(Y, \theta)) \mathrm{d}\mu(\theta) \end{array} \right. .$$

With λ denoting the Lebesgue measure on $\mathbb{R}^D$, let

$$\Gamma := \{\alpha \in\ (0, +\infty) \mid \forall \rho \ll \lambda,\ \rho P_\alpha \ll \lambda\}.$$

We will verify the following assumption for both schemes:

Assumption A.II.3. The closure of Γ contains 0.

In Proposition A.II.8, we prove a stronger result, namely that Γ contains $(0, \frac{n}{2})$, which allows us to simply take learning rates $0 < \alpha < \frac{n}{2}$, instead of having to specify $\alpha \in \Gamma$. As a weaker

alternative, Assumption A.II.3 could also be verified by noticing that for any $\theta \in \mathbb{S}^{d-1}$, the function w_θ is of class $\mathcal{C}^2$ almost everywhere (as detailed in Section A.II.2.3), which allows us to apply [BHS22, Proposition 4] and shows that Assumption A.II.3 holds.

Proposition A.II.8. For schemes Eq. (A.II.29) applied to $\mathcal{E}$ or $\mathcal{E}_p$, $\Gamma = \mathbb{R}_+^* \setminus \{\frac{n}{2}\}$.

Proof. Let $\mu \in \{\varpi,\ \varpi_p\}$. Recall the line-by-line notation $Y = (y_1, \cdots, y_n)^\top \in \mathbb{R}^{n\times d}$. We also denote $Z_\tau := (z_{\tau(1)}, \cdots, z_{\tau(n)})^\top$ for $\tau \in \mathfrak{S}_n$. Let $\rho \ll \lambda$ and $B \in \mathcal{B}(\mathbb{R}^D)$ such that $\lambda(B) = 0$. We have, with $\alpha' := 2\alpha/n$:

$$\begin{aligned}
\rho P_\alpha(B) &= \int_{\mathbb{R}^D}\int_{\mathbb{S}^{d-1}} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^\top) + \alpha' Z_{\tau_Z^\theta\circ(\tau_Y^\theta)^{-1}}\theta\theta^\top\right)\mathrm{d}\mu(\theta)\mathrm{d}\rho(Y)\\
&\leq \sum_{\tau\in\mathfrak{S}_n}\int_{\mathbb{R}^D}\int_{\mathbb{S}^{d-1}} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^\top) + \alpha' Z_\tau\theta\theta^\top\right)\mathrm{d}\mu(\theta)\mathrm{d}\rho(Y)\\
&= \sum_{\tau\in\mathfrak{S}_n}\int_{\mathbb{S}^{d-1}} I_\tau(\theta)\mathrm{d}\mu(\theta),
\end{aligned}$$

where $I_\tau(\theta) := \displaystyle\int_{\mathbb{R}^D} \mathbb{1}_B\left(Y(I-\alpha'\theta\theta^\top) + \alpha' Z_\tau\theta\theta^\top\right)\mathrm{d}\rho(Y)$, and where the last line is obtained by applying Tonelli's theorem. Let $\tau \in \mathfrak{S}_n$ and $\theta \in \mathbb{S}^{d-1}$. We now assume $\alpha' \neq 1$, which is to say $\alpha \neq n/2$. We operate the affine change of variables $X = \phi(Y) := Y(I-\alpha'\theta\theta^\top) + \alpha' Z_\tau\theta\theta^\top$, which is invertible for $\alpha' \neq 1$. We have

$$I_\tau(\theta) = \int_{\mathbb{R}^D} \mathbb{1}_B(\phi(Y))\mathrm{d}\rho(Y) = \int_{\mathbb{R}^D} \mathbb{1}_B(X)\mathrm{d}\phi\#\rho(X) = \phi\#\rho(B).$$

Now since ϕ is affine and invertible, $\phi\#\rho \ll \lambda$, thus $\phi\#\rho(B) = 0$, and finally $\rho P_\alpha(B) = 0$. This proves that $\rho P_\alpha \ll \lambda$ for $\alpha > 0$ differing from $n/2$. □

Now that we have verified Assumption A.II.1, Assumption A.II.2 and Assumption A.II.3, we can apply [BHS22, Theorem 2], to $\mathcal{E}$ and $\mathcal{E}_p$. Let $0 < \alpha_0 < n/2$.

Theorem A.II.8. [BHS22, Theorem 2], applied to $\mathcal{E}$ and $\mathcal{E}_p$: convergence of the interpolated SGD scheme Let $(Y_\alpha^{(t)}), \alpha \in]0, \alpha_0], t \in \mathbb{N}$ a collection of SGD sequences associated to Eq. (A.II.29) applied to $\mathcal{E}$ or $\mathcal{E}_p$. Consider (Y_α) their associated piecewise affine interpolations. For any $\mathcal{K}$ compact of $\mathbb{R}^D$ and any $\varepsilon > 0$, we have for $F \in \{\mathcal{E},\ \mathcal{E}_p\}$ and $\mu \in \{\varpi,\ \varpi_p\}$ respectively

$$\lim_{\substack{\alpha\longrightarrow 0\\ \alpha\in\ (0,\alpha_0]}} \nu\otimes\mu^{\otimes\mathbb{N}}\left(d_c(Y_\alpha, S_{-\partial_C F}(\mathcal{K})) > \varepsilon\right) = 0, \tag{A.II.33}$$

where d_c is the metric of uniform convergence defined in Eq. (A.II.31).

It is to be understood that when the SGD step decreases, the interpolated schemes converge towards the set of solutions of the differential inclusion related to the continuous SGD equation. This convergence is weak: the distance to this set approaches 0 in probability, and $S_{-\partial_C F}(\mathcal{K})$ is a set of solutions which we do not know how to compute, however we can study some theoretical properties of the solutions given a suitable starting point $Y(0)$, see Remark A.II.6.

Remark A.II.6. For $\mathcal{E}$, if the initial position $Y^{(0)}$ belongs to a maximal connected component $\mathcal{V}$ of the differentiability set $\mathcal{U}$ (which is open), then consider the *gradient flow* differential equation

$$\frac{\partial\gamma}{\partial t}(Y,s) = -\nabla\mathcal{E}(\gamma(Y,s)), \quad \gamma(Y,0) = Y, \quad \gamma(Y,s) \in \mathcal{V}. \tag{A.II.34}$$

Since $\mathcal{E}$ is of class $\mathcal{C}^1$ on $\mathcal{V}$ (by Theorem A.II.2), with $\nabla\mathcal{E}$ Lipschitz (locally would suffice), standard flow results show that there exists a unique solution $\gamma(Y,\cdot)$ for any $Y \in \mathcal{V}$ defined

on some interval $(a_Y, b_Y) \subset \mathbb{R}$, which defines a continuous function $\gamma : \mathcal{D} \longrightarrow \mathcal{V}$, with $\mathcal{D} = \{(s, Y) \in \mathbb{R} \times \mathcal{V} \mid s \in (a_Y, b_Y)\}$. Since in our case, we consider a *gradient* flow, and since for any $c \in \mathbb{R}$, the set $A_c := \{Y \in \mathcal{V} \mid \mathcal{E}(Y) \leq c\}$ is compact, in fact the flows $\gamma(Y, s)$ are defined for $s \in [0, +\infty)$. Furthermore, if a sequence $(\gamma(Y, s_m))_{m\in\mathbb{N}}$ were to converge to a limit Y^∞, then one would have $Y^\infty \in \mathcal{V}$ and $\nabla\mathcal{E}(Y^\infty) = 0$. Our work does not show that the set $\mathcal{Z}_\mathcal{V} := \{Y \in \mathcal{V} \mid \nabla\mathcal{E}(Y) = 0\}$ of critical points of $\mathcal{E}$ is finite, however if that were the case, then more standard euclidean gradient flow results show that for any $Y \in \mathcal{V}$, $\exists Y^\infty \in \mathcal{Z}_\mathcal{V} :\ \gamma(Y, s) \xrightarrow[s\longrightarrow+\infty]{} Y^\infty$.

Note that given a learning rate $\alpha > 0$, an SGD scheme Eq. (A.II.29) applied to $\mathcal{E}$ and starting in $\mathcal{V}$ has no reason to stay in $\mathcal{V}$, and we unfortunately do not have equality with a discretised version of the gradient flow. However, thanks to Bianchi et al. [BHS22, Theorem 1], the trajectories stay almost-surely in differentiability points of $\mathcal{E}$ and w_θ, and thus almost-surely, $\varphi(Y^{(t)}, \theta^{(t+1)}) = \nabla w_{\theta^{(t+1)}}(Y^{(t)})$.

A.II.4.3 Convergence of Noised SGD Schemes on $\mathcal{E}$ and $\mathcal{E}_p$

In order to prove stronger convergence results we need to consider noised variants of our SGD schemes. Consider $\varepsilon \sim \eta := \mathcal{N}(0, I_D)$ an independent noise, our schemes become:

$$Y^{(t+1)} = Y^{(t)} - \alpha\varphi(Y^{(t)}, \theta^{(t+1)}) + \alpha a\varepsilon^{(t+1)}, \quad (Y^{(0)}, (\theta^{(t)})_{t\in\mathbb{N}}, (\varepsilon^{(t)})_{t\in\mathbb{N}}) \sim \nu \otimes \mu^{\otimes\mathbb{N}} \otimes \eta^{\otimes\mathbb{N}} \quad \text{(A.II.35)}$$

where $\mu = \varpi$ for $\mathcal{E}$ and ϖ_p for $\mathcal{E}_p$. We follow the method from [BHS22], which suggests that adding a small perturbation (that decreases with the step size) allows us to verify additional suitable assumptions. Note that this modification does not impact our verification of the previous assumptions 1 through 3. Bianchi et al. introduce the following assumption:

Assumption A.II.4. there exists $V, p : \mathbb{R}^D \to \mathbb{R}_+$ and $\beta : \mathbb{R}_+^* \to \mathbb{R}_+^*$ measurable, as well as $C \geq 0$, such that for any $\alpha \in \Gamma \cap]0, \alpha_0]$:

i) $\exists R(\alpha) > 0,\ \delta(\alpha) > 0,\ \exists\rho(\alpha)$ a probability measure on $\mathbb{R}^D$, such that:

$$\forall Y \in \overline{B}(0, R),\ \forall A \in \mathcal{B}(\mathbb{R}^D),\ P_\alpha(Y, A) \geq \delta\rho(A).$$

ii) $\sup_{Y\in\overline{B}(0,R)} V(Y) < +\infty$ and $\inf_{Y\in B(0,R)^c} p(Y) > 0$, with:

$$\forall Y \in \mathbb{R}^D,\ P_\alpha V(Y) \leq V(Y) - \beta(\alpha)p(Y) + C\beta(\alpha)\mathbb{1}_{\overline{B}(0,R)}(Y).$$

iii) $p(Y) \xrightarrow[Y\longrightarrow\infty]{} +\infty$.

Thanks to [BHS22, Proposition 5], this noised setting implies immediately Assumption A.II.4 i), for *any* choice of $R > 0$. They also suggest more restrictions on f that imply Assumption A.II.4 ii) and iii), which our use case does not satisfy. We shall verify Assumption A.II.4 ii) and iii) for $\mathcal{E}$ and $\mathcal{E}_p$ separately, but using similar methods. Beforehand, let us remark that the Markov kernel associated to Eq. (A.II.35) is determined by the following action on measurable functions $g : \mathbb{R}^D \longrightarrow \mathbb{R}$:

$$P_\alpha g(Y) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}^D} g(Y - \alpha\varphi(Y, \theta) + \alpha aX)\mathrm{d}\mu(\theta)\mathrm{d}\eta(X).$$

Proposition A.II.9 (Drift property for noised SGD on $\mathcal{E}$)**.** Let $V := \|\cdot\|_F^2$, $\alpha_0 < 1$ and

$p(Y) := \frac{2}{dn}(1-\alpha_0)\sum_{k=1}^{n}\|y_k\|_2^2$. There exists $R>0$ and $C\geq 0$:

$$\forall \alpha \in\ (0,\alpha_0),\ \forall Y\in\mathbb{R}^D,\ P_\alpha V(Y)\leq V(Y)-\alpha p(Y)+C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y).$$

Therefore, Assumption A.II.4 is satisfied for Eq. (A.II.35) when $\mu=\varpi$.

Proof. Let $Y\in\mathbb{R}^D$, $\alpha\in(0,1)$ and $\Delta(Y) := P_\alpha V(Y)-V(Y)$. We expand the square, then leverage the fact that η is centred, and decompose, writing $\varphi_k := \varphi(Y,\theta)_{k,\cdot}$:

$$\Delta(Y)=\alpha^2a^2nd+\underbrace{\alpha^2\sum_{k=1}^{n}\int_{\mathbb{S}^{d-1}}\varphi_k^\top\varphi_k\,\mathrm{d}\varpi(\theta)}_{\Delta_1(Y)}\underbrace{-2\alpha\sum_{k=1}^{n}\int_{\mathbb{S}^{d-1}}y_k^\top\varphi_k\mathrm{d}\varpi(\theta)}_{\Delta_2(Y)}.$$

We have $\varphi_k^\top\varphi_k=\frac{4}{n^2}(y_k-z_{\tau_Z^\theta\circ(\tau_Y^\theta)^{-1}(k)})^\top\theta\theta^\top(y_k-z_{\tau_Z^\theta\circ(\tau_Y^\theta)^{-1}(k)})$. Then recall that for all $\theta\in\Theta_{k,l}^{Y,Z}$, $z_{\tau_Z^\theta\circ(\tau_Y^\theta)^{-1}(k)}=z_l$. It follows that

$$\begin{aligned}\Delta_1(Y)&=\frac{4\alpha^2}{n^2}\sum_{(k,l)\in[\![1,n]\!]^2}\int_{\Theta_{k,l}^{Y,Z}}(x_k-z_l)^\top\theta\theta^\top(y_k-z_l)\mathrm{d}\varpi(\theta)\\&=\frac{4\alpha^2}{dn^2}\sum_{(k,l)\in[\![1,n]\!]^2}(y_k-z_l)^\top S_{k,l}^{Y,Z}(y_k-z_l)\\&\leq\frac{4\alpha\alpha_0}{dn^2}\sum_{(k,l)\in[\![1,n]\!]^2}\|y_k-z_l\|_2^2\leq\frac{4\alpha\alpha_0}{dn}\left(\sum_{k=1}^{n}(\|y_k\|_2^2+2\|Z\|_{\infty,2}\|y_k\|_2)+n\|Z\|_{\infty,2}^2\right),\end{aligned}$$

where we used the inequality $S_{k,l}^{Y,Z}\preceq I_d$.

Now for Δ_2, we have $y_k^\top\varphi_k=\frac{2}{n}(\theta^\top y_k)\theta^\top(y_k-z_{\tau_Z^\theta\circ(\tau_Y^\theta)^{-1}(k)})$, hence

$$\begin{aligned}\Delta_2(Y)&=-\frac{4\alpha}{dn}\sum_{(k,l)\in[\![1,n]\!]^2}y_k^\top S_{k,l}^{Y,Z}(y_k-z_l)\\&=-\frac{4\alpha}{dn}\sum_{k=1}^{n}\|y_2\|^2+\frac{4\alpha}{dn}\sum_{(k,l)\in[\![1,n]\!]^2}y_k^\top S_{k,l}^{Y,Z}z_l\\&\leq-\frac{4\alpha}{dn}\sum_{k=1}^{n}\|y_2\|^2+\frac{4\alpha}{d}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2,\end{aligned}$$

since $\sum_{l=1}^{n}S_{k,l}^{Y,Z}=I_d$ and $S_{k,l}^{Y,Z}\preceq I_d$. Finally,

$$\Delta(Y)\leq\alpha\Bigg[\underbrace{-\frac{4}{dn}(1-\alpha_0)\sum_{k=1}^{n}\|y_k\|_2^2}_{q(Y)}+\underbrace{\alpha_0a^2nd+\frac{4\alpha_0}{d}\|Z\|_{\infty,2}^2+\frac{4}{d}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2+\frac{8\alpha_0}{dn}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2}_{r(Y)}\Bigg].$$

Now since $\frac{r(Y)}{q(Y)}\xrightarrow[\|Y\|\longrightarrow+\infty]{}0$, there exists $R>0$ such that for $Y\in\mathbb{R}^D$ such that $\|Y\|_{\infty,2}>R$, we have $r(Y)\leq q(Y)/2$. In that case, we have $\Delta(Y)\leq\alpha(-q(Y)+q(Y)/2)=-\alpha q(Y)/2$. For $Y\in\mathbb{R}^D$ such that $\|Y\|_{\infty,2}\leq R$, we have $\Delta(Y)\leq\alpha r(Y)\leq\alpha\max_{\|Y\|_{\infty,2}\leq R}r(Y)=:C\alpha$ (C exists

since r is continuous on the compact $\overline{B}(0,R)$.) This proves that for any $Y \in \mathbb{R}^D$, $\Delta(Y) \leq -\alpha q(Y)/2 + C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y)$. □

We now turn to the scheme for $\mathcal{E}_p$. Let $A := \frac{1}{p}\sum_{j=1}^{p} \theta_j \theta_j^\top$, and consider $\lambda_{\min}(A)$ its smallest eigenvalue. Note that $\lambda_{\min}(A) > 0$, since we assumed $\mathrm{Span}(\theta_j)_{j\in[\![1,p]\!]} = \mathbb{R}^d$.

Proposition A.II.10 (Drift property for noised SGD on $\mathcal{E}_p$)**.** Let $V := \|\cdot\|_F^2$, $\alpha_0 < n$ and $q(Y) := \frac{2}{n}\left(1 - \frac{\alpha_0}{n}\right)\lambda_{\min}(A)\sum_{k=1}^{n}\|y_k\|_2^2$. There exists $R > 0$ and $C \geq 0$:

$$\forall \alpha \in (0, \alpha_0],\ \forall Y \in \mathbb{R}^D,\ P_\alpha V(Y) \leq V(Y) - \alpha q(Y) + C\alpha \mathbb{1}_{\overline{B}(0,R)}(Y).$$

Therefore, Assumption A.II.4 is satisfied for Eq. (A.II.35) when $\mu = \varpi_p$.

We leverage the same strategy as Proposition A.II.9, yet the technicalities of the upper-bounds differ.

Proof. Let $Y \in \mathbb{R}^D$ and $\alpha \in (0, \alpha_0]$. We expand the squares and use that η is centred:

$$\begin{aligned}\Delta(Y) &:= P_\alpha V(Y) - V(Y)\\ &= \alpha^2 a^2 nd + \underbrace{\alpha^2 \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{n}\sum_{i=1}^{d}\varphi(Y,\theta_j)_{k,i}^2}_{\Delta_1(Y)} \underbrace{-2\alpha\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{n}\sum_{i=1}^{d} y_{k,i}\varphi(Y,\theta_j)_{k,i}}_{\Delta_2(Y)}.\end{aligned}$$

On the one hand,

$$\begin{aligned}\Delta_1(Y) &= \frac{4\alpha^2}{pn^2}\sum_{j=1}^{p}\sum_{k=1}^{n}(y_k - z_{\tau_Z^{\theta_j}\circ(\tau_Y^{\theta_j})^{-1}(k)})^\top \theta_j\theta_j^\top (y_k - z_{\tau_Z^{\theta_j}\circ(\tau_Y^{\theta_j})^{-1}(k)})\\ &\leq \frac{4\alpha\alpha_0}{n^2}\left(n\|Z\|_{\infty,2}^2 + \sum_{k=1}^{n}\left(y_k^\top A y_k + 2\|Z\|_{\infty,2}\|y_k\|_2\right)\right).\end{aligned}$$

Similarly, $\Delta_2(Y) \leq -\frac{4\alpha}{n}\sum_{k=1}^{n} y_k^\top A y_k + \frac{4\alpha}{n}\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2$. Let

$$q_0(Y) := \frac{4}{n}\left(1 - \frac{\alpha_0}{n}\right)\lambda_{\min}(A)\sum_{k=1}^{n}\|y_k\|_2^2,$$

$$r(Y) := \alpha_0 a^2 nd + \frac{4\alpha_0}{n}\|Z\|_{\infty,2}^2 + \left(\frac{8\alpha_0}{n^2} + \frac{4}{n}\right)\|Z\|_{\infty,2}\sum_{k=1}^{n}\|y_k\|_2.$$

We have $\Delta(Y) \leq \alpha(-q_0(Y) + r(Y))$, and we can conclude using the same method as Proposition A.II.9. □

Finally, we require the fairly natural assumption that F admits a "chain rule".

Assumption A.II.5.
For any $Y \in \mathcal{C}_{\mathrm{abs}}(\mathbb{R}_+, \mathbb{R}^D)$, $\underline{\forall} s > 0$, $\forall V \in \partial_C F(Y(s))$, $V^\top \dot{Y}(s) = (F \circ Y)'(s)$.

In order to satisfy Assumption A.II.5, we will use the following result:

Proposition A.II.11. Any $F : \mathbb{R}^D \longrightarrow \mathbb{R}$ locally Lipschitz and semi-concave admits a chain rule for the Clarke sub-differential, and thus satisfies Assumption A.II.5.

Proof. Let $F : \mathbb{R}^D \longrightarrow \mathbb{R}$ locally Lipschitz and semi-concave. By [Via83, Proposition 4.5], this implies that $-F$ is Clarke regular. Then, by [BP21, Proposition 2], the fact that $-F$ is Clarke regular implies that F is path differentiable, and thus admits a chain rule, by [BP21, Corollary 2]. □

Since $\mathcal{E}$ is semi-concave (Proposition A.II.4) and locally Lipschitz, Proposition A.II.11 allows us to verify Assumption A.II.5 for Eq. (A.II.35). We may follow the same line of thought for $\mathcal{E}_p$, or alternatively we may use the fact that it is semi-algebraic (Proposition A.II.5). By [BP21, Proposition 2], this implies that $\mathcal{E}_p$ is path differentiable. Then by [BP21, Corollary 2], path differentiability implies having a chain rule for the Clarke sub-differential, which is verbatim [BHS22], Assumption A.II.5. We now have all the assumptions for [BHS22, Theorem 3]:

Theorem A.II.9. Applying [BHS22, Theorem 3]: convergence of noised SGD schemes to a critical point Consider a collection of noised SGD schemes $(Y_\alpha^{(t)})$, associated to Eq. (A.II.35), respectively for $F \in \{\mathcal{E},\ \mathcal{E}_p\}$, with steps $\alpha \in\ (0, \alpha_0]$, with $\alpha_0 < 1$. Let $\mathcal{Z}$ the set of Clarke critical points of F, i.e. $\mathcal{Z} := \left\{Y \in \mathbb{R}^D \mid 0 \in \partial_C F(Y)\right\}$. For $\mu \in \{\varpi,\ \varpi_p\}$ respectively, we have:

$$\forall \varepsilon > 0,\ \overline{\lim_{t \longrightarrow +\infty}}\ \nu \otimes \mu^{\otimes \mathbb{N}} \otimes \eta^{\otimes \mathbb{N}} \left(d(Y_\alpha^{(t)}, \mathcal{Z}) > \varepsilon\right) \xrightarrow[\substack{\alpha \longrightarrow 0 \\ \alpha \in (0, \alpha_0]}]{} 0.$$

It is to be understood that the euclidean distance between any sub-sequential limit of $(Y_\alpha^{(t)})_t$ and set of Clarke critical points $\mathcal{Z}$ approaches 0 in probability as the step size decreases. The distance d in the Theorem refers to the $\|\cdot\|_2$-induced distance between the point $Y_\alpha^{(t)} \in \mathbb{R}^D$ and the set $\mathcal{Z} \subset \mathbb{R}^D$.

Computing the set of Clarke critical points of $\mathcal{E}$ remains an open problem, and seems out of reach considering the difficulty of the simpler problem of computing the points where $\mathcal{E}$ is differentiable and $\nabla \mathcal{E} = 0$ (see the discussion in Section A.II.3.1.2). The difficulty lies at the boundaries of $\mathcal{U}$ (see Eq. (A.II.12)), where there is no longer unicity of the sorting permutations of $(\theta^\top x_k)_{k=1,\dots,n}$ and $(\theta^\top z_l)_{l=1,\dots,n}$ for ϖ-almost-every $\theta \in \mathbb{S}^{d-1}$. Computing the Clarke sub-differential at such points in closed form and determining the associated critical points seems out of reach since there is already no closed form for smooth critical points Corollary A.II.1.

For $\mathcal{E}_p$, the set of Clarke critical points strictly contains the set of critical points established in Theorem A.II.6. In general, the set of Clarke critical points that lie outside of the set of differentiability $\tilde{\mathcal{Z}} := \mathcal{Z} \cap (\cup_{\mathbf{m}} \mathcal{C}_{\mathbf{m}})^c$ is not empty, yet by Theorem A.II.6 it cannot contain a local optimum, and thus only contains saddle points. We believe that these saddle points will in practice never be the limit of our noised SGD trajectories, since intuition suggests that a trajectory attaining such a point at a certain time will find a decreasing direction almost-surely. Showing such a result rigorously is out of the scope of this chapter since this question is still an active field in simpler smooth cases [Jin+17; Jin+21]. More precisely, the minimisation of generic non-smooth non-convex functions F is also still actively studied. For instance, [DD22] and [BHS23b] investigate conditions to avoid convergence to certain saddle points, for randomly initialised deterministic (and potentially noised) proximal methods for semi-convex functions under novel strict saddle conditions. Another related reference is [DDJ23], which studies the non-convergence of noised sub-gradient descent to saddle points. We illustrate in Fig. A.II.4 the Clarke critical points of $\mathcal{E}_p$ for $p = 3$, on the numerical example of Section A.II.2.6.

In full generality (without the symmetry restriction, and with larger parameters p, n, d), one may expect that the Clarke critical points will have a similar structure.

A.II.4.4 Discussion on result generalisation

Batching. One may consider a variant in which at each step t, one draws a random batch of b directions independently from a measure μ over $\mathbb{S}^{d-1}$ ($\mu \in \{\varpi, \varpi_p\}$, for our purposes).

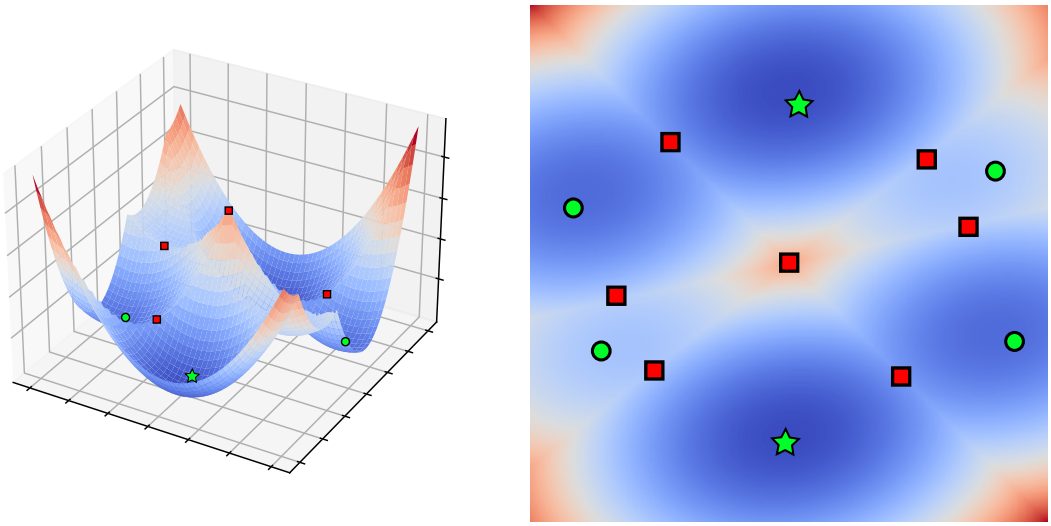


Figure A.II.4: The stars, circles and squares are the Clarke critical points of $x \mapsto \mathcal{E}_p(X = (-x, x)^\top)$, $x \in \mathbb{R}^2$ for $p = 3$. The squares do not correspond to local optima of $\mathcal{E}_p$, and are unlikely to be reached numerically. The circles and stars correspond to local optima of $\mathcal{E}_p$: the stars correspond to the global optima and satisfy the desired results $\mathcal{E}_p = 0$, while the circles are strict local optima.

Algorithmically, one does the following SGD scheme:

$$Y^{(t+1)} = Y^{(t)} - \frac{\alpha}{b}\sum_{j=1}^{b} \varphi(Y^{(t)}, \theta_j^{(t+1)}), \quad (Y^{(0)}, (\theta_j^{(t)})_{\substack{j\in[\![1,b]\!] \\ t\in\mathbb{N}}}) \sim \nu \otimes (\mu^{\otimes b})^{\otimes \mathbb{N}}. \tag{A.II.36}$$

In order to fit our theoretical framework (see Section A.II.4.1), we define

$$g(Y, (\theta_1, \cdots, \theta_b)) := \frac{1}{b}\sum_{j=1}^{b} f(Y, \theta_j).$$

Furthermore, the a.e. gradient of g becomes $\psi(\cdot, (\theta_1, \cdots, \theta_b)) := \frac{1}{b}\sum_{j=1}^{b} \varphi(\cdot, \theta_j)$ instead of $\varphi(\cdot, \theta^{(t)})$.

The function over which Eq. (A.II.36) performs SGD is:

$$\begin{aligned} G(Y) &= \int_{\mathbb{S}^{d-1}} g(Y, \theta_1, \cdots, \theta_b) \mathrm{d}\mu^{\otimes b}(\theta_1, \cdots, \theta_p) \\ &= \int_{\mathbb{S}^{d-1}} \frac{1}{b}\sum_{j=1}^{b} f(Y, \theta_j) \mathrm{d}\mu^{\otimes b}(\theta_1, \cdots, \theta_p) \\ &= \int_{\mathbb{S}^{d-1}} f(Y, \theta) \mathrm{d}\mu(\theta) = F(Y). \end{aligned}$$

One may check easily that if Assumptions 1 through 5 of Section A.II.4.1 are satisfied for (f, F), then they are satisfied for (g, G). As a consequence, all our results can be adapted without any difficulty to the batched setting.

Barycentres. If one were to replace $\mathcal{E}$ with the barycentre energy $\mathcal{E}_{\text{bar}}$ Eq. (A.II.6), the sample loss would become

$$g(Y, \theta) = \sum_{j=1}^{J} \lambda_j f_j(Y, \theta_j), \text{ where } f_j(Y, \theta) := \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_{Z^{(j)}}).$$

By sum, all of the previous results will hold, with the only technical point being path differentiability, which is stable by sum ([BP21, Corollary 4]). Note that this extension is also valid for a Monte-Carlo approximation of $\mathcal{E}$, replacing $\mathcal{E}$ with $\mathcal{E}_p$ in the barycentre formulation.

A.II.4.5 A Result for Decreasing Learning Rates

In [Dav+20], Davis et al. show the convergence of *decreasing-step* noised SGD of a function F under certain conditions. Our goal is to apply their [Dav+20, Theorem 4.2] to $F \in \{\mathcal{E}, \mathcal{E}_p\}$ with the following SGD scheme:

$$\begin{aligned} Y^{(t+1)} &= Y^{(t)} - \alpha^{(t)} \varphi(Y^{(t)}, \theta^{(t+1)}) + \alpha^{(t)} a \varepsilon^{(t+1)}, \\ (Y^{(0)}, (\theta^{(t)})_{t\in\mathbb{N}}, (\varepsilon^{(t)})_{t\in\mathbb{N}}) &\sim \nu \otimes \mu^{\otimes\mathbb{N}} \otimes \eta^{\otimes\mathbb{N}}, \end{aligned} \tag{A.II.37}$$

where as before, $\mu \in \{\varpi, \varpi_p\}$ for $\mathcal{E}, \mathcal{E}_p$ respectively, ν is the distribution of the initial position $Y^{(0)}$, $\eta := \mathcal{N}(0, I_D)$ is the noise distribution (it could be chosen more generally, but we attempt to stay close to our previous formalism for simplicity), and finally the learning rate sequence $(\alpha^{(t)})$ verify:

$$\forall t \in \mathbb{R},\ \alpha^{(t)} \geq 0, \quad \sum_{t=0}^{+\infty} \alpha^{(t)} = +\infty, \text{ and } \sum_{t=0}^{+\infty} (\alpha^{(t)})^2 < +\infty.$$

Theorem A.II.10. Consider $(Y^{(t)})$ a trajectory of Eq. (A.II.37) for $\mu \in \{\varpi, \varpi_p\}$ respectively, assume that it is almost-surely bounded. Then the sequence $F(Y^{(t)})$ is almost-surely convergent for $F \in \{\mathcal{E}, \mathcal{E}_p\}$ respectively, and almost-surely, any subsequential limit of $(Y^{(t)})$ belongs to the set of Clarke critical points of F.

Proof. We verify [Dav+20, assumptions C.1, C.2, C.3 and D.1, D.2], allowing us to apply [Dav+20, Theorem 4.2]. To begin with, the sequence $(\alpha^{(t)})$ was chosen to verify [Dav+20, assumption C.1], and Theorem A.II.10 explicitly [Dav+20, assumes C.2]. Regarding [Dav+20, p. C.3], the simple choice of independent noise verifies the martingale difference assumption trivially.

For [Dav+20, p. D.1], we need to show that the set of non-critical points of $\mathcal{E}$ and $\mathcal{E}_p$ are dense in $\mathbb{R}^D$. For $\mathcal{E}$, we can use Corollary A.II.1, which implies that critical points Y of $\mathcal{E}$ necessarily verify $\sum y_k = \sum_k z_k$, since $\sum_k S_{k,l}^{Y,Z} = I_d$, or are points of non-differentiability, which implies belonging to $\mathcal{U}$ (see Eq. (A.II.12)). In particular, critical points are necessarily within a union of two strict subspaces of $\mathbb{R}^D$, whose complementary is dense in $\mathbb{R}^D$. For $\mathcal{E}_p$, the property is easily verified using its decomposition into (non-trivial) quadratics on cells Proposition A.II.2.

For [Dav+20, p. D.2], we leverage [Dav+20, Lemma 5.2] along with the fact that $\mathcal{E}$ and $\mathcal{E}_p$ are path-differentiable (Section A.II.4.3), which shows that [Dav+20, p. D.2] is verified. □

The assumption that the trajectories are almost-surely bounded can be seen as a discrete version of [LM25, Assumption 4.4: A1)], which Li and Moosmüller require to prove the convergence of their decreasing-step SGD scheme for SW between absolutely continuous measures. While this assumption is theoretical costly, numerically we observe that the measure support Y remains bounded. Nevertheless, lifting this assumption would be of substantial mathematical interest.

A.II.5 Numerical Experiments

This section illustrates the optimisation properties of $\mathcal{E}$ and $\mathcal{E}_p$ with several numerical experiments. Section A.II.5.1 studies the optimisation of $\mathcal{E}_p$ using the BCD algorithm described in Algorithm A.II.1, which offers insights on the cell structure of $\mathcal{E}_p$ (Section A.II.2.3). Section A.II.5.2.1 focuses on stochastic gradient descent Algorithm A.II.2 and showcases various SGD trajectories on $\mathcal{E}$ and $\mathcal{E}_p$ for different learning rates, noise levels or numbers of projections, as well as the Wasserstein error along iterations. All the convergence curves shown throughout our experiments also showcase margins of error, computed by repeating the experiments several times, and corresponding to the 30% and 70% quantiles of the experiment.

In order to assess the quality of a position $Y^{(t)}$, perhaps the most germane metric is the Wasserstein distance: $\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$, which is why we will study the 2-Wasserstein error of BCD and SGD trajectories in this section. Unfortunately, this metric is not quite comparable for

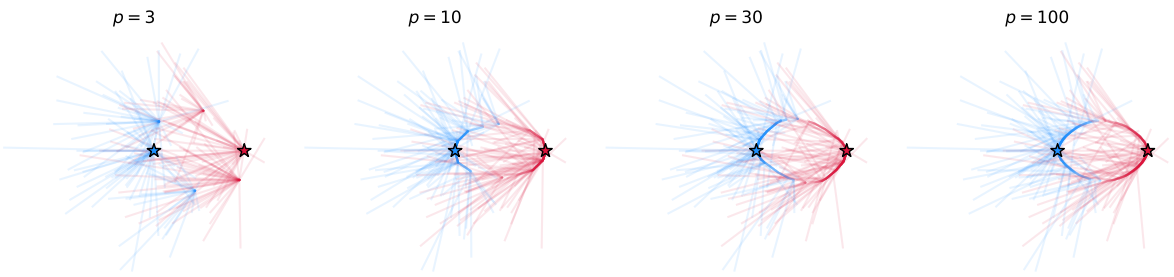


Figure A.II.5: BCD on $\mathcal{E}_p$ with different initial positions $Y^{(0)}$, with fixed projections (first sample). Each of the two points of the trajectory $Y^{(t)} = (y_1^{(t)}, y_2^{(t)})$ is coloured with respect to the point of the original measure γ_Z to which they converge.

different dimensions d, notably because $\|(1, \cdots, 1)\|_2^2 = d$. We shall attempt to compensate this phenomenon by using $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ instead, which makes the metric more comparable for measures on spaces of different dimensions.

A.II.5.1 Empirical study of Block Coordinate Descent on $\mathcal{E}_p$

In this section, we shall focus on studying the optimisation properties of the $\mathcal{E}_p$ landscape using the BCD algorithm (Algorithm A.II.1). This method leverages the cell structure of $\mathcal{E}_p$ (see Section A.II.2.3), by moving from cell to cell by computing the minimum of their associated quadratics (see the discussion in Section A.II.3.2.4). By Theorem A.II.6, all local optima of $\mathcal{E}_p$ are stable cell optima, i.e. fixed points of the BCD, which summarises briefly the ties between BCD and the optimisation properties of $\mathcal{E}_p$. As for the numerical implementation, Algorithm A.II.1 was implemented in Python with Numpy [Har+20] using the closed-form formulae for the updates.

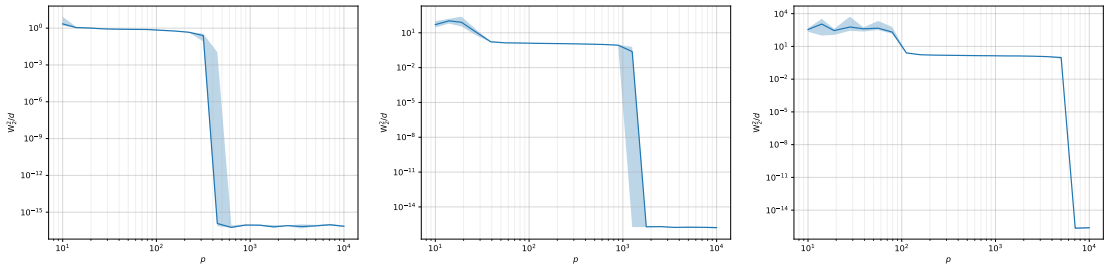
A.II.5.1.1 Illustration in 2D

Dataset and implementation details.. We start by setting a simple 2D measure γ_Z with a support of only two points represented with stars in Fig. A.II.5. The measure weights are taken as uniform. We fix sequences of p projections $(\theta_1, \cdots, \theta_p)$ for $p \in \{3, 10, 30, 100\}$ respectively. We then draw 100 BCD schemes with different initial positions $Y^{(0)} \in \mathbb{R}^{2\times 2}$, drawn with independent standard Gaussian entries. We take a stopping criterion threshold of 10^{-5} (see Algorithm A.II.1), and limit to 500 iterations.

In the case $p = 3$, we observe on Fig. A.II.5 four points which correspond to strict local optima, and the schemes appear to have a comparable probability of converging towards each of them. Note that these points are essentially the same as the ones represented in Fig. A.II.2 for $p = 3$, but that they depend on the projection sample. Between the two projection realisations, we observe that these local optima change locations. The cases $p \in \{10, 30, 100\}$ also exhibit strict local optima, however they appear to be decreasingly likely to be converged towards. For $p = 30$ and $p = 100$, notice that most trajectories end up on the same ellipsoid arcs towards the solution Z, and further remark that these arcs strongly resembles the trajectories of SGD schemes on $\mathcal{E}$ for small learning rates (see Fig. A.II.9 in Section A.II.5.2).

A.II.5.1.2 Wasserstein convergence of BCD schemes on $\mathcal{E}_p$

Final Wasserstein error of BCD Schemes.. For a dimension $d \in \{10, 30, 100\}$ and $n = 20$ points, the original measure γ_Z, $Z \in \mathbb{R}^{n\times d}$ is sampled once for all with independent standard Gaussian entries. Then, for varying numbers of projections p, we draw a starting position $Y^{(0)} \in \mathbb{R}^{n\times d}$ with entries that are uniform on $[0, 1]$; and draw p projections as input to the BCD algorithm. We set the stopping criterion threshold as $\varepsilon = 10^{-5}$ and the maximum iterations to 1000. In order to produce Fig. A.II.6, we record the normalised 2-Wasserstein discrepancy $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z)$ at the final iteration T for 10 realisations for each value of p and d.



(a) $d = 10$ (b) $d = 30$ (c) $d = 100$

Figure A.II.6: We consider BCD schemes with different amounts of projections p, and with an original measure γ_Z comprised of $n = 10$ points in dimension $d \in \{10, 30, 100\}$, which is fixed as a standard Gaussian realisation for each value of d. The stopping threshold was chosen as $\varepsilon = 10^{-5}$, and we plot the final Wasserstein errors $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^T}, \gamma_Z)$ at the final iteration T. For each set of values for the parameters, we perform 10 realisations with different initialisations $Y^{(0)}$ (drawn with uniform $[0, 1]$ entries), and different projections $(\theta_1, \cdots, \theta_p)$.

As a first estimation of the difficulty of optimising $\mathcal{E}_p$, we consider the evolution - as p increases - of final W_2^2 errors of BCD schemes. The results of the experiments presented in Fig. A.II.6 suggests the existence of a phase transition between an insufficient and a sufficient amount of projections. For instance, in the case $d = 10$, there appears to be a cutoff around $p = 400$, under which all the BCD realisations converge towards strict local optima, and past which we observe convergence up to numerical precision.

Probability of convergence of BCD schemes.. We can investigate further this empirical cutoff phenomenon by estimating the probability of convergence of a BCD algorithm. This probability is loosely related to the difficulty of optimising the landscape $\mathcal{E}_p$, since a high probability of BCD convergence indicates either a small number of strict local optima, or that their corresponding cells are extremely small and seldom reached in practice. For varying numbers of projections p and dimensions d, we run 100 realisations of BCD schemes. Each sample draws a target measure γ_Z, $Z \in \mathbb{R}^{n\times d}$ with independent standard Gaussian entries and $n = 10$ points, as well as its initialisation $Y^{(0)} \in \mathbb{R}^{n\times d}$ with entries that are uniform on $[0, 1]$ and p projections. Every BCD scheme has a stopping threshold of $\varepsilon = 10^{-5}$ and a maximum of 1000 iterations. We consider that a sample scheme has converged (towards the global optimum γ_Z) if $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z) < 10^{-5}$, which allows us to compute an empirical probability of convergence for each value of (p, d).

The findings in Fig. A.II.7 indicate that the W_2^2 error cutoffs from Fig. A.II.6 have a probabilistic counterpart: the probability of converging to a global optimum transitions from almost 0 to almost 1 relatively suddenly (in the logarithmic scale). We can conjecture that this drop in optimisation difficulty is tied to the number of iterations needed for the convergence of SGD schemes on $\mathcal{E}$, especially given the similar behaviour for the W_2^2 error in Fig. A.II.13.

A.II.5.2 Empirical study of SGD on $\mathcal{E}$ and $\mathcal{E}_p$

General numerical implementation.. In order to perform gradient descent on $\mathcal{E}$ or $\mathcal{E}_p$, we compute the gradient Eq. (A.II.30) using Pytorch's [Pas+19] Stochastic Gradient Descent optimiser, which back-propagates gradients through the loss $w_\theta := Y \mapsto \mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_Z)$, which we compute using the 1D Wasserstein solver from Python Optimal Transport [Fla+21].

A.II.5.2.1 Illustration in 2D

2D dataset and implementation details.. We define a 2D spiral dataset with the measure γ_Z, $Z = (z_1, \cdots, z_{10})^\top \in \mathbb{R}^{10\times 2}$ with $z_k = \frac{2k}{10}\left(\cos\left(2k\pi/10\right), \sin\left(2k\pi/10\right)\right)^\top$, and $k \in [\![1, 10]\!]$. The initial position $Y^{(0)}$ is fixed and remains the same across realisations. For schemes on $\mathcal{E}$, the projections $\theta^{(t)} \sim \mathfrak{o}$ are fixed beforehand and are the same across experiments. For every

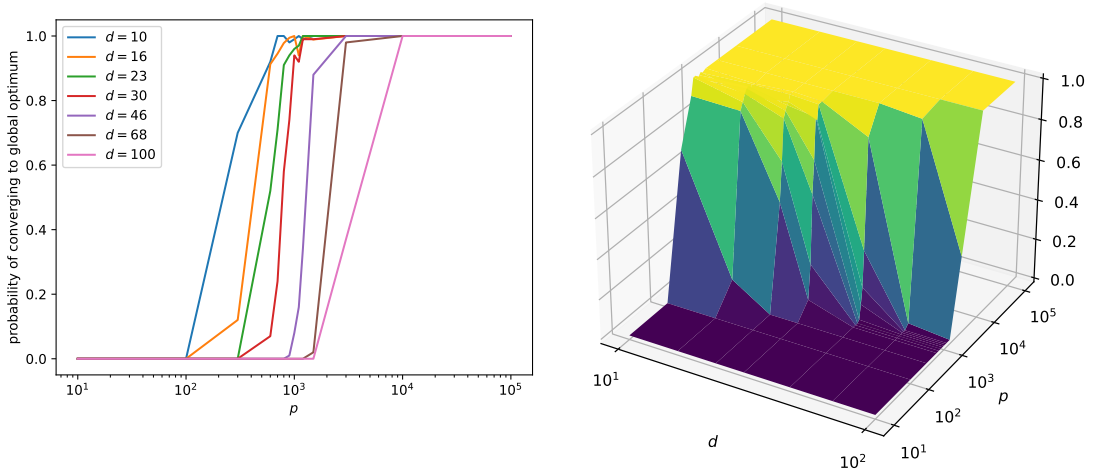


Figure A.II.7: Given a number of projections p, we run 100 BCD trials with different initial positions (with entries drawn as uniform on $[0, 1]$), projections and target measure supported by $Z \in \mathbb{R}^{n\times d}$, with $n = 10$ points in different dimensions $d \in [10, 100]$, where Z is drawn with independent standard Gaussian entries. At the final iteration T, we determine whether the optimum is global by a threshold criterion: $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(T)}}, \gamma_Z) < 10^{-5}$ and compute an empirical probability of convergence.

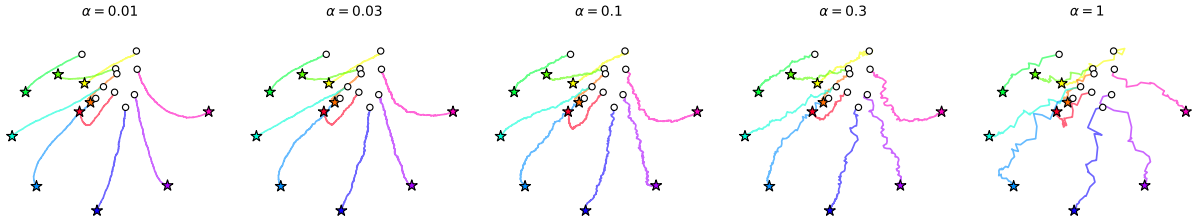


Figure A.II.8: SGD trajectories on $\mathcal{E}$ for different learning rates α. All the trajectories are computed using the same projection sequence $(\theta^{(t)})$.

realisation of a scheme on $\mathcal{E}_p$, p unique projections $(\theta_1, \cdots, \theta_p)$ are drawn, then the projections $(\theta^{(t)})$ for the iterations are drawn from these p fixed projections. For noised schemes, the only variable that is drawn at every sample is the noise $(\varepsilon^{(t)})$. Note that the associated energy landscapes are extremely similar to those illustrated in Section A.II.2.6 and in particular in Fig. A.II.2.

Fig. A.II.8 and Fig. A.II.9 illustrate the convergence of SGD schemes on $\mathcal{E}$ towards the original measure γ_Z, for different learning rate α (provided that α is under a divergence threshold). Theorem A.II.8 allowed us only to expect a convergence to a *solution of a Clarke Differential Inclusion* on $\mathcal{E}$ Eq. (A.II.32), yet in practice we seem to have convergence to a global optimum. Furthermore, Theorem A.II.8 shows that the interpolated SGD trajectories are approximately solutions of the DI $\dot{X}(t) \in -\partial_C\mathcal{E}(X(t))$, which, assuming that the trajectory stays in $\mathcal{U}$, amounts to $\dot{X}(t)+\nabla\mathcal{E}(X(t)) = 0$, which is exactly the Euclidean Gradient Flow of $\mathcal{E}$, as discussed in more detail in Remark A.II.6. This illustration suggests that the SGD schemes approach the gradient flow Eq. (A.II.34) as $\alpha \longrightarrow 0$, whereas Theorem A.II.8 predicts a (weak) convergence towards the set of solutions of the DI Eq. (A.II.32), which is equal to the gradient flow provided that the initial position $Y^{(0)}$ belongs to the differentiability set of $\mathcal{E}$ (see Remark A.II.6 for details). Note that higher learning rates lead to a "noisier" trajectory, which may impede upon the quality of the assignment. This shows that there is a trade-off: lower values of α allow for a better approximation of the (or a) gradient flow of $\mathcal{E}$ and potentially a more precise final position Y and assignment τ, however a larger value of α yields a substantially faster convergence.

Fig. A.II.10 presents a case where noised SGD schemes on $\mathcal{E}$ "converge" whatever the noise

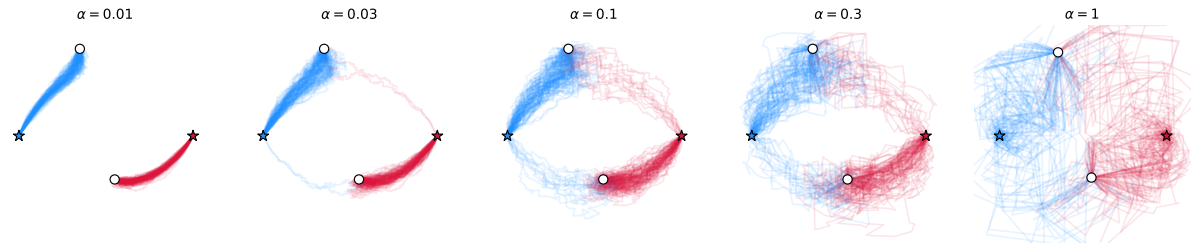


Figure A.II.9: SGD trajectories on $\mathcal{E}$ for different learning rates α. For each value of α, 100 samples are drawn with different projections $(\theta^{(t)})$, and for each realisation, each of the two points of the trajectory is coloured with respect to the point of the original measure γ_Z (represented by stars) to which they converge. The initial position $Y^{(0)}$ is represented by circles.

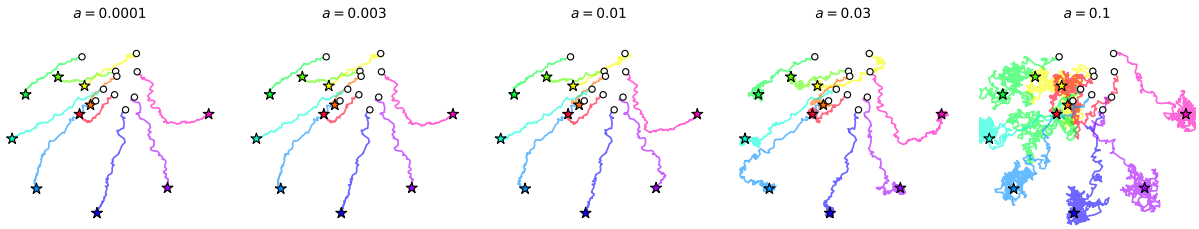


Figure A.II.10: SGD trajectories on $\mathcal{E}$ for different noise levels a. All the trajectories are computed using the same projection sequence $(\theta^{(t)})$. The learning rate is fixed at $\alpha = 0.3$.

level to a global optimum of $\mathcal{E}$. Note that the additive noise causes the scheme to oscillate around a solution, with a movement akin to Brownian motion with a scale tied to αa. Theorem A.II.9 shows that such schemes converge (as the step approaches 0) to *Clarke critical points* of $\mathcal{E}$, which could theoretically be a saddle point of strict local optimum. In this experiment, we observe convergence to a global optimum.

Fig. A.II.11 illustrates that SGD schemes on $\mathcal{E}_p$ may converge to strict local optima, which is to be expected, given how numerous they may be (see the discussion in Section A.II.2.6 and Fig. A.II.2 therein). For $p = 1$, entire lines are local optima, and for $p = 3$ and $p = 30$, we also observe convergence to strict local optima. Notice that for a large value of p such as $p = 100 \gg d = 2$, we have similar trajectories in Fig. A.II.12 compared to the $\mathcal{E}$ counterpart in Fig. A.II.9 ($\alpha = 0.03$). This observation suggests a stronger property than our results on the approximation of $\mathcal{E}$ by $\mathcal{E}_p$: uniform convergence in Theorem A.II.3 and a weak link between critical points Theorem A.II.7. To be precise, this illustration could allow one to hope for a result on the high probability for the proximity of SGD schemes on $\mathcal{E}_p$ and on $\mathcal{E}$ as $p \longrightarrow +\infty$, perhaps with conditions on the sequence of projections $(\theta^{(t)})$.

A.II.5.2.2 Wasserstein convergence of SGD schemes on $\mathcal{E}$ and $\mathcal{E}_p$

SGD on $\mathcal{E}$.. The original measure γ_Z, $Z \in \mathbb{R}^{n \times d}$ is sampled once for all with independent standard Gaussian entries. For each value of the parameter of interest (the learning rate α or

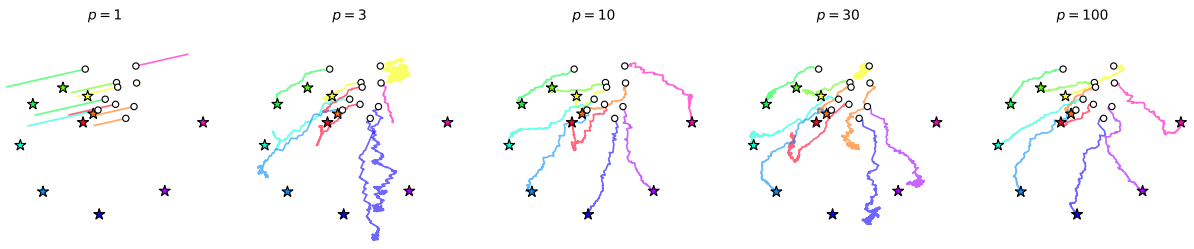


Figure A.II.11: SGD schemes on $\mathcal{E}_p$ for different number of projections p. The learning rate is fixed at $\alpha = 0.3$.

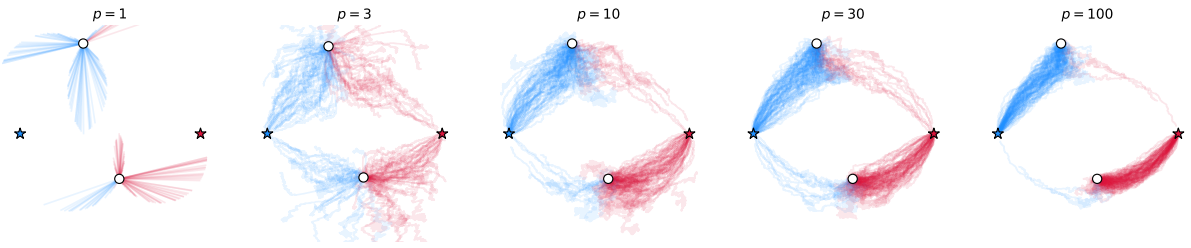


Figure A.II.12: SGD schemes on $\mathcal{E}_p$ for different number of projections p. For each value of p, 100 samples are drawn with different projections $(\theta_1, \cdots, \theta_p)$. For each realisation, each of the two points of the trajectory is coloured with respect to the point of the original measure γ_Z (represented by stars) to which they converge. The initial position $Y^{(0)}$ is represented by circles. The learning rate is fixed at $\alpha = 0.03$.

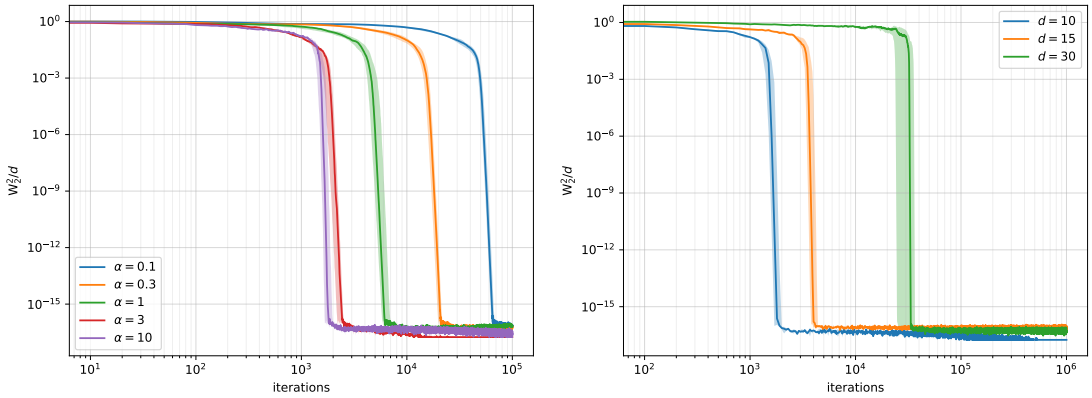


Figure A.II.13: Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for SGD iterations $Y^{(t)}$ on $\mathcal{E}$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. Left: different learning rates α for points in dimension $d = 10$. Right: different dimensions with $\alpha = 10$ (right).

the dimension d respectively), 10 realisations of the SGD schemes are computed with a different initial position $Y^{(0)}$, drawn with independent entries uniform on $[0, 1]$, and different projections $(\theta^{(t)})$. The SGD stopping criterion threshold (see Algorithm A.II.2) is set as negative, in order to always end at the maximum number of iterations, 10^6. For the experiment with varying learning rates α, we consider measures with $n = 20$ points in dimension $d = 10$. For the experiment with varying dimensions d, we still take $n = 20$ and use the learning rate $\alpha = 10$.

In Fig. A.II.13, we observe that the SGD schemes converge towards the true measure γ_Z up to numerical precision, which corresponds to a stronger convergence than the one predicted by Theorem A.II.8. The number of iterations needed for convergence obviously depends on the learning rate α, which notably can be chosen larger than $n/2$, which is a case that does not fall under the conditions for Theorem A.II.8. However, in this particular experiment, the SGD schemes diverged as soon as $\alpha \geq 30$, which could suggest that limiting oneself to $\alpha \preceq n$ is reasonable. The dimension d increases significantly the number of iterations required for convergence, furthermore we observe a transition from high W_2^2 error to low error, which is relatively sudden in logarithmic space. These first studies invites an in-depth analysis of the amount of iterations needed to reach convergence, which we propose in Fig. A.II.16. The final $\frac{1}{d}\mathrm{W}_2^2$ error does not seem to depend significantly on the dimension d, which provides empirical grounds for the $1/d$ normalisation choice.

Noised SGD on $\mathcal{E}$.. Fig. A.II.14 shows the Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for the noised SGD iterations on $\mathcal{E}$. The numerical setup is the same as above, with the addition of the noise $a\alpha\varepsilon^{(t)}$ at each iteration, where $\varepsilon^{(t)}$ has independent standard Gaussian entries, a is the noise level and α is the learning rate (set to $\alpha = 10$). This noise is drawn differently for each SGD scheme. For the experiment with different dimensions, the noise level is taken as $a = 10^{-4}$.

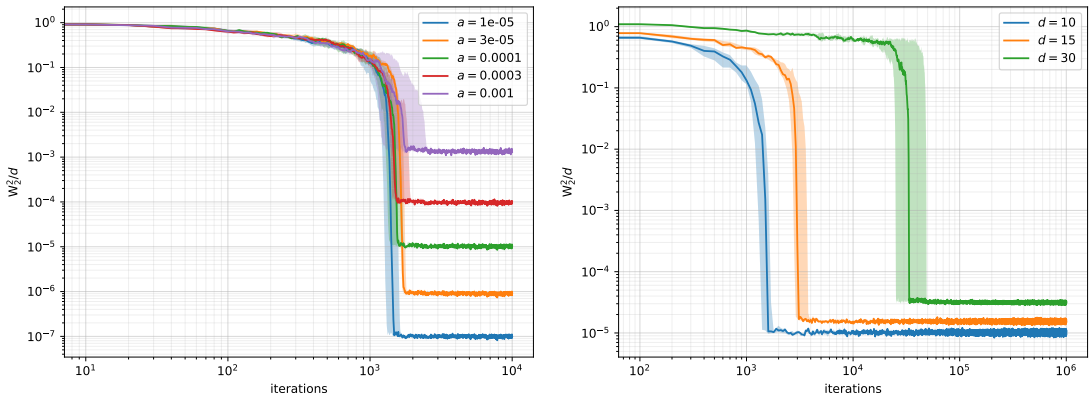


Figure A.II.14: Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for noised SGD iterations $Y^{(t)}$ on $\mathcal{E}$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. The noise is additive standard Gaussian, scaled by the learning rate $\alpha = 10$ times the noise level a. Left: different noise levels a for points in dimension $d = 10$. Right: different dimensions with $a = 10^{-4}$.

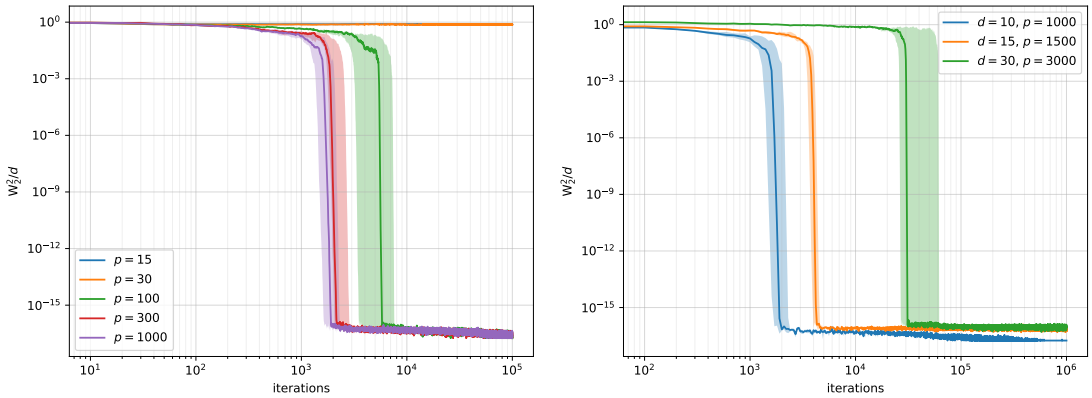


Figure A.II.15: Wasserstein error $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z)$ for SGD iterations $Y^{(t)}$ on $\mathcal{E}_p$, given a fixed measure γ_Z, $\mathbb{R}^{n\times d}$ with $n = 20$ points. The p projections in $\mathcal{E}_p$ are drawn randomly for each sample. Left: different noise levels a for points in dimension $d = 10$. Right: different dimensions with $\alpha = 10$.

The noised SGD scheme errors oscillate around a certain level which depends on the noise level, as the trajectories from Fig. A.II.10 suggest: we observed Brownian-like motion around the target points. Note that the error begins falling drastically past the same iteration threshold, albeit with a higher variance across samples for higher noise levels. At a fixed noise level, the final $\frac{1}{d}\mathrm{W}_2^2$ still depends on the noise level, despite the $1/d$ normalisation. Empirically, the final W_2^2 error seems to be smaller than the noise level a, which is reassuring since the noise is entry-wise of law $\mathcal{N}(0, a^2\alpha^2)$, where α is the learning rate.

SGD on $\mathcal{E}_p$. Fig. A.II.15 also illustrates the Wasserstein error along iterations but this time for $\mathcal{E}_p$. The general SGD setup and initial measure γ_Z remain unchanged compared to the schemes on $\mathcal{E}$ (with also a learning rate of $\alpha = 10$ in particular). In order to handle the projections $(\theta^{(t)})$, for each sample we draw p independent projections $(\theta_1, \cdots, \theta_p)$, then select the $(\theta^{(t)})$ by drawing uniformly amongst these p projections. Given this sequence of projections $(\theta^{(t)})$, the SGD algorithm is then exactly the same as for $\mathcal{E}$.

For SGD schemes on $\mathcal{E}_p$ with small values of projections p, we do not have convergence to $\gamma_{Y^{(t)}} = \gamma_Z$. Intuitively, this could be understood as the approximation $\mathcal{E}_p \approx \mathcal{E}$ being too rough, allowing for an excessive amount of numerically attainable strict local optima. This is illustrated in Fig. A.II.2 in a simple case: with $p = 3$ in dimension 2, the landscape presents numerous strict local optima that lie within large basins. However, it is notable that for p large enough ($p \geq 10d = 100$), we *do* observe convergence to $\gamma_{Y^{(t)}} = \gamma_Z$ up to numerical precision. This

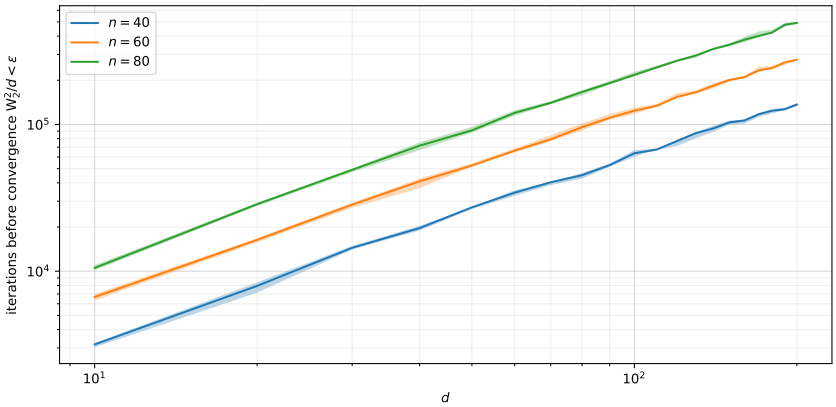


Figure A.II.16: Amount of iterations required for convergence of a SGD scheme $Y^{(t)}$ of learning rate $\alpha = 10$ on $\mathcal{E}$. Here convergence is defined as the first step t such that $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z) < 10^{-5}$. For each set of parameters (number of points n and dimension d), 10 trials are done with γ_Z, $Z \in \mathbb{R}^{n\times d}$ drawn at random (uniform $[0,1]^{n\times d}$).

convergence happens in fewer iterations as p increases, and with a smaller variance with respect to the projection samples. This suggests a stronger mode of convergence of $\mathcal{E}_p$ towards $\mathcal{E}$, as hinted at before in Fig. A.II.2 and Fig. A.II.12.

Quantifying the impact of the dimension. For different values of the number of points n and the dimension d, we run 10 samples of SGD on $\mathcal{E}$ for an original measure γ_Z drawn with standard Gaussian entries (re-drawn for each sample this time). The SGD schemes are done without additive noise, and with a learning rate of $\alpha = 10$. In order to save computation time, the SGD stopping threshold is taken as $\beta = 10^{-5}$ (see Algorithm A.II.2). For each sample, the initial position $Y^{(0)}$ is drawn with entries that are uniform on $[0,1]$. Our goal is to estimate the number of iterations required for the convergence of the SGD schemes: to this end, we define convergence as the first step t such that $\frac{1}{d}\mathrm{W}_2^2(\gamma_{Y^{(t)}}, \gamma_Z) < 10^{-5}$.

Fig. A.II.16 (cautiously) suggests that the number of iterations required for convergence its proportional to $d^{1.25}$ (where convergence means that $\frac{1}{d}\mathrm{W}_2^2$ falls below ε). Note that the exponent on d does not seem to depend on n. Obviously, the factor in front of $d^{1.25}$ depends on the number of points n, the learning rate α and the convergence threshold ε. This superlinear rule remains fairly prohibitive for large Machine Learning models, which can typically have d and n both in excess of 10^6.

A.II.6 Conclusion and Outlook

Throughout this chapter, we have investigated the properties of the Sliced Wasserstein (SW) distance between discrete measures, namely the function $\mathcal{E} : Y \longmapsto \mathrm{SW}_2^2(\gamma_Y, \gamma_Z)$, where Y and Z are supports with n points in dimension d. Due to the intractability of the expectation in $\mathcal{E}$, we introduced its Monte-Carlo empirical counterpart $\mathcal{E}_p$, computed as an average over p directions. In Section A.II.2, we showed and reminded regularity results on $\mathcal{E}$ and $\mathcal{E}_p$: they are locally-Lipschitz and differentiable on certain open sets of full measure. Leveraging the fact that $\mathcal{E}_p$ is piece-wise quadratic, we showed additional regularity results, and finally showed that the convergence of $\mathcal{E}_p$ to $\mathcal{E}$ (as $p \longrightarrow +\infty$) is almost-surely uniform on any fixed compact. Section A.II.3 furthers the study of the optimisation landscapes at hand by presenting properties of the critical points of $\mathcal{E}$ and $\mathcal{E}_p$ (points of differentiability will null gradient), and a convergence of such points of $\mathcal{E}_p$ to those of $\mathcal{E}$ as $p \longrightarrow +\infty$ (in a certain sense). In Section A.II.4, we put these theoretical results in a more practical context by showing that one can apply the SGD convergence results of [BHS22] to our optimisation landscapes. Finally, we illustrate and study these convergence results in Section A.II.5 through numerical experiments.

Further work would be welcome on the cells of $\mathcal{E}_p$ (see Section A.II.2.3), in particular the law

of their size given a fixed configuration $\mathbf{m}$ and their probability of being stable are still open problems, and would have strong consequences in practical applications such as the convergence of BCD (Algorithm A.II.1). The main difficulty stems from the link between statistical properties of the cells to the so-called Gaussian Orthant Probabilities, which can be broadly defined as the probability of a non-standard Gaussian Vector to be in the positive quadrant $\mathbb{R}^d_+$. This probability is unfortunately not tractable in high dimensions, and its estimation is a field of research in itself [AG18].

Another core limitation of our work concerns the practicality of our results on SGD convergence (Section A.II.4). Firstly, typical applications use more advanced optimisation methods, such as SGD with momentum or ADAM, which our theory does not encompass yet. Secondly, as mentioned in the introduction, practical applications actually minimise *through* $\mathcal{E}$, which is to say a loss function $F : u \longmapsto \mathrm{SW}_2^2(T_u\#\mu, \nu)$ with respect to the parameters u of a model $x \longmapsto T_u(x)$ of the input data $x \sim \mu$. Minimising F through SGD (stochastically on the projections $\theta \sim \mathfrak{o}$, the input data $x \sim \mu$ and the true data $y \sim \nu$) is beyond the scope of this chapter, and we leave this generalisation for future work.

Acknowledgements

We thank Anna Korba and Quentin Mérigot for their helpful discussions and comments on the optimisation results. We would also like to thank Antoine Chambaz for his valuable assistance with empirical processes. Finally, we are grateful for the numerous suggestions of anonymous reviewers that substantially improved this chapter.

This research was funded, in part, by the Agence nationale de la recherche (ANR), through the SOCOT project (ANR-23-CE40-0017), and the PEPR PDE-AI project (ANR-23-PEIA-0004).

A.II.7 Appendix

A.II.7.1 Proof of the Central Limit Theorem for Discrete SW

Let $c_{\mathcal{K}} := \sup_{Y\in\mathcal{K}} \|Y\|_{\infty,2}$. Consider the class of functions $\mathcal{F} := \{f_Y |\ Y \in \mathcal{K}\}$, where we define $f_Y := \theta \longmapsto \mathrm{W}_2^2(P_\theta\#\gamma_Y, P_\theta\#\gamma_Z)$ to fit the notation style of empirical processes. By Proposition A.II.1, we have for any $\theta \in \mathbb{S}^{d-1}$ and $Y, Y' \in \mathcal{K}$:

$$|f_Y(\theta) - f_{Y'}(\theta)| \leq 2n(2c_{\mathcal{K}} + c_{\mathcal{K}} + \|Z\|_{\infty,2})\|Y - Y'\|_{\infty,2},$$

where we chose the neighbourhood $B_{\|\cdot\|_{\infty,2}}(0, 2c_{\mathcal{K}}) \supset \mathcal{K}$. In particular, the Lipschitz constant $\kappa := 2n(3c_{\mathcal{K}} + \|Z\|_{\infty,2})$ is $\mathfrak{o}$-integrable (in this case, it is constant in θ), which allows us to apply [Van00, Example 19.7], which implies that the family $\mathcal{F}$ is $\mathfrak{o}$-Donsker.

To recall the definition of the property that $\mathcal{F}$ is $\mathfrak{o}$-Donsker, we recall some standard concepts and notations from empirical processes [Van00, Section 19.2], within our specific case of application. For f a function from $\mathbb{S}^{d-1}$ to $\mathbb{R}$ that is $\mathfrak{o}$-square-integrable, we introduce the real random variable

$$\mathfrak{o}_p f := \frac{1}{p}\sum_{i=1}^{p} f(\theta_i),\ (\theta_1, \cdots, \theta_p) \sim \mathfrak{o}^{\otimes p},$$

which is meant as a Monte-Carlo approximation of the true expectation $\mathfrak{o}f := \int_{\mathbb{S}^{d-1}} f(\theta)\mathrm{d}\mathfrak{o}(\theta)$. We shall study the distribution of the scaled approximation error, defined as the real random variable $\mathbb{G}_p f := \sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f)$. The central limit theorem shows that $\mathbb{G}_p f$ converges in law to a centred Gaussian (in our case the functions $f \in \mathcal{F}$ are trivially $\mathfrak{o} - L^2$), and our goal is instead to show the uniform convergence of the *random process*

$$\mathbb{G}_p := \{\sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f),\ f \in \mathcal{F}\}.$$

To study the convergence of the sequence of processes $(\mathbb{G}_p)_{p\in\mathbb{N}^*}$, we introduce the space $\ell^\infty(\mathcal{F})$ of the bounded functions $z : \mathcal{F} \longrightarrow \mathbb{R}$, equipped with the norm $\|z\|_\infty = \sup_{f\in\mathcal{F}}(|z(f)|)$. The class of functions $\mathcal{F}$ is said to be $\mathfrak{o}$-Donsker if each process $\mathbb{G}_p$ is in $\ell^\infty(\mathcal{F})$ (which is to say that it has bounded trajectories), and if the sequence $(\mathbb{G}_p)_{p\in\mathbb{N}^*}$ converges in law (in the sense

of the topology of $\ell^\infty(\mathcal{F})$) towards a tight[3] process $\mathbb{G}$. Due to the usual multivariate Central Limit Theorem, this process $\mathbb{G}$ is necessarily the $\mathfrak{o}$-*Brownian Bridge*, which is to say the centred Gaussian process indexed on $\mathcal{F}$ such that $\mathrm{Cov}[\mathbb{G}f, \mathbb{G}f'] = \mathfrak{o}(ff') - (\mathfrak{o}f)(\mathfrak{o}f')$.

We have now shown that our error process $\mathbb{G}_p$ converges in law (in $\ell^\infty(\mathcal{F})$) towards the Gaussian process $\mathbb{G}$, i.e.

$$\{\sqrt{p}(\mathfrak{o}_p f - \mathfrak{o}f),\ f \in \mathcal{F}\} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathcal{F})} \mathbb{G}. \tag{A.II.38}$$

Seeing our energy $\mathcal{E}_p : Y \longmapsto \mathfrak{o}_p f_Y$ as a process on $\mathcal{K}$, we can identify the indexes $\{f_Y,\ Y \in \mathcal{K}\}$ with $\mathcal{K}$ itself, yielding a convergence of processes on $\mathcal{K}$. To be precise, we define the space $\ell^\infty(\mathcal{K})$ as the space of bounded functions $z : \mathcal{K} \longrightarrow \mathbb{R}$, equipped with the infinite norm. Consider the map

$$\phi := \left\{ \begin{array}{ccc} \ell^\infty(\mathcal{F}) & \longrightarrow & \ell^\infty(\mathcal{K}) \\ z & \longmapsto & \left\{ \begin{array}{ccc} \mathcal{K} & \longrightarrow & \mathbb{R} \\ Y & \longmapsto & z(f_Y) \end{array} \right. \end{array} \right. ,$$

since $\mathcal{F} := \{f_Y |\ Y \in \mathcal{K}\}$, it is well defined, and it is continuous, since it is linear and verifies for any $z \in \ell^\infty(\mathcal{F})$, $\|\phi(z)\|_{\ell^\infty(K)} = \sup_{Y\in\mathcal{K}} |z(f_Y)| = \sup_{f\in\mathcal{F}} |z(f)| = \|z\|_{\ell^\infty(\mathcal{F})}$. By the continuous mapping theorem (see [Van00, Theorem 18.11] for its use on stochastic processes), we can apply ϕ to the convergence in law in Eq. (A.II.38), yielding

$$\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p\longrightarrow+\infty]{\mathcal{L},\, \ell^\infty(\mathcal{K})} \phi(\mathbb{G}), \tag{A.II.39}$$

where the process $\phi(\mathbb{G})$ is the centred Gaussian process on $\mathcal{K}$ with the covariance structure $\mathrm{Cov}\phi(\mathbb{G})[Y, Y'] = \mathfrak{o}(f_Y f_{Y'}) - (\mathfrak{o}(f_Y))(\mathfrak{o}(f_{Y'}))$. Note that $\mathcal{E}(Y) = \mathfrak{o}f_Y$ with our notations.

With the continuous mapping theorem (again [Van00, Theorem 18.11]), we can apply the continuous map $\|\cdot\|_{\ell^\infty(\mathcal{K})} : \ell^\infty(\mathcal{K}) \longrightarrow \mathbb{R}$, yielding a uniform convergence result

$$\sqrt{p}\|\mathcal{E}_p - \mathcal{E}\|_{\ell^\infty(\mathcal{K})} \xrightarrow[p\longrightarrow+\infty]{\mathcal{L}} \|\phi(\mathbb{G})\|_{\ell^\infty(\mathcal{K})}. \tag{A.II.40}$$

A.II.7.2 Computing $\mathcal{E}$, W_2^2 and $\mathcal{E}_p$ in a simple case

Computing $\mathcal{E}$. We work in polar coordinates, writing

$$\theta = \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}, \text{ and } y = \begin{pmatrix} u \\ v \end{pmatrix} = r \begin{pmatrix} \cos\psi \\ \sin\psi \end{pmatrix}.$$

By symmetry of the problem, we can assume $\psi \in [0, \pi/2]$ (i.e. the top-right quadrant $u \geq 0, v \geq 0$). Now let $\psi \in [0, 2\pi[$, let us compute $\mathrm{W}_2^2(P_\theta\#\gamma_Y, P_\theta\#\gamma_Z)$. Since we project in 1D, computing this slice amounts to sorting $(\theta^\top y_1, \theta^\top y_2)$ and $(\theta^\top z_1, \theta^\top z_2)$. Let $\tau_Z^\theta \in \mathfrak{S}_2$ such that $\theta^\top y_{\tau_Y^\theta(1)} \leq \theta^\top y_{\tau_Y^\theta(2)}$ and similarly $\theta^\top z_{\tau_Z^\theta(1)} \leq \theta^\top z_{\tau_Z^\theta(2)}$. We always have

$$\mathrm{W}_2^2(P_\theta\#\gamma_Y, P_\theta\#\gamma_Z) = \frac{1}{2}\left(\left(\theta^\top(y_{\tau_Y^\theta(1)} - z_{\tau_Z^\theta(1)})\right)^2 + \left(\theta^\top(y_{\tau_Y^\theta(2)} - z_{\tau_Z^\theta(2)})\right)^2\right).$$

We split the integral depending on the values of τ_Y^θ and τ_Z^θ, which vary depending on the angle of the projection ϕ. We begin with τ_Y^θ:

$$\theta^\top y_1 \geq \theta^\top y_2 \Longleftrightarrow \cos\phi\cos\psi + \sin\phi\sin\psi \geq 0 \Longleftrightarrow \phi \in [\psi - \pi/2, \psi + \psi/2] + 2\pi\mathbb{Z}. \tag{A.II.41}$$

The equation for τ_Z^θ is much simpler:

$$\theta^\top z_1 \geq \theta^\top z_2 \Longleftrightarrow -\sin\phi \geq 0 \Longleftrightarrow \phi \in [\pi, 2\pi] + 2\pi\mathbb{Z}. \tag{A.II.42}$$

[3] Since this technical notion is not essential for our result, we refer to [Van00, page 260] for a complete presentation.

We divide a period of 2π in four quadrants corresponding to the four possibilities for $(\tau_Y^\theta, \tau_Z^\theta)$. Since we assume $\psi \in [0, \pi/2]$, we can write this simply as:

$$\begin{aligned}\mathcal{E}(Y) = &\frac{1}{4\pi}\int_{-\pi}^{\psi-\pi/2}\left(\left(\theta^\top(y_1-z_2)\right)^2+\left(\theta^\top(y_2-z_1)\right)^2\right)\mathrm{d}\phi\\ &+\frac{1}{4\pi}\int_{\psi-\pi/2}^{0}\left(\left(\theta^\top(y_2-z_2)\right)^2+\left(\theta^\top(y_1-z_1)\right)^2\right)\mathrm{d}\phi\\ &+\frac{1}{4\pi}\int_{0}^{\psi+\pi/2}\left(\left(\theta^\top(y_2-z_1)\right)^2+\left(\theta^\top(y_1-z_2)\right)^2\right)\mathrm{d}\phi\\ &+\frac{1}{4\pi}\int_{\psi+\pi/2}^{\pi}\left(\left(\theta^\top(y_1-z_1)\right)^2+\left(\theta^\top(y_2-z_2)\right)^2\right)\mathrm{d}\phi.\end{aligned}$$

Elementary trigonometric integration yields

$$\mathcal{E}(Y) = \frac{r^2}{2}+\frac{1}{2}-\frac{2}{\pi}\left(r\cos\psi + r\psi\sin\psi\right) = \frac{u^2+v^2}{2}+\frac{1}{2}-\frac{2}{\pi}\left(u+v\operatorname{Arctan}\frac{v}{u}\right), \tag{A.II.43}$$

which holds for $\psi \in [0, \pi/2]$. By symmetry, we obtain the following expression for any $(u, v) \in \mathbb{R}^2$ (recall that we stack the vectors in Y line by line):

$$\mathcal{E}\begin{pmatrix} u & v \\ -u & -v\end{pmatrix} = \frac{u^2+v^2}{2}+\frac{1}{2}-\frac{2}{\pi}\left(|u|+|v|\operatorname{Arctan}\left|\frac{v}{u}\right|\right). \tag{A.II.44}$$

In the general case, dimension d would require the use of d-dimensional spherical coordinates, making the equations Eq. (A.II.41) and Eq. (A.II.42) intractable. Furthermore, generalising to n points would separate the integral into $(n!)^2$ parts, losing all hopes of tractability and legibility.

Computing W_2^2. In the case $n = 2$, the Kantorovich LP formulation of the Wasserstein distance can be written as:

$$\min_{a\in[0,1]}\sum_{k,l\in[\![1,2]\!]}\pi_{k,l}(a)\|y_k - z_l\|_2^2, \quad \text{with } \pi(a) := \frac{1}{2}\begin{pmatrix}1-a & a\\ a & 1-a\end{pmatrix}.$$

Substituting $y_1 = \begin{pmatrix}u\\v\end{pmatrix}$, $y_2 = \begin{pmatrix}-u\\-v\end{pmatrix}$, $z_1 = \begin{pmatrix}0\\-1\end{pmatrix}$, $z_2 = \begin{pmatrix}0\\1\end{pmatrix}$ yields:

$$\mathrm{W}_2^2(\gamma_Y, \gamma_Z) = \min_{a\in[0,1]}\left(u^2+(v+1)^2-4av\right) = u^2+(|v|-1)^2.$$

Computing $\mathcal{E}_p$. For simplicity, in the following we will only consider $\theta \in \mathbb{S}^{d-1}$ such that the $\theta^\top y_k$ are distinct, and such that the $\theta^\top z_k$ are also distinct. We will express the cases for the values of the sortings τ_Y^θ and τ_Z^θ in a different (yet equivalent) manner.

We have $\tau_Y^\theta = I$ if $\theta^\top y_1 < \theta^\top y_2$ and $\tau_Y^\theta = (2, 1)$ otherwise. Then $\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I$ if $\tau_Y^\theta = \tau_Z^\theta$, and $\tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1} = (2, 1)$ otherwise. The system

$$\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I \iff \begin{cases}\theta^\top y_1 < \theta^\top y_2 \text{ and } \theta^\top z_1 < \theta^\top z_2\\ \text{or}\\ \theta^\top y_2 < \theta^\top y_1 \text{ and } \theta^\top z_2 < \theta^\top z_1\end{cases}$$

can be simplified, yielding:

$$\tau_Z^\theta \circ (\tau_Y^\theta)^{-1} = I \iff \left(\theta\theta^\top(z_2-z_1)\right)^\top(y_2-y_1) > 0. \tag{A.II.45}$$

Eq. (A.II.45) is a linear equation in Y. Additionally, Eq. (A.II.45) only depends on $y_2 - y_1 = -2y$, which makes our symmetrical simplification inconsequential. Plugging in the specific point

values yields a more explicit definition of the cells. We write the condition on $y \in \mathbb{R}^2$, since $Y = (y, -y)^\top$.

$$\mathcal{C}_{\mathbf{m}} = \left\{ y \in \mathbb{R}^2 \mid \forall i \in [\![1, p]\!],\ -\operatorname{sign}\left[\theta_i^\top \begin{pmatrix} 0 \\ 1 \end{pmatrix} \theta_i^\top y \right] = +1 \text{ if } \mathbf{m}_i = I,\ \text{else} -1 \right\}. \tag{A.II.46}$$

Equation Eq. (A.II.46) describes $\mathcal{C}_{\mathbf{m}}$ as an intersection of p half-planes of $\mathbb{R}^2$, thus it is a polytope. Note that we use strict inequalities, which lifts configuration ambiguities, and implies that the $(\mathcal{C}_{\mathbf{m}})_{\mathbf{m} \in \mathfrak{S}_2^p}$ are disjoint, and that the union of their closure is $\mathbb{R}^2$.

Straightforward computation yields

$$\operatorname*{argmin}_{X \in \mathbb{R}^{n \times d}} q_{\mathbf{m}}(X) = (A^{-1}(B^{\mathbf{m}}_{1,1} z_1 + B^{\mathbf{m}}_{1,2} z_2), A^{-1}(B^{\mathbf{m}}_{2,1} z_1 + B^{\mathbf{m}}_{2,2} z_2)),$$

where $A := \frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top$ and $B^{\mathbf{m}}_{k,l} := \frac{1}{p} \sum_{\substack{i=1 \\ \mathbf{m}_i(k) = l}}^{p} \theta_i \theta_i^\top$.

Note that our $n = 2$ setting, we have the simplifications $B^{\mathbf{m}}_{1,2} = A - B^{\mathbf{m}}_{1,1}$, $B^{\mathbf{m}}_{2,1} = B^{\mathbf{m}}_{1,2}$ and $B^{\mathbf{m}}_{1,1} = B^{\mathbf{m}}_{2,2}$. Furthermore, $B^{\mathbf{m}}_{k,l}$ is (up to a factor), a Monte-Carlo estimation of $S^{Y,Z}_{k,l}$ (see Corollary A.II.1).

A.II.7.3 Discrete Wasserstein stability

Consider the following generic discrete Kantorovich problem, given weights $\alpha \in \triangle_n$ and $\beta \in \triangle_m$ and a generic cost matrix $C \in \mathbb{R}_+^{n \times m}$:

$$\mathrm{W}(\alpha, \beta; C) := \inf_{\pi \in \Pi(\alpha, \beta)} \pi \cdot C, \tag{A.II.47}$$

where $\Pi(\alpha, \beta, C)$ is the set of $n \times m$ matrices π with non-negative entries such that $\pi \mathbb{1} = \alpha$ and $\pi^\top \mathbb{1} = \beta$.

Lemma A.II.2 (Stability of the Wasserstein cost)**.** Let $\alpha, \overline{\alpha}, \in \triangle_n$, $\beta, \overline{\beta} \in \triangle_m$ and $C, \overline{C} \in \mathbb{R}_+^{n \times m}$, such that the weights verify $\alpha, \overline{\alpha}, \beta, \overline{\beta} > 0$ entry-wise. Then:

$$\left| \mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C}) \right| \leq \|C - \overline{C}\|_\infty + \|C\|_\infty (\|\alpha - \overline{\alpha}\|_1 + \|\beta - \overline{\beta}\|_1). \tag{A.II.48}$$

$$\left| \mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C}) \right| \leq \|C - \overline{C}\|_F + \|C\|_F \left(\|\alpha - \overline{\alpha}\|_2 + \|\beta - \overline{\beta}\|_2 \right), \tag{A.II.49}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

Note that this result is a generalisation of [BBN24, Theorem 2] (they assumes that the cost matrices are pairwise distances, which amount to the W_1 case), but requires the weights to have positive entries (as opposed to non-negative entries).

Proof. We split the difference in two terms:

$$\begin{aligned} \left| \mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C}) \right| \leq & \left| \mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\alpha, \beta; \overline{C}) \right| =: \mathrm{I} \\ & + \left| \mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; C) \right| =: \mathrm{II} \end{aligned}$$

— *Step 1*: Controlling I Using the primal formulation

We use Equation Eq. (A.II.47): let $\overline{\pi}^*$ optimal for $\mathrm{W}(\alpha, \beta, \overline{C})$. In particular, $\overline{\pi}^*$ is admissible

for the problem $\mathrm{W}(\alpha, \beta, C)$. We have:

$$\begin{aligned}\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\alpha,\beta;\overline{C}) &= \min_{\pi\in\Pi(\alpha,\beta)} \pi\cdot C - \min_{\pi\in\Pi(\alpha,\beta)} \pi\cdot\overline{C} \\ &\leq \overline{\pi}^*\cdot C - \overline{\pi}^*\cdot\overline{C} = \sum_{i=1}^n\sum_{j=1}^m \overline{\pi}^*_{i,j}(C_{i,j}-\overline{C}_{i,j}) \\ &\leq \|C-\overline{C}\|_\infty \sum_{i=1}^n\sum_{j=1}^m \overline{\pi}^*_{i,j} = \|C-\overline{C}\|_\infty,\end{aligned}$$

where the property $\pi\in\Pi(\alpha,\beta)$ implied $\sum_{i,j}\overline{\pi}^*_{i,j}=1$. By using the same argument symmetrically, we obtain

$$\mathrm{I} \leq \|C-\overline{C}\|_\infty.$$

— *Step 2*: Controlling the dual variables

Consider the Legendre dual problem associated to Eq. (A.II.47):

$$\mathrm{W}(\alpha,\beta;C) = \sup_{\substack{f\in\mathbb{R}^n, g\in\mathbb{R}^m \\ f\oplus g\leq C}} f^\top\alpha + g^\top\beta. \tag{A.II.50}$$

Let f^*, g^* optimal for the dual formulation, our objective is to bound this dual solution in a set which depends on C. First, notice that the value and constraints remain unchanged if we replace (f^*, g^*) with $(f^* - t\mathbb{1}, g^* + t\mathbb{1})$ for $t\in\mathbb{R}$, which allows us to assume $f^*_1 = 0$. We now leverage the complementary slackness property (which characterises the primal-dual optimality conditions for this linear problem, see [PC19b, Section 3.3]): for any π^* optimal for the primal problem Eq. (A.II.47), we have the implication

$$\pi^*_{i,j} \neq 0 \Longrightarrow f^*_i + g^*_j = C_{i,j}.$$

The primal constraints imply that $\sum_j \pi^*_{1,j} = \alpha_1 > 0$ and that $\pi^*\geq 0$ entry-wise, there exists $j_1\in[\![1,m]\!]$ such that $\pi^*_{1,j_1}\neq 0$. Using the complementary slackness implication, we obtain $0 + g^*_{j_1} = C_{1,j_1}$. We now use the dual constraint $f^*\oplus g^*\leq C$ at $i=1$ to show that $\forall j\in[\![1,m]\!]$, $g^*_j\leq C_{1,j}$. This allows us to find a lower-bound on f^*: since $\forall i\in[\![2,n]\!]$, $\sum_j\pi^*_{i,j}=\alpha_i>0$, thus there exists a $j_i\in[\![1,m]\!]$ such that $\pi^*_{i,j_i}\neq 0$, yielding $f^*_i + g^*_{j_i} = C_{i,j_i}$, then since $g^*_{j_i}\leq C_{1,j_i}$, this yields the lower-bound $f^*_i\geq C_{i,j_i} - C_{1,j_i}$. For an upper-bound on f^*, we use the dual constraint at (i,j_1): we have $f^*_i + g^*_{j_1}\leq C_{i,j_1}$, then we use $g^*_{j_1} = C_{1,j_1}$ proved earlier to show $f^*_i\leq C_{i,j_1} - C_{1,j_1}$. At this point, we have the following control on f^*_i:

$$f^*_1 = 0, \quad \forall i\in[\![2,n]\!],\ C_{i,j_i} - C_{1,j_i} \leq f^*_i \leq C_{i,j_1} - C_{1,j_1}.$$

Regarding g^*, we already have $\forall j\in[\![1,m]\!]$, $g_j\leq C_{1,j}$. For a lower bound, since $\sum_i\pi_{i,j}=\beta_j>0$, there exists $i_j\in[\![1,n]\!]$ such that $\pi^*_{i_j,j}\neq 0$, so by complementary slackness $f^*_{i_j} + g^*_j = C_{i_j,j}$, thus by the upper-bound on f^* we have if $i_j\neq 1$ that $g^*_j\geq C_{i_j,j} - C_{i_j,j_1} + C_{1,j_1}$. If $i_j = 1$ then $f^*_{i_j} = 0$ and $g^*_j = C_{i_1,j}$. Our control on g^* is the following:

$$\forall j\in[\![1,m]\!], \begin{cases} C_{i_j,j} - C_{i_j,j_1} + C_{1,j_1} \leq g^*_j \leq C_{1,j} & \text{if } i_j\neq 1 \\ g^*_j = C_{i_1,j} & \text{if } i_j = 1 \end{cases}.$$

We summarise our bounds in the following (weaker) statement, which holds thanks to the condition $C\geq 0$ (entry-wise):

$$\|f^*\|_\infty \leq \|C\|_\infty, \quad \|g^*\|_\infty\leq\|C\|_\infty.$$

— *Step 3*: Bounding II using the dual formulation

Let f^*, g^* optimal for the dual formulation Eq. (A.II.50) of $\mathrm{W}(\alpha,\beta,C)$, which by Step 2 we can choose to verify $\|f^*\|_\infty\leq\|C\|_\infty$ and $\|g^*\|_\infty\leq\|C\|_\infty$. In particular, (f^*, g^*) is admissible

for the dual formulation of $\mathrm{W}(\overline{\alpha}, \overline{\beta}, C)$.

$$\begin{aligned}\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};C) &= \max_{f\oplus g\le C} f^\top\alpha + g^\top\beta - \max_{\overline{f}\oplus\overline{g}\le C} \overline{f}^\top\overline{\alpha} + \overline{g}^\top\overline{\beta} \\ &\le (f^*)^\top\alpha + (g^*)^\top\beta - (f^*)^\top\overline{\alpha} - (g^*)^\top\overline{\beta} \\ &= (f^*)^\top(\alpha-\overline{\alpha}) + (g^*)^\top(\beta-\overline{\beta}) \\ &\le \|f^*\|_\infty\|\alpha-\overline{\alpha}\|_1 + \|g^*\|_\infty\|\beta-\overline{\beta}\|_1 \\ &\le \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1).\end{aligned}$$

By symmetry, we obtain $|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};C)| \le \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1)$.

— *Step 4*: Wrapping up

By Step 1 and Step 3 combined we conclude:

$$\left|\mathrm{W}(\alpha,\beta;C) - \mathrm{W}(\overline{\alpha},\overline{\beta};\overline{C})\right| \le \mathrm{I} + \mathrm{II} \le \|C-\overline{C}\|_\infty + \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1).$$

— Details for the proof of Eq. (A.II.49)

For the first term, to get the Frobenius norm $\|C-\overline{C}\|_F$ instead of the infinite norm, it suffices to use that $\|M\|_\infty \le \|M\|_F$.

For the second term, note that the penultimate inequality of Step 3 can also be written with the Cauchy-Schwarz inequality, yielding $\|C\|_F(\|\alpha-\overline{\alpha}\|_2 + \|\beta-\overline{\beta}\|_2)$, where the upper-bound on $\|f^*\|_2$ and $\|g^*\|_2$ by $\|C\|_F$ are obtained using the element-wise bounds on f^* and g^* from Step 2. □

A.II.7.4 Proof of Theorem A.II.7 and convergence rate

The proof of Theorem A.II.7 requires matrix concentration technicalities. In the following, $\|\cdot\|_{\mathrm{op}}$ denotes the $\|\cdot\|_2$-induced operator norm on $\mathbb{R}^{d\times d}$, and $S_d(\mathbb{R})$ denotes the space of symmetric $d\times d$ matrices. We write $\preceq$ for the Loewner order of positive semi-definite symmetric matrices ($A \preceq B$ means that $B-A$ is positive semi-definite). We recall the following Hoeffding inequality.

Theorem A.II.11. Matrix Hoeffding Inequality [Tro12, Theorem 1.3]
Let $q \in \mathbb{N}^*$, $(X_i)_{i\in[\![1,q]\!]}$ independent random variables with values in $S_d(\mathbb{R})$, such that $\mathbb{E}[X_i] = 0$. Suppose that $\forall i \in [\![1,q]\!]$, $\exists A_i \in S_d(\mathbb{R}) : X_i^2 \preceq A_i^2$. Let $\sigma^2 := \left\|\sum_i A_i^2\right\|_{\mathrm{op}}$, then for any $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^q X_i\right\|_{\mathrm{op}} \ge t\right) \le d\exp\left(-\frac{t^2}{8\sigma^2}\right).$$

We deduce from Theorem A.II.11 the following lemma, where the X_i follow a uniform law on $\Theta \subset \mathbb{S}^{d-1}$.

Lemma A.II.3 (Hoeffding applied to $\theta \sim \mathcal{U}(\Theta)$)**.**
Let $(\theta_i)_{i\in[\![1,q]\!]}$, independent random vectors following the uniform law on $\Theta \subset \mathbb{S}^{d-1}$, where Θ is $\mathfrak{o}$-measurable with $\mathfrak{o}(\Theta) > 0$. Let $S_\Theta := \frac{1}{s_\Theta}\int_\Theta \theta\theta^\top \mathrm{d}\mathfrak{o}(\theta)$, where $s_\Theta := \mathfrak{o}(\Theta)$. S_Θ is the covariance matrix of $\theta \sim \mathcal{U}(\Theta)$. Let $\eta \in]0,1[$ and $t > 0$. Then with probability exceeding $1-\eta$ we have

$$q \ge \frac{32\log(d/\eta)}{t^2} \Longrightarrow \left\|\frac{1}{q}\sum_{i=1}^q \theta_i\theta_i^\top - S_\Theta\right\|_{\mathrm{op}} \le t.$$

In the case $\Theta = \mathbb{S}^{d-1}$, the condition $q \ge \dfrac{8\log(d/\eta)}{t^2}$ is sufficient.

Proof. The idea is to apply Theorem A.II.11 to $X_i := \frac{1}{q}\theta_i\theta_i^\top - \frac{1}{q}S_\Theta$. First, by definition, $\mathbb{E}[X_i] = 0$.

We now find $A \in S_d^+(\mathbb{R})$ such that $X_i^2 \preceq A$. Let $u \in \mathbb{S}^{d-1}$, we compute:

$$u^\top X_i^2 u = \frac{1}{q^2}\left(u^\top \theta_i\theta_i^\top u - u^\top \theta_i\theta_i^\top S_\Theta u - u^\top S_\Theta \theta_i\theta_i^\top u + u^\top S_\Theta^2 u\right) \leq \left(\frac{1+\|S_\Theta\|_{\mathrm{op}}}{q}\right)^2.$$

Moreover, $\|S_\Theta\|_{\mathrm{op}} \leq 1$, since

$$\forall u \in \mathbb{S}^{d-1},\ u^\top S_\Theta u = \frac{1}{s_\Theta}\int_\Theta u^\top \theta\theta^\top u \mathrm{d}\varpi(\theta) \leq \frac{1}{s_\Theta}\int_\Theta 1 \mathrm{d}\varpi(\theta) = 1.$$

In conclusion $X_i^2 \preceq \frac{4}{q^2} I$. Using the notations of Theorem A.II.11, we compute $\sigma^2 = 4/q$, and apply the Matrix Hoeffding inequality with $\Delta := \sum_i X_i = \frac{1}{q}\sum_i \theta_i\theta_i^\top - S_\Theta$. It follows that for any $t > 0$, $\mathbb{P}\left(\|\Delta\|_{\mathrm{op}} \geq t\right) \leq d\exp\left(-\frac{qt^2}{32}\right)$. In order to have the event $\|\Delta\|_{\mathrm{op}} \leq t$ with probability exceeding $1-\eta$, it is therefore sufficient that $\eta \geq d\exp\left(-\frac{qt^2}{32}\right)$, which is equivalent to $q \geq \frac{32\log(d/\eta)}{t^2}$.

In the case $\Theta = \mathbb{S}^{d-1}$, one has $S_\Theta = I/d$, and a finer Loewner upper-bound can be established, since

$$u^\top X_i^2 u = \frac{1}{q^2}\left(u^\top \theta_i\theta_i^\top u - \frac{2}{d}u^\top \theta_i\theta_i^\top u + \frac{1}{d^2}\right) \leq \left(\frac{1-\frac{1}{d}}{q}\right)^2 \leq \frac{1}{q^2},$$

and thus $\sigma^2 = 1/q$. This yields the Hoeffding inequality $\mathbb{P}\left(\|\Delta\|_{\mathrm{op}} \geq t\right) \leq d\exp\left(-\frac{qt^2}{8}\right)$, which in turn provides the announced weaker condition on q. □

With this tool at hand, we now prove a quantitative concentration result:

Theorem A.II.12 (Concentration of cell optima)**.**
Let $\mathbf{m} = (\sigma_1, \cdots, \sigma_p)$ be a fixed matching configuration (see Section A.II.2.3) and let $(\theta_i)_{i\in[\![1,p]\!]} \sim \mathfrak{v}^{\otimes p}$ (uniform on $\mathbb{S}^{d-1}$). We introduce the following notations and variables:

- For $(k,l) \in [\![1,n]\!]^2$, let $q_{k,l} := \#\{i \in [\![1,p]\!] \mid k = \sigma_i(l)\}$;
- Let $\bar{c}_Z := \max_{l\in[\![1,n]\!]} \|z_l\|_2$;
- Let $\varepsilon \in (0, \frac{4}{3}n\bar{c}_Z]$;
- Let $\eta \in (0,1)$.

Assume the following:

- (H_q) : $\forall (k,l) \in [\![1,n]\!]^2$, $q_{k,l} \geq \overline{q}$ or $q_{k,l} < \underline{q}$, with $1 \leq \underline{q} \leq \overline{q} \leq p$;
- (H_1) : $p \geq \dfrac{697 d^2 n^2 \bar{c}_Z^2 \log(3d/\eta)}{\varepsilon^2}$;
- (H_2) : $\overline{q} \geq \dfrac{512 d^2 \bar{c}_Z^2 \log(3dnn^+/\eta)}{\varepsilon^2}$; $\quad n^+ := \max_{k\in[\![1,n]\!]} \#\{l \in [\![1,n]\!] \mid q_{k,l} \geq \overline{q}\}$;
- (H_3) : $\underline{q} \leq \dfrac{\varepsilon}{8dn^-\bar{c}_Z} p$; $\quad n^- := \max_{k\in[\![1,n]\!]} \#\{l \in [\![1,n]\!] \mid q_{k,l} \leq \underline{q}\}$;
- (H_4) : $p \geq \dfrac{8 d^2 n^2 \bar{c}_Z^2 \log(6n^2/\eta)}{\varepsilon^2}$.

Then with probability exceeding $1-\eta$, writing $Y^* := \operatorname{argmin}_{Y'\in\mathbb{R}^{n\times d}} q_{\mathbf{m}}(Y')$, we have

$$\forall k \in [\![1,n]\!],\ \left\| y_k^* - \sum_{l=1}^n S_{k,l} z_l \right\|_2 \leq \varepsilon, \tag{A.II.51}$$

where the normalized conditional covariance matrices $S_{k,l}$ are defined in Corollary A.II.1 (we omit the Y^*, Z exponent here for legibility).

Proof. — *Step 1*: Re-writing Eq. (A.II.26).

Remind that the matching configuration $\mathbf{m}$ is fixed here. Let $Y^* := \underset{Y' \in \mathbb{R}^{n \times d}}{\operatorname{argmin}}\ q_{\mathbf{m}}(Y')$ and $k \in [\![1, n]\!]$. By Eq. (A.II.26), we have

$$y_k^* = A^{-1} \left(\frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top z_{\sigma_i(k)} \right), \text{ with } A = \frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top.$$

Let $I_{k,l} := \{ i \in [\![1, p]\!] \mid \sigma_i(k) = l \}$. Since the σ_i are permutations, we have $[\![1, p]\!] = \bigcup_{l=1}^{n} I_{k,l} = \bigcup_{k=1}^{n} I_{k,l}$ and $k \neq k' \Rightarrow I_{k,l} \cap I_{k',l} = \varnothing$; $l \neq l' \Rightarrow I_{k,l} \cap I_{k,l'} = \varnothing$. We re-order the sum:

$$\frac{1}{p} \sum_{i=1}^{p} \theta_i \theta_i^\top z_{\sigma_i(k)} = \sum_{l=1}^{n} \frac{1}{p} \sum_{i \in I_{k,l}} \theta_i \theta_i^\top z_l = \sum_{l=1}^{n} \frac{q_{k,l}}{p} B_{k,l} z_l,$$

where $q_{k,l} := \# I_{k,l}$ and $B_{k,l} := \dfrac{1}{q_{k,l}} \displaystyle\sum_{i \in I_{k,l}} \theta_i \theta_i^\top$. This invites the definition of the matrix $R = (r_{k,l})$, $r_{k,l} := \frac{q_{k,l}}{p}$, which is bi-stochastic by construction.

— *Step 2*: Separating the terms in y_k^*.

We will see later that the empirical covariance matrix A concentrates towards the covariance matrix of $\theta \sim \varpi$, which is I/d. In order to quantify the impact of this concentration on y_k^*, we introduce the error term: $\delta A^- := A^{-1} - dI$.

A similar concentration will be observed for $B_{k,l}$, but the θ_i in the sum are *selected* such that $i \in I_{k,l}$. Recall that since we project in 1D, the permutations σ_i arise from a sorting problem, namely $\sigma_i = \tau_Z^{\theta_i} \circ (\tau_Y^{\theta_i})^{-1}$, where we recall that τ_Y^θ is a permutation sorting the numbers $(y_1^\top \theta, \cdots, y_n^\top \theta)$.

By definition, we have $\sigma_i(k) = l \iff \theta_i \in \Theta_{k,l} = \left\{ \theta \in \mathbb{S}^{d-1} \mid \tau_Z^\theta \circ (\tau_Y^\theta)^{-1}(k) = l \right\}$, where we omit again the Y, Z exponent on $\Theta_{k,l}$ for legibility.

Since the θ_i in $B_{k,l}$ are drawn under the condition $\theta_i \in \Theta_{k,l}$, we study the concentration $B_{k,l} \approx C_{k,l}$, where $C_{k,l} := \frac{1}{d\varpi(\Theta_{k,l})} S_{k,l}$. In order to quantify this approximation, we define the error term $\delta B_{k,l} := B_{k,l} - C_{k,l}$. Similarly, the $r_{k,l} := \frac{q_{k,l}}{p}$ are Monte-Carlo approximations of $\varpi(\Theta_{k,l})$, which leads to the definition $\delta r_{k,l} := r_{k,l} - \varpi(\Theta_{k,l})$.

We may now separate the terms in the result from Step 1:

$$\begin{aligned}
y_k^* &= (dI + \delta A^-) \left(\sum_{l=1}^{n} r_{k,l} \underbrace{(C_{k,l} + \delta B_{k,l})}_{B_{k,l}} z_l \right) \\
&= \underbrace{d \sum_{l=1}^{n} \varpi(\Theta_{k,l}) C_{k,l} z_l}_{v} + \underbrace{\delta A^- \left(\sum_{l=1}^{n} r_{k,l} B_{k,l} z_l \right)}_{\delta v_1} + \underbrace{d \sum_{\substack{l=1 \\ q_{k,l} \geq \underline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l}_{\delta v_2} \\
&\quad + \underbrace{d \sum_{\substack{l=1 \\ q_{k,l} < \underline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l}_{\delta v_3} + \underbrace{d \sum_{l=1}^{n} \delta r_{k,l} C_{k,l} z_l}_{\delta v_4}.
\end{aligned}$$

The separation of the terms in the second equality arises from (H_q), formulated in the theorem. Observe that the first term v is exactly $\Psi(Y^*)$, with Ψ defined in Section A.II.3.1.2. Our objective is to provide conditions under which $\forall i \in \{1, 2, 3, 4\}$, $\|\delta v_i\|_2 \leq \varepsilon/4$ with probability exceeding $1 - \eta$. To that end, we let $\varepsilon > 0$ and $\eta \in (0, 1)$.

— *Step 3*: Condition for $\|\delta v_2\|_2 \leq \dfrac{\varepsilon}{4}$.

First of all, note that if the sum defining δv_2 is empty, the condition holds trivially almost-surely. In the following, we suppose that the sum has at least one non-zero term. We have from Step 2,

$$\|\delta v_2\|_2 = \left\| d \sum_{\substack{l=1 \\ q_{k,l} \geq \overline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l \right\|_2 \leq d\bar{c}_Z \sum_{\substack{l=1 \\ q_{k,l} \geq \overline{q}}}^{n} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}}.$$

Let the shorthands $n_k^+ := \#J_k^+$ and $J_k^+ := \{l \in [\![1,n]\!] \mid q_{k,l} \geq \overline{q}\}$. We upper-bound the right term by $\sum_{l \in J_k^+} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}} \leq \sum_{l \in J_k^+} r_{k,l} \max_{l \in J_k^+} \|\delta B_{k,l}\|_{\mathrm{op}} \leq \max_{l \in J_k^+} \|\delta B_{k,l}\|_{\mathrm{op}}$.

For $l \in J_k^+$, by Lemma A.II.3, we have $\|\delta B_{k,l}\|_{\mathrm{op}} \leq t$ with probability exceeding $1 - \eta/(3nn_k^+)$ provided that $q_{k,l} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}$. Since the probability of $\bigcup_{l \in J_k^+} \{\|\delta B_{k,l}\|_{\mathrm{op}} > t\}$ can be upper bounded by the sum of the probabilities of each of the n_k^+ terms, it is upper bounded by $\eta/(3n)$. Therefore, writing the event $\{\forall l \in J_k^+,\ \|\delta B_{k,l}\|_{\mathrm{op}} \leq t\}$ as the complementary of this union, we conclude that it holds with probability exceeding $1 - \eta/(3n)$, provided that

$$\forall l \in J_k^+,\ q_{k,l} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}.$$

A sufficient condition for this last assumption to hold is $(H_2^k):\ \overline{q} \geq \frac{32 \log\left(3dnn_k^+/\eta\right)}{t^2}$. Applying this result to $t := \frac{\varepsilon}{4d\bar{c}_Z}$, and by letting $n^+ := \max_{k \in [\![1,n]\!]} n_k^+$, a sufficient condition to have $\|\delta v_2\|_2 \leq \frac{\varepsilon}{4}$ with probability exceeding $1 - \eta/(3n)$ is

$$(H_2):\quad \overline{q} \geq \frac{512 d^2 \bar{c}_Z^2 \log(3dnn^+/\eta)}{\varepsilon^2}.$$

— *Step 4*: Condition for $\|\delta v_3\|_2 \leq \frac{\varepsilon}{4}$.

With a computation analogous to Step 3, we write

$$\|\delta v_3\|_2 = \left\| d \sum_{\substack{l=1 \\ q_{k,l} < \underline{q}}}^{n} r_{k,l} \delta B_{k,l} z_l \right\|_2 \leq d\bar{c}_Z \sum_{l \in J_k^-} r_{k,l} \|\delta B_{k,l}\|_{\mathrm{op}},$$

where, like in Step 3, we define $n_k^- := \#J_k^-$ and $J_k^- := \{l \in [\![1,n]\!] \mid q_{k,l} \leq \underline{q}\}$. If $n_k^- = 0$ then the objective holds almost-surely, thus we suppose $n_k^- \geq 1$. In this setting, the $q_{k,l}$ are small, thus we have little control over $\|\delta B_{k,l}\|_{\mathrm{op}}$, which can be upper bounded by 2.

Leveraging the condition $q_{k,l} \leq \underline{q}$, which holds for $l \in J_k^-$, we have $r_{k,l} = q_{k,l}/p \leq \underline{q}/p$. In order to have $\|\delta v_3\|_2 \leq \frac{\varepsilon}{4}$ almost-surely, it is sufficient to have $(H_3^k):\quad \underline{q} \leq \frac{\varepsilon}{8dn_k^- \bar{c}_Z}\, p$. Again, with $n^- := \max_{k \in [\![1,n]\!]} n_k^-$, we obtain the sufficient condition:

$$(H_3):\quad \underline{q} \leq \frac{\varepsilon}{8dn^- \bar{c}_Z}\, p.$$

— *Step 5*: Condition for $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$.

By definition, $\delta v_4 = d \sum_{l=1}^{n} \delta r_{k,l} C_{k,l} z_l$, then $\|\delta v_4\|_2 \leq \bar{c}_Z d \sum_{l=1}^{n} |\delta r_{k,l}| \|C_{k,l}\|_{\mathrm{op}}$. We use the upper-bound $\|C_{k,l}\|_{\mathrm{op}} \leq 1$ (observe that $\|C_{k,l}\|_{\mathrm{op}}$ can be made as close to 1 as desired by choosing $\Theta_{k,l}$ as a very small portion of the sphere). In order to have $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$, it is sufficient to have $\forall l \in [\![1,n]\!],\ |\delta r_{k,l}| \leq \frac{\varepsilon}{4dn\bar{c}_Z} =: t$. Our objective is to quantify the Monte-Carlo error

$$\delta r_{k,l} = \frac{\#\{i \in [\![1,p]\!] \mid \theta_i \in \Theta_{k,l}\}}{p} - \varpi(\Theta_{k,l}).$$

To that end, we fix $l \in [\![1,n]\!]$ and apply the standard Bernoulli Chernoff concentration inequality (additive form) to $X_i := \mathbb{1}(\theta_i \in \Theta_{k,l})$. By definition, $\mathbb{E}[X_i] = \varpi(\Theta_{k,l})$, hence by Chernoff

$$\mathbb{P}\left(\left|\frac{1}{p}\sum_{i=1}^{p} X_i - \varpi(\Theta_{k,l})\right| > t\right) \leq 2e^{-2pt^2}.$$

It follows that the inequality $p \geq \frac{\log(6n^2/\eta)}{2t^2}$ implies $|\delta r_{k,l}| \leq t$ with probability exceeding $1 - \frac{\eta}{3n^2}$. Substituting $t = \frac{\varepsilon}{4dn\bar{c}_Z}$ yields

$$(H_4): \; p \geq \frac{8d^2n^2\bar{c}_Z^2 \log(6n^2/\eta)}{\varepsilon^2}.$$

Using the same reasoning as in previous steps, under (H_4), the event $\{\forall l \in [\![1,n]\!], \; |\delta r_{k,l}| \leq \frac{\varepsilon}{4dn\bar{c}_Z}\}$ holds with probability exceeding $1 - \frac{\eta}{3n}$, which implies that our objective $\|\delta v_4\|_2 \leq \frac{\varepsilon}{4}$ also holds with the same probability.

— *Step 6*: Condition for $\|\delta v_1\|_2 \leq \dfrac{\varepsilon}{4}$.

We have

$$\|\delta v_1\|_2 \leq \|\delta A^-\|_{\mathrm{op}} \left\|\sum_{l=1}^{n} r_{k,l} B_{k,l} z_l\right\|_2 \leq \frac{\|\delta A^-\|_{\mathrm{op}}}{d}\left(\|v\|_2 + \|\delta v_2\|_2 + \|\delta v_3\|_2 + \|\delta v_4\|_2\right).$$

In the following, we continue conditionally on the three events "$\|\delta v_i\|_2 \leq \frac{\varepsilon}{4}$", $i \in \{2,3,4\}$, under which:

$$\|\delta v_1\|_2 \leq \frac{\|\delta A^-\|_{\mathrm{op}}}{d}\left(\|v\|_2 + \frac{3\varepsilon}{4}\right).$$

We now dominate $\|v\|_2 = \left\|\sum_{l=1}^{n} S_{k,l} z_l\right\|_2$. Recall that the $(\Theta_{k,l})_{l \in [\![1,n]\!]}$ are disjoint, with $\bigcup_{l=1}^{n} \Theta_{k,l} = \mathbb{S}^{d-1}$, which implies $\sum_{l=1}^{n} S_{k,l} = d\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\varpi(\theta) = I$. Since the $S_{k,l}$ are symmetric semi-definite, the previous equation provides $\|S_{k,l}\|_{\mathrm{op}} \leq 1$, which in turn yields $\|v\|_2 \leq n\bar{c}_Z$. Assuming $\varepsilon \leq \frac{4}{3}n\bar{c}_Z$, we get finally $\|\delta v_1\|_2 \leq \|\delta A^-\|_{\mathrm{op}} \frac{2n\bar{c}_Z}{d}$.

It is sufficient to find a condition under which $\|\delta A^-\|_{\mathrm{op}} \leq \frac{d\varepsilon}{8n\bar{c}_Z} =: t$. We cannot apply Lemma A.II.3 directly since δA^- has an inverse operation. First, $\|\delta A^-\|_{\mathrm{op}} = \|A^{-1} - dI\|_{\mathrm{op}} = \left\|d(I - d\delta A)^{-1} - dI\right\|_{\mathrm{op}}$, with $\delta A := I/d - A$. Then, assuming $(H_{\delta A}): d\|\delta A\|_{\mathrm{op}} < 1$, we use a Neumann series for the inverse:

$$\|\delta A^-\|_{\mathrm{op}} = \left\|\sum_{k=1}^{+\infty} (d\delta A)^k\right\|_{\mathrm{op}} \leq \sum_{k=1}^{+\infty} (d\|\delta A\|_{\mathrm{op}})^k,$$

and finally $\|\delta A^-\|_{\mathrm{op}} \leq \dfrac{d^2\|\delta A\|_{\mathrm{op}}}{1 - d\|\delta A\|_{\mathrm{op}}}$. Consider $f := \left\{\begin{array}{ccc} [0, \frac{1}{d}) & \longrightarrow & [0, +\infty) \\ u & \longmapsto & \frac{d^2 u}{1-du} \end{array}\right.$.

The function f is bijective and increasing, with $f^{-1} = \left\{\begin{array}{ccc} [0, +\infty) & \longrightarrow & [0, \frac{1}{d}) \\ v & \longmapsto & \frac{v}{d(d+v)} \end{array}\right.$. This analysis yields under $(H_{\delta A})$, $\|\delta A^-\|_{\mathrm{op}} \leq t \Longleftarrow \|\delta A\|_{\mathrm{op}} \leq \dfrac{t}{d(d+t)}$.

Conveniently, by Lemma A.II.3, $\|\delta A\|_{\mathrm{op}} \leq s$ with probability $1 - \eta/3$ if $p \geq \frac{8\log(3d/\eta)}{s^2}$. We can apply this to

$$\frac{t}{d(d+t)} = \frac{\varepsilon}{8dn\bar{c}_Z(1 + \frac{\varepsilon}{8n\bar{c}_Z})},$$

but in order to simplify the expression, we apply it to

$$s := \frac{3\varepsilon}{28dn\bar{c}_Z} \leq \frac{t}{d(d+t)},$$

where the inequality holds thanks to $\varepsilon \leq \frac{4}{3} n \bar{c}_Z$.

Now we must quantify the assumption $(H_{\delta A}) : \|\delta A\|_{\mathrm{op}} < 1/d$. Notice that $s \leq 1/d$ and thus the event $\|\delta A\|_{\mathrm{op}} < s$ is contained in the event $\|\delta A\|_{\mathrm{op}} < 1/d$, hence it is sufficient to satisfy (H_1), which we write (after upper-bounding $8 \times 28^2/9 \leq 697$):

$$(H_1): \quad p \geq \frac{697 d^2 n^2 \bar{c}_Z^2 \log(3d/\eta)}{\varepsilon^2}.$$

To summarise, under (H_1), we have $\|\delta A\|_{\mathrm{op}} \leq s$ with probability exceeding $1-\eta/3$. Conditionally to the events "$\|\delta A\|_{\mathrm{op}} \leq s$", "$\|\delta v_i\|_2 \leq \frac{\varepsilon}{4}$", $i \in \{2,3,4\}$, this step shows $\|\delta v_1\|_2 \leq \frac{\varepsilon}{4}$.

— *Step 7*: Wrapping up.

We now work under the conditions (H_i), $i \in \{1,2,3,4\}$. By Step 1,

$$\|y_k^* - v_k\|_2 \leq \|\delta v_1^k\|_2 + \|\delta v_2^k\|_2 + \|\delta v_3^k\|_2 + \|\delta v_4^k\|_2,$$

where we restore the omitted k indices. By Step 3, with probability exceeding $1 - \eta/(3n)$, we have $\|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4}$, thus with probability $1-\eta/3$ we have $\forall k \in [\![1,n]\!]$, $\|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4}$. By Step 4, we have almost-surely $\forall k \in [\![1,n]\!]$, $\|\delta v_3^k\|_2 \leq \frac{\varepsilon}{4}$. By Step 5, with probability $1-\eta/3$, $\|\delta A\|_{\mathrm{op}} \leq s$. Putting this together yields that with probability $1-\eta$, we have:

$$\forall k \in [\![1,n]\!],\ \|\delta v_2^k\|_2 \leq \frac{\varepsilon}{4},\ \|\delta v_3^k\|_2 \leq \frac{\varepsilon}{4},\ \|\delta v_4^k\|_2 \leq \frac{\varepsilon}{4} \text{ and } \|\delta A\|_{\mathrm{op}} \leq s.$$

Finally, Step 5 shows that conditionally to the events above, $\|\delta v_1^k\|_2 \leq \frac{\varepsilon}{4}$ almost-surely. Thus with probability exceeding $1-\eta$, $\forall k \in [\![1,n]\!], \|y_k^* - v_k\|_2 \leq \varepsilon$. Since $v_k = \sum_{l=1}^n S_{k,l} z_l$, with probability over $1-\eta$: $\forall k \in [\![1,n]\!]$, $\left\| y_k^* - \sum_{l=1}^n S_{k,l} z_l \right\|_2 \leq \varepsilon$. □

In order to get the summarised result from Section A.II.3.2.3, we simplify the conditions as follows.

Corollary A.II.2 (Simplified conditions for Theorem A.II.12)**.** With the notations of Theorem A.II.12, the condition:

$$(H_p): \quad p \geq \left(\frac{4096 d^3 n \bar{c}_Z^3 \log(3dn^2/\eta)}{\varepsilon^3}\right) \vee \left(\frac{697 d^2 n^2 \bar{c}_Z^2 \log(3d/\eta)}{\varepsilon^2}\right) \vee \left(\frac{8 d^2 n^2 \bar{c}_Z^2 \log(6n^2/\eta)}{\varepsilon^2}\right) \tag{A.II.52}$$

implies (H_q) and $(H_i)_{i\in\{1,2,3,4\}}$, and thus is sufficient in order to have Eq. (A.II.25).

Proof. The second and third terms of Eq. (A.II.52) correspond to (H_1) and (H_4) respectively. Then, using $n^+, n^- \leq n$, we have

$$(H_2) \Longleftarrow \overline{q} \geq \frac{512 d^2 \bar{c}_Z^2 \log(3dn^2\eta)}{\varepsilon^2},$$

$$(H_3) \Longleftarrow \underline{q} \leq \frac{\varepsilon}{8dn\bar{c}_Z}\, p.$$

Let $q := \frac{512 d^2 \bar{c}_Z^2 \log(3dn^2/\eta)}{\varepsilon^2}$; $\overline{q} = \underline{q} = q$. (H_q) and (H_2) are automatically satisfied by this choice. For q to satisfy (H_3), it is sufficient to have

$$\frac{512 d^2 \bar{c}_Z^2 \log(3dn^2/\eta)}{\varepsilon^2} \leq \frac{\varepsilon}{8dn\bar{c}_Z}\, p, \text{ i.e. } p \geq \frac{4096 d^3 n \bar{c}_Z^3 \log(3dn^2/\eta)}{\varepsilon^3}$$

□

A.II.7.5 Closed-form expression for Block-Coordinate Descent

In Algorithm A.II.1, we mention in line 4 the minimisation $\min_Y J(\pi, Y)$, where

$$J := \left\{ \begin{array}{ccc} \mathbb{U}^p \times \mathbb{R}^{n\times d} & \longrightarrow & \mathbb{R}_+ \\ (\pi^{(1)}, \cdots, \pi^{(p)}), Y & \longmapsto & \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\sum_{l=1}^{n}(\theta_i^\top y_k - \theta_i^\top z_l)^2 \pi_{k,l}^{(i)} \end{array} \right. ,$$

and claim that it can in fact be done explicitly. We provide the formula below, which stems from a straightforward quadratic minimisation: let $Y^* = ((y_1^*)^\top, \cdots, (y_n^*)^\top)^\top = \operatorname{argmin}_Y J(\pi, Y)$, we obtain

$$\forall k \in [\![1, n]\!],\ y_k^* = \left(\frac{1}{n}\sum_{i=1}^{p}\theta_i\theta_i^\top\right)^{-1}\left(\sum_{i=1}^{p}\sum_{l=1}^{n}\pi_{k,l}^{(i)}\theta_i\theta_i^\top z_l\right),$$

where we used the constraint $\pi \in \mathbb{U}^p$ which implies $\sum_l \pi_{k,l}^{(i)} = \frac{1}{n}$.

A.III

Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses

A.III.1 Introduction . . . 102
A.III.1.1 Optimal Transport in Machine Learning 102
A.III.1.2 The Sliced Wasserstein Distance as an Alternative 102
A.III.1.3 Related Works 103
A.III.1.4 Contributions 104
A.III.2 Stochastic Gradient Descent with SW as Loss . . . 104
A.III.3 Convergence of Interpolated SGD Trajectories on F . . . 108
A.III.4 Convergence of Noised Projected SGD Schemes on F . . . 110
A.III.5 Convergence for Decreasing Steps in the Semi-Algebraic Setting . . . 112
A.III.6 Conclusion and Outlook . . . 113
A.III.7 Appendix . . . 115
A.III.7.1 Table of Notations 115
A.III.7.2 Postponed Proofs 115
A.III.7.3 Background on Non-Smooth and Non-Convex Analysis 117
A.III.7.4 Suitable Neural Networks 118
A.III.7.5 Generalisation to Other Sliced Wasserstein Orders 121

Abstract

Optimal Transport has sparked vivid interest in recent years, in particular thanks to the Wasserstein distance, which provides a geometrically sensible and intuitive way of comparing probability measures. For computational reasons, the Sliced Wasserstein (SW) distance was introduced as an alternative to the Wasserstein distance, and has seen uses for training generative Neural Networks (NNs). While convergence of Stochastic Gradient Descent (SGD) has been observed practically in such a setting, there is to our knowledge no theoretical guarantee for this observation. Leveraging recent works on convergence of SGD on non-smooth and non-convex functions by [BHS22], we aim to bridge that knowledge gap, and provide a realistic context under which fixed-step SGD trajectories for the SW loss on NN parameters converge. More precisely, we show that the trajectories approach the set of (sub)-gradient flow equations as the step decreases. Under stricter assumptions, we show a much stronger convergence result for noised and projected SGD schemes, namely that the long-run limits of the trajectories approach a set of generalised critical points of the loss function. Finally, we provide another convergence result for the decreasing-step SGD case under mild semi-algebraicity conditions. This chapter is based on the paper:

[Tan23] Eloi Tanguy.

"Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses".
Transactions on Machine Learning Research (Oct. 2023).

A.III.1 Introduction

A.III.1.1 Optimal Transport in Machine Learning

Optimal Transport (OT) allows the comparison of measures on a metric space by generalising the use of the ground metric. Typical applications use the so-called 2-Wasserstein distance, defined as

$$\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{W}_2^2(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \mathrm{d}\pi(x,y), \tag{W2}$$

where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ admitting a second-order moment and where $\Pi(\mu,\nu)$ is the set of measures of $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ of first marginal μ and second marginal ν. One may find a thorough presentation of its properties in classical monographs such as [PC19b; San15; Vil09]

The ability to compare probability measures is useful in probability density fitting problems, which are a sub-genre of generation tasks. In this formalism, one considers a probability measure parametrised by a vector u which is designed to approach a target data distribution ν (typically the real-world dataset). In order to determine suitable parameters, one may choose any probability discrepancy (Kullback-Leibler, Csiszár divergences, f-divergences or Maximum Mean Discrepancy [Gre+06]), or in our case, the Wasserstein distance. In the case of Generative Adversarial Networks, the optimisation problem which trains the "Wasserstein GAN" [ACB17] stems from the Kantorovich-Rubinstein dual expression of the 1-Wasserstein distance.

A.III.1.2 The Sliced Wasserstein Distance as an Alternative

The Wasserstein distance suffers from the curse of dimensionality, in the sense that the sample complexity for n samples in dimension d is of the order $\mathcal{O}(n^{1/d})$ [Dud69]. Due to this practical limitation and to the computational cost of the Wasserstein distance, the study of cheaper alternatives has become a prominent field of research. A prominent example is Entropic OT introduced by [Cut13], which adds an entropic regularisation term, advantageously making the problem strongly convex. Sample complexity bounds have been derived by [Gen+19], showing a convergence in $\mathcal{O}(\sqrt{n})$ with a constant depending on the regularisation factor.

Another alternative is the Sliced Wasserstein (SW) Distance introduced by [Rab+12], which consists in computing the 1D Wasserstein distances between projections of input measures, and averaging over the projections. The aforementioned projection of a measure μ on $\mathbb{R}^d$ is done by the *push-forward* operation by the map $P_\theta : x \longmapsto \theta^\top x$. Formally, $P_\theta \# \mu$ is the measure on $\mathbb{R}$ such that for any Borel set $B \subset \mathbb{R}$, $P_\theta \# \mu(B) = \mu(P_\theta^{-1}(B))$. Once the measures are projected onto a line $\mathbb{R}\theta$, the computation of the Wasserstein distance becomes substantially simpler numerically. We illustrate this fact in the discrete case, which arises in practical optimisation settings. Let two discrete measures on $\mathbb{R}^d$: $\gamma_X := \frac{1}{n}\sum_k \delta_{x_k}$, $\gamma_Y := \frac{1}{n}\sum_k \delta_{y_k}$ with supports $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n) \in \mathbb{R}^{n \times d}$. Their push-forwards by P_θ are simply computed by the formula $P_\theta \# \gamma_X = \frac{1}{n}\sum_k \delta_{P_\theta(x_k)}$, and the 2-Wasserstein distance between their projections can be computed by sorting their supports: let σ a permutation sorting $(\theta^\top x_1, \cdots, \theta^\top x_n)$, and τ a permutation sorting $(\theta^\top y_1, \cdots, \theta^\top y_n)$, one has the simple expression

$$\mathrm{W}_2^2(P_\theta \# \gamma_X, P_\theta \# \gamma_Y) = \frac{1}{n}\sum_{k=1}^{n} (\theta^\top x_{\sigma(k)} - \theta^\top y_{\tau(k)})^2. \tag{A.III.1}$$

The SW distance is the expectation of this quantity with respect to $\theta \sim \varpi$, i.e. uniform on the sphere: $\mathrm{SW}_2^2(\gamma_X, \gamma_Y) = \mathbb{E}_{\theta \sim \varpi}\left[\mathrm{W}_2^2(P_\theta \# \gamma_X, P_\theta \# \gamma_Y)\right]$. The 2-SW distance is also defined more generally between two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\mathrm{SW}_2^2(\mu,\nu) := \int_{\theta \in \mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \mu, P_\theta \# \nu) \mathrm{d}\varpi(\theta). \tag{SW}$$

In addition to its computational accessibility, the SW distance enjoys a dimension-free sample complexity [Nad+20b]. Additional statistical, computational and robustness properties of SW have been explored by [Nie+22]. Moreover, central-limit results have been shown by [XH22] for 1-SW and the 1-max-SW distance (a variant of SW introduced by [Des+19]), and related work by [XN22] shows the convergence of the sliced error process

$$\theta \longmapsto \sqrt{n}\left(\mathrm{W}_p^p(P_\theta \# \gamma_X, P_\theta \# \gamma_Y) - \mathrm{W}_p^p(P_\theta \# \mu, P_\theta \# \nu)\right),$$

where the samples $X \sim \mu^{\otimes n}$ and $Y \sim \nu^{\otimes n}$ are drawn for each θ. Another salient field of research for SW is its metric properties, and while it has been shown to be weaker than the Wasserstein distance in general by [Bon13], and metric comparisons with Wasserstein and max-SW have been undergone by [BG21] and [PC19a].

A.III.1.3 Related Works

Our subject of interest is the theoretical properties of SW as a loss for implicit generative modelling, which leads to minimising $\mathrm{SW}_2^2(T_u \# \mu, \nu)$ in the parameters u, where ν is the target distribution, and $T_u \# \mu$ is the image by the NN[1] of μ, a low-dimensional input distribution (often chosen as Gaussian or uniform noise). In order to train a NN in this manner, at each iteration one draws n samples from μ and ν (denoted γ_X and γ_Y as discrete measures with n points), as well as a projection θ (or a batch of projections) and performs an SGD step on the sample loss

$$\mathcal{L}(u) = \mathrm{SW}_2^2(P_\theta \# T_u \# \gamma_X, P_\theta \# \gamma_Y) = \frac{1}{n}\sum_{k=1}^{n} (\theta^\top T_u(x_{\sigma(k)}) - \theta^\top y_{\tau(k)})^2. \tag{A.III.2}$$

Taking the expectation of this loss over the samples yields the minibatch Sliced-Wasserstein discrepancy, a member of the minibatch variants of the OT distances, introduced formally by Fatras et al. [Fat+21b]. The framework Eq. (A.III.2) fits several Machine Learning applications, for instance, [DZS18] trains GANs and auto-encoders with this method, and [Wu+19] consider related dual formulations. Other examples within this formalism include the synthesis of images by minimising the SW distance between features of the optimised image and a target image, as done by [Hei+21] for textures with neural features, and by [TPG16] with wavelet features (amongst other methods).

The general study of convergence of SGD in the context of non-smooth, non-convex functions (as is the case of $\mathcal{L}$ from Eq. (A.III.2)) is an active field of research: [MMM18] and [Dav+20] show the convergence of diminishing-step SGD under regularity constraints, while [BP21] leverage conservative field theory to show convergence results for training with back-propagation. Finally, the recent work by [BHS22] shows the convergence of fixed-step SGD schemes on a general function F under weaker regularity assumptions.

More specifically, the study of convergence for OT-based generative NNs has been tackled by [Fat+21b], who prove strong convergence results for minibatch variants of classical OT distances, namely the Wasserstein distance, the Entropic OT and the Gromov Wasserstein distance (another OT variant introduced by [Mém11]). A related study on GANs by [Hua+23] derive optimisation properties for one layer and one dimensional Wasserstein-GANs and generalise to higher dimensions by turning to SW-GANs. Another work by [Bré+23] focuses on the theoretical properties of linear NNs trained with the Bures-Wasserstein loss (introduced by [Bur69]; see also [BJL19] for reference on this metric). Finally, the regularity and optimisation properties of the simpler energy $\mathrm{SW}_2^2(\gamma_X, \gamma_Y)$ have been studied in Chapter A.II.

In practice, it has been observed that SGD in such settings always converges (in the loose numerical sense, see [DZS18, Section 5], or [Hei+21, Figure 3], yet this property is not known theoretically. The aim of this work is to bridge the gap between theory and practical observation by proving convergence results for SGD on (minibatch) Sliced Wasserstein generative losses of the form $F(u) = \mathbb{E}_{X\sim\mu^{\otimes n}, Y\sim\nu^{\otimes n}} \mathrm{SW}_2^2(T_u \# \gamma_X, \gamma_Y)$.

[1] Similarly to the 1D case, $T_u \# \mu$ is the push-forward measure of μ by T_u, i.e. the law of $T_u(x)$ when $x \sim \mu$.

A.III.1.4 Contributions

Convergence of Interpolated SGD Under Practical Assumptions. Under practically realistic assumptions, we prove in Theorem A.III.1 that piecewise affine interpolations (defined in Equation Eq. (A.III.10)) of constant-step SGD schemes on $u \longmapsto F(u)$ (formalised in Equation Eq. (A.III.7)) converge towards the set of sub-gradient flow solutions (see Equation Eq. (A.III.9)) as the gradient step decreases. This results signifies that with very small learning rates, SGD trajectories will be close to sub-gradient flows, which themselves converge to critical points of F (omitting serious technicalities).

The assumptions for this result are practically reasonable: the input measure μ and the true data measure ν are assumed to be compactly supported. As for the network $(u, x) \longmapsto T(u, x)$, we assume that for a fixed datum x, $T(\cdot, x)$ is piecewise $\mathcal{C}^2$-smooth and that it is Lipschitz jointly in both variables.

We require additional assumptions on T which are more costly, but are verified as long as T is a NN composed of typical activations and linear units, with the constraint that the parameters u and data x stay both stay within a fixed bounded domains. We discuss a class of neural networks that satisfy all of the assumptions of the chapter in the Appendix (Section A.III.7.4). Furthermore, this result can be extended to other orders $p \neq 2$ of SW: we present the tools for this generalisation in Section A.III.7.5.

Stronger Convergence Under Stricter Assumptions. In order to obtain a stronger convergence result, we consider a variant of SGD where each iteration receives an additive noise (scaled by the learning rate) which allows for better space exploration, and where each iteration is projected on a ball $B(0, r)$ in order to ensure boundedness. This alternative SGD scheme remains within the realm of practical applications, and we show in Theorem A.III.2 that long-run limits of such trajectories converge towards a set of generalised critical points of F, as the gradient step approaches 0. This result is substantially stronger, and can serve as an explanation of the convergence of practical SGD trajectories, specifically towards a set of critical points which amounts to the stationary points of the energy (barring theoretical technicalities).

Unfortunately, we require additional assumptions in order to obtain this stronger convergence result, the most important of which is that the input data measure μ and the dataset measure ν are discrete. For the latter, this is always the case in practice, however the former assumption is more problematic, since it is common to envision generative NNs as taking an argument from a continuous space (the input is often Gaussian of Uniform noise), thus a discrete setting is a substantial theoretical drawback. For practical concerns, one may argue that the discrete μ can have an arbitrary fixed amount of points, and leverage strong sample complexity results to ascertain that the discretisation is not costly if the number of samples is large enough.

Convergence of Decreasing-Step SGD Under Semi-Algebraic Assumptions. Finally, we show in Section A.III.5 that under mild (semi-algebraic) conditions, accumulation points of the decreasing-step SGD scheme converges a set of critical points. This result works under the mild assumption that the NN's activations are semi-algebraic, which is the case for the common ReLU activation. We also require that the source measure μ and target measure ν be either discrete of admit a semi-algebraic density, which is the case in essentially all practical applications. The result does not enforce the NN parameters to be bounded, but works under the typical condition that the trajectories of parameters remain bounded, which is observed in practice, in particular thanks to regularisation techniques.

A.III.2 Stochastic Gradient Descent with SW as Loss

Training Sliced-Wasserstein generative models consists in training a neural network

$$T : \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ (u, x) & \longmapsto & T_u(x) := T(u, x) \end{array} \right. \tag{A.III.3}$$

by minimising the SW minibatch loss $u \longmapsto \mathbb{E}_{X\sim\mu^{\otimes n},Y\sim\nu^{\otimes n}}\left[\mathrm{SW}_2^2(T_u\#\gamma_X,\gamma_Y)\right]$ through Stochastic Gradient Descent (as described in Algorithm A.III.1). The probability distribution $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ is the law of the input of the generator $T(u,\cdot)$. The distribution $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$ is the data distribution, which T aims to simulate. Finally, $\mathfrak{o}$ will denote the uniform measure on the unit sphere of $\mathbb{R}^{d_y}$, denoted by $\mathbb{S}^{d_y-1}$. Given a list of points $X = (x_1,\cdots,x_n) \in \mathbb{R}^{n\times d_x}$, denote the associated discrete uniform measure $\gamma_X := \frac{1}{n}\sum_i \delta_{x_i}$. By abuse of notation, we write $T_u(X) := (T_u(x_1),\cdots,T_u(x_n)) \in \mathbb{R}^{n\times d_y}$. The reader may find a summary of this chapter's notations in Table A.III.1.

Algorithm A.III.1: Training a NN on the SW loss with Stochastic Gradient Descent

Data: Learning rate $\alpha > 0$, probability distributions $\mu \in \mathcal{P}_2(\mathbb{R}^{d_x})$ and $\nu \in \mathcal{P}_2(\mathbb{R}^{d_y})$.
Initialisation: Draw $u^{(0)} \in \mathbb{R}^{d_u}$;
for $t \in [\![0, T_{\max}-1]\!]$ **do**
  Draw $\theta^{(t+1)} \sim \mathfrak{o}$, $X^{(t+1)} \sim \mu^{\otimes n}$ $Y^{(t+1)} \sim \nu^{\otimes n}$. SGD update:
  $$u^{(t+1)} = u^{(t)} - \alpha\left[\frac{\partial}{\partial u}\mathrm{W}_2^2(P_{\theta^{(t+1)}}\#T_u\#\gamma_{X^{(t+1)}}, P_{\theta^{(t+1)}}\#\gamma_{Y^{(t+1)}})\right]_{u=u^{(t)}}$$
end

In the following, we will apply results from [BHS22], and we pave the way to the application of these results by presenting their theoretical framework. Consider a sample loss function $f : \mathbb{R}^{d_u}\times\mathcal{Z} \longrightarrow \mathbb{R}$ that is locally Lipschitz in the first variable, and ζ a probability measure on $\mathcal{Z} \subset \mathbb{R}^d$ which is the law of the samples drawn at each SGD iteration. Consider $\varphi : \mathbb{R}^{d_u}\times\mathcal{Z} \longrightarrow \mathbb{R}^{d_u}$ an *almost-everywhere gradient* of f, which is to say that for almost every $(u,z) \in \mathbb{R}^{d_u}\times\mathcal{Z}$, $\varphi(u,z) = \partial_u f(u,z)$ (since each $f(\cdot,z)$ is locally Lipschitz, it is differentiable almost-everywhere by Rademacher's theorem). The complete loss function is the expectation of the sample loss, $F := u \longmapsto \int_{\mathcal{Z}} f(u,z)\mathrm{d}\zeta(z)$. An SGD trajectory of step $\alpha > 0$ for F is a sequence $(u^{(t)}) \in (\mathbb{R}^{d_u})^{\mathbb{N}}$ of the form:

$$u^{(t+1)} = u^{(t)} - \alpha\varphi(u^{(t)}, z^{(t+1)}), \quad \left(u^{(0)}, (z^{(t)})_{t\in\mathbb{N}}\right) \sim \rho_0 \otimes \zeta^{\otimes\mathbb{N}},$$

where ρ_0 is the distribution of the initial position $u^{(0)}$. Within this framework, we define an SGD scheme described by Algorithm A.III.1, with $\zeta := \mu^{\otimes n}\otimes\nu^{\otimes n}\otimes\mathfrak{o}$ and the minibatch SW sample loss

$$f := \left\{\begin{array}{ccc} \mathbb{R}^{d_u}\times\mathbb{R}^{n\times d_x}\times\mathbb{R}^{n\times d_y}\times\mathbb{S}^{d_y-1} & \longrightarrow & \mathbb{R}^{d_y} \\ (u,X,Y,\theta) & \longmapsto & \mathrm{W}_2^2(P_\theta\#T_u\#\gamma_X, P_\theta\#\gamma_Y) \end{array}\right. . \tag{A.III.4}$$

With this definition for f, we have

$$F(u) = \mathbb{E}_{(X,Y,\theta)\sim\zeta}\left[f(u,X,Y,\theta)\right] = \mathbb{E}_{(X,Y)\sim\mu^{\otimes n}\otimes\nu^{\otimes n}}\left[\mathrm{SW}_2^2(T_u\#\gamma_X,\gamma_Y)\right], \tag{A.III.5}$$

thus the population loss compares the "true" data ν with the model's generation $T_u\#\mu$ using (minibatch) SW. We now wish to define an almost-everywhere gradient of f. To this end, notice that one may write $f(u,X,Y,\theta) = w_\theta(T(u,X),Y)$, where for $X,Y \in \mathbb{R}^{n\times d_y}$ and $\theta \in \mathbb{S}^{d_y-1}$, $w_\theta(X,Y) := \mathrm{W}_2^2(P_\theta\#\gamma_X, P_\theta\#\gamma_Y)$. The differentiability properties of $w_\theta(\cdot,Y)$ are already known (they were studied in Chapter A.II: see in particular Proposition A.II.1), in particular one has the following almost-everywhere gradient of $w_\theta(\cdot,Y)$:

$$\frac{\partial w_\theta}{\partial X}(X,Y) = \left(\frac{2}{n}\theta\theta^\top(x_k - y_{\sigma_\theta^{X,Y}(k)})\right)_{k\in[\![1,n]\!]} \in \mathbb{R}^{n\times d_y},$$

where the permutation $\sigma_\theta^{X,Y} \in \mathfrak{S}_n$ is $\tau_Y^\theta \circ (\tau_X^\theta)^{-1}$, with $\tau_Y^\theta \in \mathfrak{S}_n$ being a sorting permutation of the list $(\theta^\top y_1,\cdots,\theta^\top y_n)$. The sorting permutations are chosen arbitrarily when there is ambiguity. To define an almost-everywhere gradient, we must differentiate $f(\cdot,X,Y,\theta) = u \longmapsto$

$w_\theta(T(u,X),Y)$ for which we need regularity assumptions on T: this is the goal of Assumption A.III.1. In the following, $\overline{A}$ denotes the topological closure of a set A, ∂A its boundary, and $\mathscr{L}^{d_u}$ denotes the Lebesgue measure of $\mathbb{R}^{d_u}$.

Assumption A.III.1. For every $x \in \mathbb{R}^{d_x}$, there exists a family of disjoint connected open sets $(\mathcal{U}_j(x))_{j\in J(x)}$ such that:

$$\forall j \in J(x),\ T(\cdot,x) \in \mathcal{C}^2(\mathcal{U}_j(x),\mathbb{R}^{d_y}),\quad \bigcup_{j\in J(x)} \overline{\mathcal{U}_j(x)} = \mathbb{R}^{d_u} \text{ and } \mathscr{L}^{d_u}\Big(\bigcup_{j\in J(x)} \partial\mathcal{U}_j(x)\Big) = 0.$$

Note that for measure-theoretic reasons, the sets $J(x)$ are assumed countable. One may understand this assumption broadly as the neural networks T being piecewise smooth with respect to the parameters u, where the pieces depend on the input data x. In practice, Assumption A.III.1 is an assumption on the activation functions of the neural network. For instance, it is of course satisfied in the case of smooth activations, or in the common case of piecewise polynomial activations. We detail suitable neural networks in the Appendix (Section A.III.7.4).

Assumption A.III.1 implies that given X, Y, θ fixed, $f(\cdot, X, Y, \theta)$ is differentiable almost-everywhere, and that one may define the following almost-everywhere gradient Eq. (A.III.6).

$$\varphi : \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{n\times d_x} \times \mathbb{R}^{n\times d_y} \times \mathbb{S}^{d_y-1} & \longrightarrow & \mathbb{R}^{d_u} \\ (u,X,Y,\theta) & \longmapsto & \displaystyle\sum_{k=1}^{n} \frac{2}{n} \left(\frac{\partial T}{\partial u}(u,x_k)\right)^\top \theta\theta^\top (T(u,x_k) - y_{\sigma_\theta^{T(u,X),Y}(k)}) \end{array} \right. , \tag{A.III.6}$$

where for $x \in \mathbb{R}^{d_x}$, $\dfrac{\partial T}{\partial u}(u,x) \in \mathbb{R}^{d_y\times d_u}$ denotes the matrix of the differential of $u \longmapsto T(u,x)$, which is defined for almost-every u. Given $u \in \partial\mathcal{U}_j(x)$ (a point of potential non-differentiability), take instead 0. (Any choice at such points would still define an a.e. gradient, and will make no difference). Given a step $\alpha > 0$, and an initial position $u^{(0)} \sim \rho_0$, we may now define formally the following fixed-step SGD scheme for F:

$$\begin{aligned} u^{(t+1)} &= u^{(t)} - \alpha\varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}), \\ \left(u^{(0)}, (X^{(t)})_{t\in\mathbb{N}}\ (Y^{(t)})_{t\in\mathbb{N}}\ (\theta^{(t)})_{t\in\mathbb{N}}\right) &\sim \rho_0 \otimes \mu^{\otimes\mathbb{N}} \otimes \nu^{\otimes\mathbb{N}} \otimes \mathfrak{o}^{\otimes\mathbb{N}}. \end{aligned} \tag{A.III.7}$$

An important technicality that we must verify in order to apply [BHS22]'s results is that $u \longmapsto f(u,X,Y,\theta)$ and F are locally Lipschitz. Before proving those claims, we recall a useful property from Chapter A.II. In the following, $\|X\|_{\infty,2}$ denotes $\max_{k\in[\![1,n]\!]} \|x_k\|_2$ given $X = (x_1,\cdots,x_n) \in \mathbb{R}^{n\times d_x}$, and $B_{\mathcal{N}}(x,r)$ for $\mathcal{N}$ a norm on $\mathbb{R}^{d_x}$, $x \in \mathbb{R}^{d_x}$ and $r > 0$ shall denote the open ball of $\mathbb{R}^{d_x}$ of centre x and radius r for the norm $\mathcal{N}$ (if $\mathcal{N}$ is omitted, then B is an euclidean ball).

Proposition A.III.1. The $(w_\theta(\cdot,Y))_{\theta\in\mathbb{S}^{d_y-1}}$ are uniformly locally Lipschitz
Let $K_w(r,X,Y) := 2n(r + \|X\|_{\infty,2} + \|Y\|_{\infty,2})$, for $X, Y \in \mathbb{R}^{n\times d_y}$ and $r > 0$. Then $w_\theta(\cdot,Y)$ is $K_w(r,X,Y)$-Lipschitz in the neighbourhood $B_{\|\cdot\|_{\infty,2}}(X,r)$:

$$\forall Y', Y'' \in B_{\|\cdot\|_{\infty,2}}(X,r),\ \forall\theta \in \mathbb{S}^{d_y-1},\ |w_\theta(Y',Y) - w_\theta(Y'',Y)| \le K_w(r,X,Y)\|Y' - Y''\|_{\infty,2}.$$

In order to deduce regularity results on f and F from Proposition A.III.1, we will make the assumption that T is globally Lipschitz in (u,x). In practice, this is the case when both parameters are enforced to stay within a fixed bounded domain, for instance by multiplying a typical NN with the indicator of such a set. We present this in detail in the Appendix (Section A.III.7.4).

Assumption A.III.2. There exists $L > 0$ such that:

$$\forall (u_1, u_2, x_1, x_2) \in (\mathbb{R}^{d_u})^2 \times (\mathbb{R}^{d_x})^2,\ \|T(u_1, x_1) - T(u_2, x_2)\|_2 \leq L\left(\|u_1 - u_2\|_2 + \|x_1 - x_2\|_2\right).$$

Proposition A.III.2. Under Assumption A.III.2, for $\varepsilon > 0,\ u_0 \in \mathbb{R}^{d_u},\ X \in \mathbb{R}^{n \times d_x},\ Y \in \mathbb{R}^{n \times d_y}$ and $\theta \in \mathbb{S}^{d_y - 1}$, let $K_f(\varepsilon, u_0, X, Y) := 2Ln(\varepsilon L + \|T(u_0, X)\|_{\infty,2} + \|Y\|_{\infty,2})$. Then $f(\cdot, X, Y, \theta)$ is $K_f(\varepsilon, u_0, X, Y)$-Lipschitz in $B(u_0, \varepsilon)$:

$$\forall u, u' \in B(u_0, \varepsilon),\ |f(u, X, Y, \theta) - f(u', X, Y, \theta)| \leq K_f(\varepsilon, u_0, X, Y)\|u - u'\|_2.$$

Proof. Let $\varepsilon > 0,\ u_0 \in \mathbb{R}^{d_u},\ X \in \mathbb{R}^{n \times d_x},\ Y \in \mathbb{R}^{n \times d_y}$ and $\theta \in \mathbb{S}^{d_y - 1}$. Let $u, u' \in B(u_0, \varepsilon)$. Using Assumption A.III.2, we have $T(u, X), T(u', X) \in B_{\|\cdot\|_{\infty,2}}(T(u_0, X), r)$, with $r := \varepsilon L$.

Denoting $L := L_{\overline{B}(u_0, \varepsilon), \overline{B}(0_{\mathbb{R}^{d_x}}, \|X\|_{\infty,2})}$, we apply successively Proposition A.III.1 (first inequality), then Assumption A.III.2 (second inequality):

$$\begin{aligned}
|f(u, X, Y, \theta) - f(u', X, Y, \theta)| &= |w_\theta(T(u, X), Y) - w_\theta(T(u', X), Y)| \\
&\leq K_w(r, T(u_0, X), Y)\|T(u, X) - T(u', X)\|_{\infty,2} \\
&\leq 2n(\varepsilon L + \|T(u_0, X)\|_{\infty,2} + \|Y\|_{\infty,2})L\|u - u'\|_2.
\end{aligned}$$

□

Proposition A.III.2 shows that f is locally Lipschitz in u. We now assume some conditions on the measures μ and ν in order to prove that F is also locally Lipschitz. Specifically, we require that the data measures μ and ν be supported on bounded domains, which imposes little restriction in practice.

Assumption A.III.3. μ and ν are Radon probability measures on $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$ respectively, supported by the compact sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Denote $R_x := \sup_{x \in \mathcal{X}} \|x\|_2$ and $R_y := \sup_{y \in \mathcal{Y}} \|y\|_2$.

Proposition A.III.3. Assume Assumptions A.III.2 and A.III.3. For $\varepsilon > 0,\ u_0 \in \mathbb{R}^{d_u}$, let $C_1(u_0) := \int_{\mathcal{X}^n} \|T(u_0, X)\|_{\infty,2} \mathrm{d}\mu^{\otimes n}(X)$ and $C_2 := \int_{\mathcal{Y}^n} \|Y\|_{\infty,2} \mathrm{d}\nu^{\otimes n}(Y)$.
Let $K_F(\varepsilon, u_0) := 2Ln(\varepsilon L + C_1(u_0) + C_2)$. We have:

$$\forall u, u' \in B(u_0, \varepsilon),\ |F(u) - F(u')| \leq K_F(\varepsilon, u_0)\|u - u'\|_2.$$

Proof. Let $\varepsilon > 0,\ u_0 \in \mathbb{R}^{d_u}$ and $u, u' \in B(u_0, \varepsilon)$. We have

$$\begin{aligned}
|F(u) - F(u')| &\leq \int_{\mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y - 1}} |f(u, X, Y, \theta) - f(u', X, Y, \theta)| \mathrm{d}\mu^{\otimes n}(X) \mathrm{d}\nu^{\otimes n}(Y) \mathrm{d}\varpi(\theta) \\
&\leq \int_{\mathcal{X}^n \times \mathcal{Y}^n} K_f(\varepsilon, u_0, X, Y)\|u - u'\|_2 \mathrm{d}\mu^{\otimes n}(X) \mathrm{d}\nu^{\otimes n}(Y) \\
&\leq \int_{\mathcal{X}^n \times \mathcal{Y}^n} 2Ln(\varepsilon L + \|T(u_0, X)\|_{\infty,2} + \|Y\|_{\infty,2})\|u - u'\|_2 \mathrm{d}\mu^{\otimes n}(X) \mathrm{d}\nu^{\otimes n}(Y).
\end{aligned}$$

Now by Assumption A.III.2, $X \longmapsto \|T(u_0, X)\|_{\infty,2}$ is continuous on the compact set $\mathcal{X}^n$, thus upper-bounded by a certain $M(u_0) > 0$. We can define $C_1(u_0) := \int_{\mathcal{X}^n} \|T(u_0, X)\|_{\infty,2} \mathrm{d}\mu^{\otimes n}(X)$, which verifies $C_1(u_0) \leq M(u_0)\mu(\mathcal{X})^n$. Since $\mathcal{X}$ is compact and μ is a Radon probability measure by Assumption A.III.3, $\mu(\mathcal{X})$ is well-defined and finite, thus $C_1(u_0)$ is finite. Likewise, let $C_2 := \int_{\mathcal{Y}^n} \|Y\|_{\infty,2} \mathrm{d}\nu^{\otimes n}(Y) < +\infty$. Finally, $|F(u) - F(u')| \leq 2Ln(\varepsilon L + C_1(u_0) + C_2)\|u - u'\|_2$. □

Having shown that our losses are locally Lipschitz, we can now turn to convergence results. These conclusions are placed in the context of non-smooth and non-convex optimisation, thus

will be tied to the Clarke sub-differential of F, which we denote $\partial_C F$. The set of Clarke sub-gradients at a point u is the convex hull of the limits of gradients of F:

$$\partial_C F(u) := \operatorname{conv}\left\{v \in \mathbb{R}^{d_u} : \exists (u^{(t)}) \in (\mathcal{D}_F)^{\mathbb{N}} : u^{(t)} \xrightarrow[t\longrightarrow+\infty]{} u \text{ and } \nabla F(u^{(t)}) \xrightarrow[t\longrightarrow+\infty]{} v\right\}, \tag{A.III.8}$$

where $\mathcal{D}_F$ is the set of differentiability of F. At points u where F is differentiable, $\partial_C F(u) = \{\nabla F(u)\}$, and if F is convex in a neighbourhood of u, then the Clarke differential at u is the set of its convex sub-gradients. The interested reader may turn to Section A.III.7.3 for further context on non-smooth and non-convex optimisation, which is relevant to this chapter.

A.III.3 Convergence of Interpolated SGD Trajectories on F

In general, the idea behind SGD is a discretisation of the gradient flow equation $\dot{u}(s) = -\nabla F(u(s))$. In our non-smooth setting, the underlying continuous-time problem is instead the Clarke differential inclusion $\dot{u}(s) \in -\partial_C F(u(s))$. Our objective is to show that in a certain sense, the SGD trajectories approach the set of solutions of this inclusion problem, as the step size decreases. We consider solutions that are absolutely continuous (we will write $u(\cdot) \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u})$) and start within $\mathcal{K} \subset \mathbb{R}^{d_u}$, a fixed compact set. We can now define the solution set formally as

$$S_{-\partial_C F}(\mathcal{K}) := \left\{u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u}) \mid \underline{\forall} s \in \mathbb{R}_+,\ \dot{u}(s) \in -\partial_C F(u(s));\ u(0) \in \mathcal{K}\right\}, \tag{A.III.9}$$

where we write $\underline{\forall}$ for "almost every". In order to compare the discrete SGD trajectories to this set of continuous-time trajectories, we interpolate the discrete points in an affine manner: Equation Eq. (A.III.10) defines the *piecewise-affine interpolated SGD trajectory* associated to a discrete SGD trajectory $(u_\alpha^{(t)})_{t\in\mathbb{N}}$ of learning rate α.

$$u_\alpha(s) = u_\alpha^{(t)} + \left(\frac{s}{\alpha} - t\right)(u_\alpha^{(t+1)} - u_\alpha^{(t)}), \quad \forall s \in [t\alpha, (t+1)\alpha), \quad \forall t \in \mathbb{N}. \tag{A.III.10}$$

In order to compare our interpolated trajectories with the solutions, we consider the metric of uniform convergence on all segments

$$d_c(u, u') := \sum_{k\in\mathbb{N}^*} \frac{1}{2^k} \min\left(1, \max_{s\in[0,k]} \|u(s) - u'(s)\|_2\right). \tag{A.III.11}$$

In order to prove a convergence result on the interpolated trajectories, we will leverage the work of [BHS22] which hinges on three conditions on the loss F that we reproduce and verify successively. Firstly, Condition A.III.1 assumes mild regularity on the sample loss function f.

Condition A.III.1.

i) There exists $\kappa : \mathbb{R}^{d_u} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$ measurable such that each $\kappa(u, \cdot)$ is ζ-integrable, and:

$$\exists \varepsilon > 0,\ \forall u, u' \in B(u_0, \varepsilon),\ \forall z \in \mathcal{Z},\ |f(u, z) - f(u', z)| \leq \kappa(u_0, z)\|u - u'\|_2.$$

ii) There exists $u \in \mathbb{R}^{d_u}$ such that $f(u, \cdot)$ is ζ-integrable.

Our regularity result on f Proposition A.III.2 allows us to verify Condition A.III.1, by letting $\varepsilon := 1$ and $\kappa(u_0, z) := K_f(1, u_0, X, Y)$. Condition A.III.1 ii) is immediate since for *all* $u \in \mathbb{R}^{d_u}$, $(X, Y, \theta) \longmapsto w_\theta(T(u, X), Y)$ is continuous in each variable separately, thanks to the regularity of T provided by Assumption A.III.2, and to the regularities of w. This continuity implies that all $f(u, \cdot)$ are ζ-integrable, since $\zeta = \mu^{\otimes n} \otimes \nu^{\otimes n} \otimes \varpi$ is a compactly supported probability measure under Assumption A.III.3. Secondly, Condition A.III.2 concerns the local Lipschitz constant κ introduced in Condition A.III.1: it is assumed to increase slowly with respect to the network parameters u.

Condition A.III.2. The function κ of Condition A.III.1 verifies:

i) There exists $c \geq 0$ such that $\forall u \in \mathbb{R}^{d_u}$, $\int_{\mathcal{Z}} \kappa(u,z)\mathrm{d}\zeta(z) \leq c(1+\|u\|_2)$.

ii) For every compact $\mathcal{K} \subset \mathbb{R}^{d_u}$, $\sup_{u\in\mathcal{K}} \int_{\mathcal{Z}} \kappa(u,z)^2\mathrm{d}\zeta(z) < +\infty$.

Condition A.III.2.ii) is verified by κ given its regularity. However, Condition A.III.2.i) requires that $T(u,x)$ increase slowly as $\|u\|_2$ increases, which is more costly.

Assumption A.III.4. There exists an μ-integrable function $g : \mathbb{R}^{d_x} \longrightarrow \mathbb{R}_+$ such that $\forall u \in \mathbb{R}^{d_u}$, $\forall x \in \mathbb{R}^{d_x}$, $\|T(u,x)\|_2 \leq g(x)(1+\|u\|_2)$.

Assumption A.III.4 is satisfied in particular as soon as $T(\cdot,x)$ is bounded (which is the case for a neural network with bounded activation functions), or if T is of the form $T(u,x) = \widetilde{T}(u,x)\mathbb{1}_{B(0,R)}(u)$, i.e. limiting the network parameters u to be bounded. This second case does not yield substantial restrictions in practice (see Section A.III.7.4 for a class of NNs that satisfy all of the assumptions), yet vastly simplifies theory. Under Assumption A.III.4, we have for any $u \in \mathbb{R}^{d_u}$, with $\kappa(u,z) = K_f(1,u,X,Y)$ from Proposition A.III.2 and C_2 from Proposition A.III.3,

$$\begin{aligned}
&\int_{\mathcal{X}^n\times\mathcal{Y}^n\times\mathbb{S}^{d_y-1}} K_f(1,u,X,Y)\mathrm{d}\mu^{\otimes n}(X)\mathrm{d}\nu^{\otimes n}(Y)\mathrm{d}\varpi(\theta)\\
&\leq 4Ln\left(\varepsilon L + (1+\|u\|_2)\int_{\mathcal{X}^n} \max_{k\in[\![1,n]\!]} g(x_k)\mathrm{d}\mu^{\otimes n}(X) + C_2\right)\\
&\leq c(1+\|u\|_2).
\end{aligned}$$

As a consequence, Condition A.III.2 holds under our assumptions. We now consider the Markov kernel associated to the SGD schemes:

$$P_\alpha : \left\{\begin{array}{ccc} \mathbb{R}^{d_u}\times\mathcal{B}(\mathbb{R}^{d_u}) & \longrightarrow & [0,1] \\ u, B & \longmapsto & \int_{\mathcal{Z}} \mathbb{1}_B(u-\alpha\varphi(u,z))\mathrm{d}\zeta(z) \end{array}\right. .$$

Given $u \in \mathbb{R}^{d_u}$, $P_\alpha(u,\cdot)$ is a probability measure on $\mathbb{R}^{d_u}$ which dictates the law of the positions of the next SGD iteration $u^{(t+1)}$, conditionally to $u^{(t)} = u$. With $\mathscr{L}^{d_u}$ denoting the Lebesgue measure on $\mathbb{R}^{d_u}$, let $\Gamma := \left\{\alpha \in (0,+\infty) \mid \forall \rho \ll \mathscr{L}^{d_u},\ \rho P_\alpha \ll \mathscr{L}^{d_u}\right\}$. Γ is the set of learning rates α for which the kernel P_α maps any absolutely continuous probability measure ρ to another such measure. We will verify the following condition, which can be interpreted as the SGD trajectories continuing to explore the entire space for a small enough learning rate α:

Condition A.III.3. The closure of Γ contains 0.

In order to satisfy Condition A.III.3, we require an additional regularity condition on the neural network T which we formulate in Assumption A.III.5.

Assumption A.III.5. There exists a constant $M > 0$, such that (with the notations of Assumptions A.III.1 and A.III.3) $\forall x \in \mathcal{X}$, $\forall j \in J(x)$, $\forall u \in \mathcal{U}_j(x)$, $\forall (i_1,i_2,i_3,i_4) \in [\![1,d_u]\!]^2 \times [\![1,d_y]\!]^2$,

$$\left|\frac{\partial^2}{\partial u_{i_1}\partial u_{i_2}}\Big([T(u,x)]_{i_3}[T(u,x)]_{i_4}\Big)\right| \leq M, \text{ and } \left\|\frac{\partial^2 T}{\partial u_{i_1}\partial u_{i_2}}(u,x)\right\|_2 \leq M.$$

The upper bounds in Assumption A.III.5 bear strong consequences on the behaviour of T for $\|u\|_2 \gg 1$, and are only practical for networks of the form $T(u,x) = \widetilde{T}(u,x)\mathbb{1}_{B(0,R)}(u,x)$,

similarly to Assumption A.III.4. We detail the technicalities of verifying this assumption along with the others in the Appendix (Section A.III.7.4).

Proposition A.III.4. Under Assumptions A.III.1, A.III.3 and A.III.5, for the SGD trajectories Eq. (A.III.7), Γ contains $(0, \alpha_0)$, where $\alpha_0 := \left(({d_y}^2 + 2R_y) d_u M \right)^{-1}$.

We postpone the proof to Section A.III.7.2. Now that we have verified Condition A.III.1, Condition A.III.2 and Condition A.III.3, we can apply [BHS22, Theorem 2], to F, showing a convergence result on interpolated SGD trajectories.

Theorem A.III.1. Consider a neural network T and measures μ, ν satisfying Assumptions A.III.1 to A.III.5. Let $\alpha_1 < \alpha_0$ (see Proposition A.III.4).
Let $(u_\alpha^{(t)}), \alpha \in (0, \alpha_1], t \in \mathbb{N}$ a collection of SGD trajectories associated to Eq. (A.III.7). Consider (u_α) their associated interpolations. For any compact $\mathcal{K} \subset \mathbb{R}^{d_u}$ and any $\eta > 0$, we have:

$$\lim_{\substack{\alpha \longrightarrow 0 \\ \alpha \in\ (0, \alpha_1]}} \rho_0 \otimes \mu^{\otimes \mathbb{N}} \otimes \nu^{\otimes \mathbb{N}} \otimes \varpi^{\otimes \mathbb{N}} \left(d_c(u_\alpha, S_{-\partial_C F}(\mathcal{K})) > \eta \right) = 0. \tag{A.III.12}$$

The distance d_c is defined in Eq. (A.III.11). As the learning rate decreases, the interpolated trajectories approach the trajectory set $S_{-\partial_C F}$, which is essentially a solution of the *gradient flow equation* $\dot{u}(s) = -\nabla F(u(s))$ (ignoring the set of non-differentiability, which is $\mathscr{L}^{d_u}$-null). To get a tangible idea of the concepts at play, if F was $\mathcal{C}^2$ and had a finite amount of critical points, then one would have the convergence of a solution $u(s)$ to a critical point of F, as $s \longrightarrow +\infty$. These results have implicit consequences on the value of the parameters at the "end" of training for low learning rates, which is why we will consider a variant of SGD for which we can say more precise results on the convergence of the parameters.

A.III.4 Convergence of Noised Projected SGD Schemes on F

In practice, it is seldom desirable for the parameters of a neural network to reach extremely large values during training. Weight clipping is a common (although contentious) method of enforcing that $T(u, \cdot)$ stay Lipschitz, which is desirable for theoretical reasons. For instance the 1-Wasserstein duality in Wasserstein GANs [ACB17] requires Lipschitz networks, and similarly, Sliced-Wasserstein GANs [DZS18] use weight clipping and enforce their networks to be Lipschitz.

Given a radius $r > 0$, we consider SGD schemes that are restricted to $u \in \overline{B}(0, r) =: B_r$, by performing *projected* SGD. At each step t, we also add a noise $a\varepsilon^{(t+1)}$, where $\varepsilon^{(t+1)}$ is an additive noise of law $\xi \ll \mathscr{L}^{d_u}$, which is often taken as standard Gaussian in practice. These additions yield the following SGD scheme:

$$\begin{aligned} u^{(t+1)} &= \pi_r \left(u^{(t)} - \alpha \varphi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \theta^{(t+1)}) + \alpha a \varepsilon^{(t+1)} \right), \\ \left(u^{(0)}, (X^{(t)})_{t \in \mathbb{N}}\ (Y^{(t)})_{t \in \mathbb{N}},\ (\theta^{(t)})_{t \in \mathbb{N}},\ (\varepsilon^{(t)})_{t \in \mathbb{N}} \right) &\sim \rho_0 \otimes \mu^{\otimes \mathbb{N}} \otimes \nu^{\otimes \mathbb{N}} \otimes \varpi^{\otimes \mathbb{N}} \otimes \xi^{\otimes \mathbb{N}}, \end{aligned} \tag{A.III.13}$$

where $\pi_r : \mathbb{R}^u \longrightarrow B_r$ denotes the orthogonal projection on the ball $B_r := \overline{B}(0, r)$. Thanks to Conditions A.III.1 and A.III.2 and the additional noise, we can verify the assumptions for [BHS22, Theorem 4], yielding the same result as Theorem A.III.1 for the noised projected scheme Eq. (A.III.13). In fact, under additional assumptions, we shall prove a stronger mode of convergence for the aforementioned trajectories. The natural context in which to perform gradient descent is on functions that admit a chain rule, which is formalised in the case of almost-everywhere differentiability by the notion of *path differentiability*, as studied thoroughly in [BP21]. We also provide a brief presentation in the Appendix (Section A.III.7.3.1). We formulate this condition from [BHS22] before presenting sufficient conditions on T under which path differentiability shall hold.

Condition A.III.4. F is path differentiable, which is to say that for any $u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^{d_u})$, for almost all $s > 0$, $\forall v \in \partial_C F(u(s))$, $v^\top \dot{u}(s) = (F \circ u)'(s)$.

Remark A.III.1. There are alternate equivalent formulations for Condition A.III.4. Indeed, as presented in further detail in Section A.III.7.3.1, F is path differentiable if and only if $\partial_C F$ is a conservative field for F if and only if F has a chain rule for ∂_C (the latter is the formulation chosen above in Condition A.III.4).

In order to satisfy Condition A.III.4, we need to make the assumption that the NN input measure μ and the data measure ν are discrete measures, which is the case for ν in the case of generative neural networks, but is less realistic for μ in practice. We define $\triangle_n$ the n-simplex: its elements are the $a \in \mathbb{R}^n$ s.t. $\forall i \in [\![1, n]\!]$, $a_i > 0$ and $\sum_i a_i = 1$.

Assumption A.III.6. One may write $\mu = \sum_{k=1}^{n_x} a_k \delta_{x_k}$ and $\nu = \sum_{k=1}^{n_y} b_k \delta_{y_k}$, with the coefficient vectors $a \in \triangle_{n_x}, b \in \triangle_{n_y}$, $\mathcal{X} = \{x_1, \cdots, x_{n_x}\} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} = \{y_1, \cdots, y_{n_y}\} \subset \mathbb{R}^{d_y}$.

There is little practical reason to consider non-uniform measures, however the generalisation to any discrete measure makes no theoretical difference. Note that Assumption A.III.3 is clearly implied by Assumption A.III.6.

In order to show that F is path differentiable, we require the natural assumption that each $T(\cdot, x)$ be path differentiable. Since $T(\cdot, x)$ is a vector-valued function, we need to extend the notion of path-differentiability. Thankfully, [BP21, Definition 4] define *conservative mappings* for vector-valued locally Lipschitz functions, which allows us to define naturally path differentiability of a vector-valued function as the path-differentiability of all of its coordinate functions. See Section A.III.7.3.2 for a detailed presentation.

Assumption A.III.7. For any $x \in \mathbb{R}^{d_x}$, $T(\cdot, x)$ is path differentiable.

Assumption A.III.7 holds as soon as each the neural network has the typical structure of compositions of linear units and typical activations, as was proved by [Dav+20, Corollary 5.11], and [BP21, Section 6.2]. We provide a more specific class of NNs that are path differentiable and satisfy all our other assumptions in Section A.III.7.4. We can now leverage results from Chapter A.II and use the assumptions to show Condition A.III.4.

Proposition A.III.5. Under Assumptions A.III.2, A.III.6 and A.III.7, F is path differentiable.

Proof. We shall use repeatedly the property that the composition of path differentiable functions remains path differentiable, which is proved in [BP21, Lemma 6]. Let

$$\mathcal{E} : \left\{ \begin{array}{ccc} \mathbb{R}^{n \times d_y} \times \mathbb{R}^{n \times d_y} & \longrightarrow & \mathbb{R}_+ \\ Y, Y' & \longmapsto & \mathrm{SW}_2^2(\gamma_Y, \gamma_{Y'}) \end{array} \right. ,$$

by Proposition A.II.4, each $\mathcal{E}(\cdot, Y)$ is semi-concave and thus is path differentiable by Proposition A.II.11. Thanks to Assumption A.III.6, $\mu^{\otimes n}$ and $\nu^{\otimes n}$ are discrete measures on $\mathbb{R}^{n \times d_x}$ and $\mathbb{R}^{n \times d_y}$ respectively, allowing one to write $\mu^{\otimes n} = \sum_k a_k \delta_{X_k}$ and $\nu^{\otimes n} = \sum_l b_l \delta_{Y_l}$. Then $F = u \longmapsto \sum_{k,l} a_k b_l \mathcal{E}(T(u, X_k), Y_l)$ is path differentiable as a sum ([BP21, Corollary 4]) of compositions ([BP21, Lemma 6]) of path differentiable functions. □

We have now satisfied all the assumptions to apply [BHS22, Theorem 6], showing that trajectories of Eq. (A.III.13) converge towards to a set of generalised critical points[2] $\mathcal{C}_r$ defined

[2] Typically referred to as the set of *Karush-Kahn-Tucker* points of the differential inclusion $\dot{u}(s) \in -\partial_C F(u(s)) -$

as

$$\mathcal{C}_r := \left\{u \in \mathbb{R}^{d_u} \mid 0 \in -\partial_C F(u) - \mathcal{N}_r(u)\right\}, \quad \mathcal{N}_r(u) = \begin{cases} \{0\} \text{ if } \|u\|_2 < r \\ \{su \mid s \geq 0\} \text{ if } \|u\|_2 = r \\ \varnothing \text{ if } \|u\|_2 > r \end{cases}, \tag{A.III.14}$$

where $\mathcal{N}_r(u)$ refers to the *normal cone* of the ball $\overline{B}(0,r)$ at x. The term $\mathcal{N}_r(u)$ in Eq. (A.III.14) only makes a difference in the pathological case $\|u\|_2 = r$, which never happens in practice since the idea behind projecting is to do so on a very large ball, in order to avoid gradient explosion, to limit the Lipschitz constant and to satisfy theoretical assumptions. Omitting the $\mathcal{N}_r(u)$ term, and denoting $\mathcal{D}$ the points where F is differentiable, Eq. (A.III.14) simplifies to $\mathcal{C}_r \cap \mathcal{D} = \{u \in \mathcal{D} \mid \nabla F(u) = 0\}$, i.e. the critical points of F for the usual differential. Like in Theorem A.III.1, we let $\alpha_1 < \alpha_0$, where α_0 is defined in Proposition A.III.4. We have met the conditions to apply [BHS22, Theorem 6], showing a long-run convergence results on the SGD trajectories Eq. (A.III.13).

Theorem A.III.2. Consider a neural network T and measures μ, ν satisfying Assumptions A.III.1, A.III.2 and A.III.4 to A.III.7. Let $(u_\alpha^{(t)})_{t\in\mathbb{N}}$ be SGD trajectories defined by Eq. (A.III.13) for $r > 0$ and $\alpha \in (0, \alpha_1]$. One has

$$\forall \eta > 0, \ \overline{\lim_{t\longrightarrow+\infty}} \ \rho_0 \otimes \mu^{\otimes\mathbb{N}} \otimes \nu^{\otimes\mathbb{N}} \otimes \varpi^{\otimes\mathbb{N}} \otimes \xi^{\otimes\mathbb{N}} \left(d(u_\alpha^{(t)}, \mathcal{C}_r) > \eta\right) \xrightarrow[\substack{\alpha\longrightarrow 0 \\ \alpha\in(0,\alpha_1]}]{} 0.$$

The distance d above is the usual euclidean distance. Theorem A.III.2 shows essentially that as the learning rate approaches 0, the long-run limits of the SGD trajectories approach the set of $\mathcal{C}_r$ in probability. Omitting the points of non-differentiability and the pathological case $\|u\|_2 = r$, the general idea is that $u_\alpha^{(\infty)} \xrightarrow[\alpha\longrightarrow 0]{} \{u \ : \ \nabla F(u) = 0\}$, which is the convergence that would be achieved by the gradient flow of F, in the simpler case of $\mathcal{C}^2$ smoothness.

A.III.5 Convergence for Decreasing Steps in the Semi-Algebraic Setting

In this section, we present additional convergence results (novel to this thesis and not present in the paper on which this chapter is based), in the case where the neural network T is semi-algebraic, which is the case for NNs that are compositions of linear units and semi-algebraic activation functions, such as the ReLU. See Section A.III.7.3.4 for a detailed presentation of semi-algebraic sets and functions, and Section A.III.7.4 for a detailed proof of semi-algebraicity for semi-algebraic activations.

Assumption A.III.8. The neural network T is semi-algebraic.

The main result of this section hinges on the recent work [BLP23], which only came to our knowledge when working on the paper on which Chapter B.I is based. To circumvent technical difficulties, we will study a variant of the Sliced Wasserstein distance which takes the expectation over directions $\eta \sim \varpi_B := \mathcal{U}(\mathcal{B}_{d_y})$, the uniform distribution on the closed unit ball $\mathcal{B}_{d_y} \subset \mathbb{R}^{d_y}$. This way, the probability measure ϖ_B has a semi-algebraic density with respect to the Lebesgue measure on $\mathbb{R}^{d_y}$, which falls under the conditions of [BLP23]. Our sample loss function thus reads:

$$g := \begin{cases} \mathbb{R}^{d_u} \times \mathbb{R}^{n\times d_x} \times \mathbb{R}^{n\times d_y} \times \mathcal{B}_{d_y} & \longrightarrow & \mathbb{R}^{d_y} \\ (u, X, Y, \eta) & \longmapsto & \mathrm{W}_2^2(P_\eta \# T_u \# \gamma_X, P_\eta \# \gamma_Y) \end{cases}, \tag{A.III.15}$$

and the associated population loss is defined as

$$G(u) := \int_{\mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{B}_{d_y}} g(u, X, Y, \eta) \mathrm{d}\mu^{\otimes n}(X) \mathrm{d}\nu^{\otimes n}(Y) \mathrm{d}\varpi_B(\eta), \tag{A.III.16}$$

$\mathcal{N}_r(u(s))$.

and a simple change-of-variables shows that there exists a constant $c > 0$ such that $F = cG$. Noticing for $\eta \in \mathcal{B}_{d_y} \setminus \{0\}$ and $Y, Y' \in \mathbb{R}^{n \times d_y}$, the simple computation

$$w_\eta(Y, Y') := \mathrm{W}_2^2(P_\eta \# T_u \# \gamma_X, P_\eta \# \gamma_Y) = \|\eta\|_2 w_{\eta/\|\eta\|_2}(X, Y),$$

all our previous considerations on w can be adapted immediately to this setting. In particular, under Assumption A.III.8, we can introduce the function ψ defined by:

$$\psi : \begin{cases} \mathbb{R}^{d_u} \times \mathbb{R}^{n \times d_x} \times \mathbb{R}^{n \times d_y} \times \mathcal{B}_{d_y} & \longrightarrow & \mathbb{R}^{d_u} \\ (u, X, Y, \eta) & \longmapsto & \displaystyle\sum_{k=1}^{n} \frac{2}{n} \left(\frac{\partial T}{\partial u}(u, x_k) \right)^\top \eta\eta^\top (T(u, x_k) - y_{\sigma_\eta^{T(u,X),Y}(k)}) \end{cases}, \tag{A.III.17}$$

where $(u, x) \longmapsto \frac{\partial T}{\partial u}(u, x)$ refers to a fixed semi-algebraic selection of $(u, x) \longmapsto [\partial_C T(\cdot, x)](u)$, which exists since T is semi-algebraic. The map ψ defines a semi-algebraic selection of the Clarke differential of g with respect to the parameters u. The resulting decreasing-step SGD scheme is the following:

$$\begin{aligned} u^{(t+1)} &= u^{(t)} - \alpha_t \psi(u^{(t)}, X^{(t+1)}, Y^{(t+1)}, \eta^{(t+1)}), \\ u_0 \in \mathbb{R}^{d_u},\ &\left((X^{(t)})_{t \in \mathbb{N}}\ (Y^{(t)})_{t \in \mathbb{N}}\ (\eta^{(t)})_{t \in \mathbb{N}} \right) \sim \mu^{\otimes \mathbb{N}} \otimes \nu^{\otimes \mathbb{N}} \otimes \mathfrak{o}_B^{\otimes \mathbb{N}}, \end{aligned} \tag{A.III.18}$$

To apply [BLP23, Theorem 3], we need will require assumptions on the data measures μ and ν:

Assumption A.III.9. The measures μ and ν are either discrete or admit a semi-algebraic density with respect to the Lebesgue measure.

Assumption A.III.9 is very mild, since in practice, target datasets ν are discrete and source μ are either uniform of Gaussian (in practice approximated semi-algebraically by computers to numerical precision). We now have all the elements to apply [BLP23, Theorem 3]:

Theorem A.III.3. Consider a NN T and measures μ, ν verifying Assumptions A.III.8 and A.III.9. Assume that the gradient steps $(\alpha)_{t \in \mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ are such that $\sum_t \alpha_t = +\infty$ and $\alpha_t = o(1/\log(t))$.
Then there exists a set $A \subset (0, +\infty)$ of possible steps with finite complementary and a set $U_0 \subset \mathbb{R}^{d_u}$ of full measure of possible initialisations u_0, such that iterates (u_t) of the SGD scheme Eq. (A.III.18) with $u_0 \in U_0$ and $(\alpha_t) \subset A$ verify the following properties almost-surely, conditionally to (u_t) bounded:

- The sequence $(G(u_t))$ converges;
- Any accumulation point $\overline{u}$ of (u_t) is Clarke critical: $0 \in \partial_C G(\overline{u})$.

Proof. We apply [BLP23, Theorem 3], noticing that the sample loss g is semi-algebraic, thanks to Assumption A.III.8. Using the boundedness conditioning formulated in the result statement, Assumption A.III.9, Proposition B.I.8 and the construction of a semi-algebraic sub-gradient selection ψ, we have collected all the conditions for [BLP23, Theorem 3], yielding the result. For the case where one or both of $\{\mu, \nu\}$ is/are discrete, we applied [BLP23, Remark 3]. □

A.III.6 Conclusion and Outlook

Under reasonable assumptions, we have shown that SGD trajectories of parameters of generative NNs with a minibatch SW loss converge towards the desired sub-gradient flow solutions, implying in a weak sense the convergence of said trajectories. Under stronger assumptions, we have shown that trajectories of a mildly modified SGD scheme converge towards a set of generalised critical points of the loss, which provides a missing convergence result for such optimisation problems.

The core limitation of Theorem A.III.1 is the assumption that the input data measure μ is discrete (Assumption A.III.6), which we required in order to prove that the loss F is path differentiable. Using [BLP23], we were able to circumvent the issue of path differentiability by proving convergence for decreasing-step SGD under mild semi-algebraicity assumptions.

Our studies focus on the 2-SW distance, but our results from Section A.III.2 can be extended to $p \in [1, +\infty)$, as presented in the appendix (Section A.III.7.5). However, as also discussed in the Appendix, the generalisation of Section A.III.4 is still an open problem, since it has not yet be proven that $X \longmapsto \mathrm{SW}_p^p(\gamma_X, \gamma_Y)$ is path differentiable for $p \neq 2$.

This chapter studies the use of the *average* SW distance as a loss, and an extension to related distances would be worth considering. The average SW distance aggregates the projected distances through an expectation, while the closely-related *max*-Sliced Wasserstein distance introduced by [Des+19] aggregates the projections via a maximisation on the axis $\theta \in \mathbb{S}^{d-1}$. The training paradigm presented in [Des+19] differs strongly from our formalism since it applies to GANs, however one could consider an extension of our formalism in which the optimal projection θ becomes a learned parameter of the neural network. A related extension is the Subspace-Robust Wasserstein distance [PC19a], which can take the following formulation

$$\mathcal{S}_k^2(\mu, \nu) = \max_{\substack{0 \preceq \Omega \preceq I_d \\ \mathrm{trace}(\Omega) = k}} \mathrm{W}_2^2(\Omega^{1/2} \# \mu, \Omega^{1/2} \# \nu),$$

for which one could consider a similar extension where the positive semi-definite Ω becomes a learned parameter of T.

Another avenue for future study would be to tie the flow approximation result from Theorem A.III.1 to Sliced Wasserstein Flows [Liu+19; Bon+22]. The difficulty in seeing the differential inclusion Eq. (A.III.9) as a flow of F lies in the non-differentiable nature of the functions at play, as well as the presence of the composition between SW and the neural network T, which bodes poorly with Clarke sub-differentials.

Acknowledgements

We thank Julie Delon for proof-reading and general feedback, as well as Rémi Flamary and Alain Durmus for fruitful discussions.

A.III.7 Appendix

A.III.7.1 Table of Notations

Table A.III.1: List of Notations

Symbol	Explanation
γ_X	Given $X = (x_1, \cdots, x_n) \in \mathbb{R}^{n\times d}$, $\gamma_X = \frac{1}{n}\sum_i \delta_{x_i}$
X	$(x_1, \cdots, x_n) \in \mathbb{R}^{n\times d_x}$ an input data sample of law $\mu^{\otimes n}$
μ	input data probability measure on $\mathbb{R}^{d_x}$, supported on $\mathcal{X}$
Y	$(y_1, \cdots, y_n) \in \mathbb{R}^{n\times d_y}$ a target data sample of law $\nu^{\otimes n}$
ν	target data probability measure on $\mathbb{R}^{d_y}$, supported on $\mathcal{Y}$
θ	direction in $\mathbb{S}^{d_y-1}$
ϖ	uniform measure on $\mathbb{S}^{d_y-1}$
$z := (X, Y, \theta)$	sample in X, Y and θ
$\zeta := \mu^{\otimes n} \otimes \nu^{\otimes n} \otimes \varpi$	probability measure for the samples z, supported on $\mathcal{Z} := \mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y-1}$
u	neural network parameters in $\mathbb{R}^{d_u}$
$T(u, X)$	neural network function defined in Eq. (A.III.3)
$f(u, X, Y, \theta)$	sample loss function defined in Eq. (A.III.4)
$F(u)$	population loss function defined in Eq. (A.III.5)
$w_\theta(Y, Y')$	discrete and projected 2-Wasserstein distance $\mathrm{W}_2^2(P_\theta \# \gamma_Y, P_\theta \# \gamma_{Y'})$
$\varphi(u, X, Y, \theta)$	almost-everywhere gradient of $f(\cdot, X, Y, \theta)$ defined in Eq. (A.III.6)
K_w, K_f, K_F	local Lipschitz constants of w, f, F respectively (see Propositions 1, 2, 3)
$\alpha; a$	SGD learning rate; noise level
$\mathscr{L}^d; \rho \ll \mathscr{L}^d$	Lebesgue measure on $\mathbb{R}^d$; a measure ρ absolutely continuous w.r.t. $\mathscr{L}^d$
∂_C	Clarke differential, defined in Eq. (A.III.8)
ρ_0	probability measure of SGD initialisation $u^{(0)}$
$\varepsilon^{(t)}$	additive noise in $\mathbb{R}^{d_u}$ at SGD step t
ξ	additive noise probability measure on $\mathbb{R}^{d_u}$
$B_{\|\cdot\|}(x, R), \overline{B}_{\|\cdot\|}(x, R)$	open (resp. closed) ball of centre x and radius R for the norm $\|\cdot\|$

A.III.7.2 Postponed Proofs

Proof of Proposition A.III.4. For the sake of legibility, we will use the concise notation $\partial_x f(x, y)$ for $\frac{\partial f}{\partial x}(x, y)$ for partial derivatives of a function, and $\partial_x(f(x, y))$ for partial derivatives of an expression.

Proof. Let $\rho \ll \mathscr{L}^{d_y}$ and $B \in \mathcal{B}(\mathbb{R}^{d_u})$ such that $\mathscr{L}(B) = 0$. We have, with $\alpha' := 2\alpha/n$, $z := (X, Y, \theta)$, $\zeta := \mu^{\otimes n} \otimes \nu^{\otimes n} \otimes \varpi$ and $\mathcal{Z} := \mathcal{X}^n \times \mathcal{Y}^n \times \mathbb{S}^{d_y-1}$,

$$\begin{aligned}\rho P_\alpha(B) &= \int_{\mathbb{R}^{d_u}\times\mathcal{Z}} \mathbb{1}_B \left[u - \alpha' \sum_{k=1}^n \left(\partial_u T(u, x_k)\right)^\top \theta\theta^\top (T(u, x_k) - y_{\sigma_\theta^{T(u,X),Y}(k)})\right] \mathrm{d}\rho(u)\mathrm{d}\zeta(z) \\ &\leq \sum_{\tau\in\mathfrak{S}_n} \int_{\mathcal{Z}} I_\tau(z)\mathrm{d}\zeta(z),\end{aligned}$$

where $I_\tau(z) := \int_{\mathbb{R}^{d_u}} \mathbb{1}_B\left(\phi_{\tau,z}(u)\right) \mathrm{d}\rho(u)$, with $\phi_{\tau,z} := u - \alpha' \underbrace{\sum_{k=1}^n \left(\partial_u(u, x_k)\right)^\top \theta\theta^\top (T(u, x_k) - y_{\tau(k)})}_{\psi_{\tau,z} :=}$.

Let $\tau \in \mathfrak{S}_n$ and $(X, Y, \theta) \in \mathcal{Z}$. Using Assumption A.III.1, separate

$$I_\tau(z) = \sum_{j\in J} \int_{\mathcal{U}_j(X)} \mathbb{1}_B\left(u - \psi_{\tau,z}(u)\right) \mathrm{d}\rho(u),$$

where the differentiability structure $(\mathcal{U}_j(X))_{j\in J(X)}$ is obtained using the respective differentiability structures: for each $k \in [\![1,n]\!]$, Assumption A.III.1 yields a structure $(\mathcal{U}_{j_k}(x_k))_{j_k\in J_k(x_k)}$ of $u \longmapsto T(u,x_k)$, which depends on x_k, hence the k indices.

To be precise, define for $j=(j_1,\cdots,j_n)\in J_1(x_1)\times\cdots\times J_n(x_n)$, $\mathcal{U}_j(X) := \bigcap_{k=1}^n \mathcal{U}_{j_k}(x_k)$, and $J(X) := \{(j_1,\cdots,j_n)\in J_1(x_1)\times\cdots\times J_n(x_n) \mid \mathcal{U}_j(X)\neq\varnothing\}$. In particular, for any $k\in[\![1,n]\!]$, $T(\cdot,x_k)$ is $\mathcal{C}^2$ on $\mathcal{U}_j(X)$. Notice that the derivatives are not necessarily defined on the border $\partial\mathcal{U}_j(X)$, which is of Lebesgue measure 0 by Assumption A.III.1, thus the values of the derivatives on the border do not change the value of the integrals (the integrals may have the value $+\infty$, depending on the behaviour of $\phi_{\tau,s}$, but we shall see that they are all finite when α is small enough).

We drop the z,τ index in the notation, and focus on the properties of ϕ and ψ as functions of u. Our first objective is to determine a constant $K>0$, independent of u,z,τ, such that ψ is K-Lipschitz on $\mathcal{U}_j(X)$. First, let

$$\chi := \left\{\begin{array}{ccc} \mathcal{U}_j(X) & \longrightarrow & \mathbb{R}^{d_u} \\ u & \longmapsto & \left(\partial_u T(u,x_k)\right)^\top \theta\theta^\top T(u,x_k)\end{array}\right. .$$

The function χ is of class $\mathcal{C}^1$, therefore we determine its Lipschitz constant by upper-bounding the $\|\cdot\|_2$-induced operator norm of its differential, denoted by $|||\partial_u\chi(u)|||_2$. Notice that $\chi(u) = \frac{1}{2}\partial_u\left(\theta^\top T(u,x_k)\right)^2$. First we upper-bound:

$$\left|\left|\left|\partial_u^2\left(\theta^\top T(u,x_k)\right)^2\right|\right|\right|_2 \leq d_u \max_{(i_1,i_2)\in[\![1,d_u]\!]^2}\left|\partial_{u_{i_1}}\partial_{u_{i_2}}\left(\theta^\top T(u,x_k)\right)^2\right|,$$

then we use Assumption A.III.5 and $|\theta_i|\leq 1$,

$$\left|\partial_{u_{i_1}}\partial_{u_{i_2}}\left(\theta^\top T(u,x_k)\right)^2\right| \leq \sum_{(i_3,i_4)\in[\![1,d_y]\!]^2}\left|\theta_{i_3}\theta_{i_4}\times\partial_{u_{i_1}}\partial_{u_{i_2}}\left([T(u,x_k)]_{i_3}[T(u,x_k)]_{i_4}\right)\right| \leq {d_y}^2 M.$$

We obtain that χ is $\frac{1}{2}d_u{d_y}^2M$-Lipschitz. Second, let

$$\omega : u\in\mathcal{U}_j(X)\longmapsto \left(\partial_u T(u,x_k)\right)^\top\theta\theta^\top y_{\tau(k)},$$

also of class $\mathcal{C}^1$. We re-write

$$[\partial_u\omega(u)]_{i_1,i_2} = y_{\tau(k)}^\top\theta\theta^\top\partial_{u_{i_1}}\partial_{u_{i_2}}T(u,x_k),$$

and conclude similarly by Assumption A.III.5 that ω is $\|y_{\tau(k)}\|_2 d_u M$-Lipschitz.

Finally, $\psi=\sum_{k=1}^n(\chi_k-\omega_k)$, and is therefore $K := (\frac{1}{2}{d_y}^2+R_y)d_u nM$-Lipschitz, with R_y from Assumption A.III.3. We have proven that $|||\partial_u\psi(u)|||_2\leq K$ for any $u\in\mathcal{U}_j(X)$, and that K does not depend on X,Y,θ,j or u.

We now suppose that $\alpha'<\frac{1}{K}$, which is to say $\alpha<\frac{n}{2K}$. Under this condition, $\phi:\mathcal{U}_j(X)\longrightarrow\mathbb{R}^{d_u}$ is injective. Indeed, if $\phi(u_1)=\phi(u_2)$, then $\|u_1-u_2\|_2=\alpha'\|\psi(u_1)-\psi(u_2)\|_2\leq\alpha' K\|u_1-u_2\|_2$, thus $u_1=u_2$. Furthermore, for any $u\in\mathcal{U}_j(X)$, $\partial_u\phi(u)=\mathrm{Id}_{\mathbb{R}^{d_u}}-\alpha'\partial_u\psi(u)$, with $|||\alpha'\partial_u\psi(u)|||_2<1$, thus the matrix $\partial_u\psi(u)$ is invertible (using the Neumann series method). By the global inverse function theorem, $\phi:\mathcal{U}_j(X)\longrightarrow\phi(\mathcal{U}_j(X))$ is a $\mathcal{C}^1$-diffeomorphism. Using the change-of-variables formula, we have

$$\int_{\mathcal{U}_j(X)}\mathbb{1}_B(\phi(u))\mathrm{d}\rho(u) = \int_{\mathcal{U}_j(X)}\mathbb{1}_B(u')\mathrm{d}\phi\#\rho(u') = \phi\#\rho(B),$$

we have now shown that ϕ is a $\mathcal{C}^1$-diffeomorphism, thus since $\rho\ll\mathscr{L}$, $\phi\#\rho\ll\mathscr{L}$. ($\alpha\ll\beta$ denoting that α is absolutely continuous with respect to β). Since $\mathscr{L}(B)=0$, it follows that the integral is 0, then by sum over j, $I_\tau(z)=0$ and finally $\rho P_\alpha(B)=0$ by integration over z and sum over τ. □

A.III.7.3 Background on Non-Smooth and Non-Convex Analysis

This work is placed within the context of non-smooth optimisation, a field of study in part introduced by Clarke with the so-called Clarke differential, which we introduced in Equation Eq. (A.III.8) (see [Cla90] for a general reference on this object). The purpose of this appendix is to present several adjacent objects that can be useful to the application of our results, even though we do not need them in order to prove our theorems.

A.III.7.3.1 Conservative Fields

The Clarke differential ∂_C of a locally Lipschitz function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ (defined in Equation Eq. (A.III.8)) is an example of a *set-valued map*. Such a map is a function $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ from the subsets of $\mathbb{R}^p$ to the subsets of $\mathbb{R}^q$, for instance in the case of the Clarke differential, we have the signature $\partial_C g : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. A set-valued map D is *graph closed* if its graph $\{(u, v) \mid u \in \mathbb{R}^p,\ v \in D(u)\}$ is a closed set of $\mathbb{R}^{p+q}$. A set-valued map D is said to be a *conservative field*, when it is graph closed, has non-empty compact values and for any absolutely continuous loop $\gamma \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p)$ with $\gamma(0) = \gamma(1)$, we have

$$\int_0^1 \max_{v \in D(\gamma(s))} \langle \dot{\gamma}(s), v \rangle \mathrm{d}s = 0.$$

Similarly to primitive functions in calculus, one may define a function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ using a conservative field $D : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ up to an additive constant through following expression:

$$g(u) = g(0) + \int_0^1 \max_{v \in D(\gamma(s))} \langle \dot{\gamma}(s), v \rangle \mathrm{d}s, \quad \forall \gamma \in \mathcal{C}_{\text{abs}}([0, 1], \mathbb{R}^p) \text{ such that } \gamma(0) = 1 \text{ and } \gamma(1) = u. \tag{A.III.19}$$

In this case, we say that g is a *potential function* for the field D. This notion allows us to define a new regularity class: a function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ is called *path differentiable* when there exists a conservative field of which it is a potential. A standard result in non-smooth optimisation is the following equivalence between different notions of regularity:

Proposition A.III.6. [BP21, Corollary 2] Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ locally Lipschitz. We have the equivalence between the following statements:

- g is path differentiable
- $\partial_C g$ is a conservative field
- g has a *chain rule* for the Clarke differential ∂_C:

$$\forall u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^d),\ \underline{\forall} s > 0,\ \forall v \in \partial_C g(u(s)),\ v^\top \dot{u}(s) = (g \circ u)'(s). \tag{A.III.20}$$

This equivalence justifies the terminology used in Condition A.III.4. The reader seeking a complete presentation of conservative field theory may refer to [BP21].

A.III.7.3.2 Conservative Mappings

The notion of conservative fields for real-valued locally Lipschitz functions $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ can be generalised to *conservative mappings* for vector-valued locally Lipschitz functions $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$, which one may see as a generalised Jacobian matrix (see [BP21, Section 3.3] for further details). A set-valued map $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{q \times p}$ is a conservative mapping for such a g if

$$\forall u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^p),\ \underline{\forall} s > 0,\ (g \circ u)'(s) = M \dot{u}(t),\ \forall M \in J(u(s)). \tag{A.III.21}$$

In this case, we shall say that g is path differentiable. Note that if each coordinate function g_i is the potential of a conservative field D_i, then the set-valued map

$$J(u) = \left\{ \begin{pmatrix} v_1^\top \\ \vdots \\ v_q^\top \end{pmatrix} \ : \ \forall i \in [\![1, q]\!], \ v_i \in D_i(u) \right\}$$

is a conservative mapping for g (although not all conservative mappings for g can be written in this manner). As a consequence, one could interpret (simplistically) vector-valued path differentiability as coordinate-wise path differentiability.

A.III.7.3.3 Clarke Regularity

Another notion of regularity for locally Lipschitz functions is that of *Clarke regularity*. Let $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ and $u \in \mathbb{R}^p$, g is said to be *Clarke regular* at u if the two quantities

$$g^\circ(u; v) := \limsup_{\substack{u' \to u \\ t \searrow 0}} \frac{g(u' + tv) - g(u')}{t} \quad \text{and} \quad g'(u; v) := \lim_{t \searrow 0} \frac{g(u + tv) - g(u)}{t}$$

exist and are equal for all $v \in \mathbb{R}^p$. Note that this notion implies path differentiability by [BP21, Proposition 2]. Clarke regularity is the central concept of Clarke's monograph [Cla90].

A.III.7.3.4 Semi-Algebraic Functions

In non-smooth analysis, one of the simplest regularity cases is the class of *semi-algebraic* functions, which are essentially piecewise polynomial functions defined on polynomial pieces. To be precise, a set $\mathcal{A} \subset \mathbb{R}^d$ is *semi-algebraic* if it can be written under the form

$$\mathcal{A} = \bigcup_{i=1}^{n} \bigcap_{j=1}^{m} \left\{ u \in \mathbb{R}^d \mid P_{i,j}(u) < 0, \ Q_{i,j}(u) = 0 \right\},$$

where the $P_{i,j}$ and $Q_{i,j}$ are real multivariate polynomials. A function $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ is *semi-algebraic* if its graph $\mathcal{G} := \{(u, g(u)) \mid u \in \mathbb{R}^p\}$ is semi-algebraic.

A locally Lipschitz real-valued semi-algebraic function is path differentiable (see for instance [BP21, Proposition 2]), and in the light of [BP21, Lemma 3], this is also the case in the vector-valued case. Another useful property of semi-algebraic functions is that their class is stable by composition and product. The interested reader may consult [Wak08] for additional properties of semi-algebraic objects, or [Cos99; VM96], for a presentation of o-minimal structures, a generalisation of this concept.

A.III.7.4 Suitable Neural Networks

In this section, we detail our claim that typical NN structures satisfy our conditions. To this end, we define a class of practical neural networks whose properties are sufficient (not all NNs that satisfy our assumptions are within this framework). Consider $\mathcal{T}$ the set of NNs T of the form

$$T : \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_y} \\ (u, x) & \longmapsto & \widetilde{T}(u, x) \mathbb{1}^{\varepsilon}_{B(0,R_u)}(u) \mathbb{1}^{\varepsilon}_{B(0,R_x)}(x) \end{array} \right. ,$$

with $R_u, R_x > 0$ and $\varepsilon > 0$. The function $\mathbb{1}^{\varepsilon}_{B(0,R)}$ is a smoothed version of the usual indicator function $\mathbb{1}_{B(0,R)}$: it is any function that has value 1 in $B(0, R - \varepsilon)$, 0 outside $B(0, R + \varepsilon)$ and is $\mathcal{C}^2$-smooth (see Remark A.III.2 for a possible construction). Given that one may take arbitrarily large radii, these indicators are added for theoretical purposes and impose no realistic constraints in practice. Additionally, $\widetilde{T} = h_N$, the N-th layer of a recursive NN structure defined by

$$h_0(u, x) = x, \quad \forall n \in [\![1, N]\!], \ h_n = \left\{ \begin{array}{ccc} \mathbb{R}^{d_u} \times \mathbb{R}^{d_x} & \longrightarrow & \mathbb{R}^{d_n} \\ (u, x) & \longmapsto & a_n \left(\sum_{i=0}^{n-1} A_{n,i}(u) h_i(u, x) + B_n u \right) \end{array} \right. ,$$

where:

- *All* functions $a_n : \mathbb{R} \longrightarrow \mathbb{R}$ are $\mathcal{C}^2$-smooth, or *all* locally Lipschitz semi-algebraic activation functions (applied entry-wise). The former condition is satisfied by the common sigmoid, hyperbolic tangent or softplus activations. The latter condition applies to the non-differentiable ReLU activation, its "Leaky ReLU" extension, and continuous piecewise polynomial activations. Note that other non-linearities such as softmax can also be considered under the same regularity restrictions, but we limit ourselves to entry-wise non-linearities for notational consistency.
- Each dimension d_n is a positive integer, with obviously $d_N = d_y$, the output dimension.
- Each $A_{n,i}$ is a linear map: $\mathbb{R}^{d_u} \longrightarrow \mathbb{R}^{d_n \times d_i}$, which maps a parameter vector u to a $d_n \times d_i$ matrix. Since the entire parameter vector u is given at each layer, this allows the architecture to only use certain parameters at each layer (as is more typical in practice). One may see this map as a 3-tensor of shape (d_n, d_i, d_u), as specified in the formulation

$$\forall u \in \mathbb{R}^{d_u},\ \forall h \in \mathbb{R}^{d_i},\ A_{n,i}(u)h = \left(\sum_{j_2=1}^{d_i} \sum_{j_3=1}^{d_u} A^{(n,i)}_{j_1,j_2,j_3} h_{j_2} u_{j_3} \right)_{j_1 \in [\![1,d_n]\!]} \in \mathbb{R}^{d_n}. \tag{A.III.22}$$

- The matrix $B_n \in \mathbb{R}^{d_n \times d_u}$ determines the intercept from the full parameter vector u.

In this model, each layer depends on all the previous layers, allowing for residual inputs for instance. Overall, all typical networks fit this description, once bounded using the indicator functions, with only a technicality on the regularities of the activations which need to be *all* $\mathcal{C}^2$-smooth, or *all* semi-algebraic. One could extend this class of NNs to those with *definable* activations within the same o-minimal structure (similarly to [Dav+20] and [BP21]), however mixing these concepts with our piecewise $\mathcal{C}^2$ assumption (Assumption A.III.1) is beyond the scope of this chapter. The difficulty lies in the condition that the borders of the differentiable structure are Lebesgue-null.

Remark A.III.2. We mention an explicit construction of a $\mathcal{C}^\infty$-smooth indicator $\mathbb{1}^\varepsilon_{B(0,R)}$:

$$f(s) := \begin{cases} e^{-1/s} \text{ if } s > 0 \\ \text{else } 0 \end{cases}, \quad g(s) := \frac{f(s)}{f(s) + f(1-s)},$$
$$\mathbb{1}^\varepsilon_{B(0,R)} := \begin{cases} \mathbb{R}^d & \longrightarrow & [0,1] \\ u & \longmapsto & g\left(\dfrac{(R+\varepsilon)^2 - \|u\|_2^2}{4R\varepsilon} \right) \end{cases}.$$

Before proving the properties of NNs from the class $\mathcal{T}$, we require a technical result on path differentiable functions.

Proposition A.III.7. Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ path differentiable, and $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ of class $\mathcal{C}^1$. Then their product fg is path differentiable.

Proof. Our objective is to apply [BP21, Corollary 2] (stated in Proposition A.III.6), which is to say that $h := fg$ admits a chain rule for $\partial_C h$. First, we apply the definition of the Clarke differential and compute

$$\forall u \in \mathbb{R}^d,\ \partial_C f(u) = f(u)\nabla g(u) + g(u)\partial_C f(u) := \{ f(u)\nabla g(u) + g(u)v \mid v \in \partial_C f(u) \}.$$

Note that we used the smoothness of g. We now consider an absolutely continuous curve $u \in \mathcal{C}_{\text{abs}}(\mathbb{R}_+, \mathbb{R}^d)$. By [BP21, Lemma 2], since f is path differentiable, $f \circ u$ is differentiable almost everywhere. Let D the associated set of differentiability, then let $s \in D$ and $v \in \partial_C h(u(s))$, writing $v = f(u(s))\nabla g(u(s)) + g(u(s))w$ with $w \in \partial_C f(u(s))$. We compute $(h \circ u)'(s) = (f \circ$

$u)'(s)g(u(s)) + f(u(s))(g \circ u)'(s)$. Now since f is path differentiable and $w \in \partial_C f(u(s))$, by Proposition A.III.6 item 3, we have $(f \circ u)'(s) = \langle w, \dot{u}(s)\rangle$. On the other hand, $(g \circ u)'(s) = \langle \nabla g(u(s)), \dot{u}(s)\rangle$ since g is $\mathcal{C}^1$. Finally by definition of v and bilinearity of $\langle \cdot, \cdot \rangle$,

$$(h \circ u)'(s) = \langle w, \dot{u}(s)\rangle g(u(s)) + f(u(s))\langle \nabla g(u(s)), \dot{u}(s)\rangle = \langle v, \dot{u}(s)\rangle. \qquad \square$$

We now have all the tools to prove that the class of NNs $\mathcal{T}$ satisfies all of the assumptions of our paper.

Proposition A.III.8. All networks of the class $\mathcal{T}$ verify Assumption A.III.1, Assumption A.III.2, Assumption A.III.4, Assumption A.III.5 and Assumption A.III.7.

Proof. Let $T \in \mathcal{T}$, and $\widetilde{T}$ its associated underlying network. We begin with regularity considerations.

Verifying Assumptions 1 and 7 in the $\mathcal{C}^2$ Case. In the case where the activations are $\mathcal{C}^2$-smooth, then each $\widetilde{T}(\cdot, x)$ is also of class $\mathcal{C}^2$. Furthermore, the smooth indicator $\mathbb{1}^{\varepsilon}_{B(0,R_u)}$ is $\mathcal{C}^\infty$-smooth, thus we can conclude that $T(\cdot, x)$ is $\mathcal{C}^2$-smooth, and thus satisfies Assumption A.III.1 trivially. In particular, $T(\cdot, x)$ is path differentiable for any $x \in \mathbb{R}^{d_x}$, thus T also satisfies Assumption A.III.7.

Verifying Assumptions 1 and 7 in the Semi-Algebraic Case. In the case where the activations are locally Lipschitz and semi-algebraic, it follows that each $\widetilde{T}(\cdot, x)$ is semi-algebraic, which yields naturally a differentiability structure associated to the polynomial pieces, satisfying Assumption A.III.1. Furthermore, this regularity yields path differentiability by [BP21, Proposition 2]. By product with the smooth indicator function, T is path differentiable by Proposition A.III.7, therefore it satisfies Assumption A.III.7 .

Verifying Assumption 2 in the $\mathcal{C}^2$ Case. In the case where the activations are $\mathcal{C}^2$-smooth, it is clear that by composition and product $(u, x) \longmapsto \widetilde{T}(u, x)$ is *jointly* $\mathcal{C}^2$-smooth. As a consequence, it is Lipschitz jointly in (u, x) on any compact of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_y}$, and by product with the smooth indicators, so is T. Since T is zero outside $\overline{B}(0, R_u + \varepsilon) \times \overline{B}(0, R_x + \varepsilon)$, we conclude that it is globally Lipschitz in (u, x).

Verifying Assumption 2 in the Semi-Algebraic Case. In the case of locally Lipschitz and semi-algebraic activations, we prove that $\widetilde{T}$ is jointly Lipschitz on any compact $\mathcal{K}$ by strong induction on $n \in [\![1, N]\!]$. Let $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$ a product compact of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_y}$, and P_n : "$\exists L_n > 0 : h_n$ is L_n-Lipschitz on $\mathcal{K}$". The initialisation P_0 is trivial, since $z(u, x) = x$. Let $n \in [\![1, N]\!]$ and assume P_i to hold true for $i \in [\![0, n-1]\!]$. In particular, the h_i are jointly continuous in (u, x), allowing the definition of

$$M := \max_{(u,x)\in\mathcal{K}} \left| \sum_{i=0}^{n-1} A_{n,i}(u) h_i(u, x) + B_n u \right|.$$

Since a_n is locally Lipschitz, a covering argument shows that there exists $L_{a_n} > 0$ such that a_n is L_{a_n}-Lipschitz on $[-M, M]$. Now let $(u_1, u_2) \in \mathcal{K}_1^2$ and $(x_1, x_2) \in \mathcal{K}_2^2$. We have

$$\begin{aligned}
&\|h_n(u_1, x_1) - h_n(u_2, x_2)\|_2 \\
&\leq L_{a_n} \left\| \sum_{i=0}^{n-1} A_{n,i}(u_1) h_i(u_1, x_1) + B_n u_1 - \sum_{i=0}^{n-1} A_{n,i}(u_2) h_i(u_2, x_2) - B_n u_2 \right\|_2 \\
&\leq L_{a_n} \left(\|B_n\|_{\mathrm{op}} \|u_1 - u_2\|_2 + \sum_{i=0}^{n-1} \|A_{n,i}(u_1) h_i(u_1, x_1) - A_{n,i}(u_2) h_i(u_2, x_2)\|_2 \right), \qquad \text{(A.III.23)}
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the $\|\cdot\|_2$-induced operator norm. Let $i \in [\![0, n-1]\!]$, we separate the norm:

$$\begin{aligned}\|A_{n,i}(u_1)h_i(u_1,x_1) - A_{n,i}(u_2)h_i(u_2,x_2)\|_2 \leq \|A_{n,i}(u_1)h_i(u_1,x_1) - A_{n,i}(u_2)h_i(u_1,x_1)\|_2 =: \Delta_1 \\ + \|A_{n,i}(u_2)h_i(u_1,x_1) - A_{n,i}(u_2)h_i(u_2,x_2)\|_2 =: \Delta_2.\end{aligned} \tag{A.III.24}$$

For Δ_1, use the tensor form Eq. (A.III.22) and the inequality $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$ for $x \in \mathbb{R}^d$, then $\|u\|_\infty \leq \|u\|_2$:

$$\begin{aligned}\Delta_1 &\leq \sqrt{d_n}\left\| \left(\sum_{j_2=1}^{d_i} \sum_{j_3=1}^{d_u} A^{(n,i)}_{j_1,j_2,j_3} h_i(u_1,x_1)_{j_2}(u^{(1)}_{j_3} - u^{(2)}_{j_3}) \right)_{j_1 \in [\![1,d_n]\!]} \right\|_\infty \\ &\leq \sqrt{d_n} \max_{j_1,j_2,j_3} |A^{(n,i)}_{j_1,j_2,j_3}| \max_{(u,x)\in\mathcal{K}_1\times\mathcal{K}_2} \|h_i(u,x)\|_\infty \, \|u_1 - u_2\|_\infty \\ &\leq L_{\Delta_1}\|u_1 - u_2\|_2.\end{aligned} \tag{A.III.25}$$

For Δ_2, we leverage P_i and obtain

$$\Delta_2 \leq \max_{u\in\mathcal{K}_1} \|A_i(u)\|_{\mathrm{op}} \, \|h_i(u_1,x_1) - h_i(u_2,x_2)\|_2 \leq \max_{u\in\mathcal{K}_1} \|A_i(u)\|_{\mathrm{op}} \, L_i \left(\|u_1 - u_2\|_2 + \|x_1 - x_2\|_2\right). \tag{A.III.26}$$

Combining Eqs. (A.III.23) to (A.III.26) shows P_n and concludes the induction, which in turn shows that $\widetilde{T}$ is jointly Lipschitz on any compact. Like in the smooth case, we conclude that T is globally Lipschitz, and thus that Assumption A.III.2 holds.

Verifying Assumption 4. Under both cases of regularity for the activations,

$$g := x \longmapsto \max_{u\in\overline{B}(0,R_u+\varepsilon)} \|\widetilde{T}(u,x)\|_2 \mathbb{1}^{\varepsilon}_{B(0,R_x)}(x)$$

is measurable and bounded. Furthermore, observe that for $u, x \in \mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$, $\|T(u,x)\|_2 \leq g(x)$. As a consequence, Assumption A.III.4 holds.

Verifying Assumption 5 in the $\mathcal{C}^2$ case. If all activations are $\mathcal{C}^2$-smooth, both $\widetilde{T}$ and its coordinate-wise products $T_i \times T_j$ are $\mathcal{C}^2$-smooth jointly in (u,x). As a result, one may bound these terms on $(u,x) \in \overline{B}(0,R_u+\varepsilon) \times \overline{B}(0,R_x+\varepsilon)$ by a constant M, independent of u, x, and the assumption is verified.

Verifying Assumption 5 in the semi-algebraic case. In the semi-algebraic case, there exists a structure $(\mathcal{U}_j)_{j\in J}$ of open sets of $\mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$ whose closures cover the entire space, such that $\widetilde{T}$ is polynomial in (u,x) on each $\mathcal{U}_j$, with J finite (this is possible since $\widetilde{T}$ is jointly semi-algebraic). The NN can be written $T(u,x) = \widetilde{T}(u,x)\mathbb{1}^{\varepsilon}_{B(0,R_u)}(u)\mathbb{1}^{\varepsilon}_{B(0,R_x)}(x)$, and is therefore $\mathcal{C}^2$-smooth on each $\mathcal{U}_j$. Furthermore, its restriction to $\mathcal{U}_j$ is extendable $\mathcal{C}^2$-smoothly to $\overline{\mathcal{U}_j}$ (we shall not introduce a different notation to these extensions, for legibility). As a result, one may introduce the following bounds on the derivatives of the coordinate functions on the intersection of the compact $\mathcal{K} := \overline{B}(0,R_u+\varepsilon) \times \overline{B}(0,R_x+\varepsilon)$ and $\overline{\mathcal{U}_j}$: there exists an $M_j > 0$ such that

$$\forall (u,x) \in \mathcal{K} \cap \overline{\mathcal{U}_j}, \left| \frac{\partial^2}{\partial u_{i_1} \partial u_{i_2}} \Big([T(u,x)]_{i_3} [T(u,x)]_{i_4} \Big) \right| \leq M_j \text{ and } \left\| \frac{\partial^2 T}{\partial u_{i_1} \partial u_{i_2}}(u,x) \right\|_2 \leq M_j.$$

Since J is finite and the $(\mathcal{U}_j)_{j\in J}$ cover $\mathcal{K}$, we deduce that this bound holds for $(u,x) \in \mathcal{K}$ for a common constant $M > 0$. Moreover, since T is the zero function outside of $\mathcal{K}$, this bounds also holds for any $(u,x) \in \mathbb{R}^{d_u} \times \mathbb{R}^{d_x}$. Finally, this shows that Assumption A.III.5 holds. □

A.III.7.5 Generalisation to Other Sliced Wasserstein Orders

In this section, we shall discuss how some of our results can be extended by replacing the 2-SW term SW_2^2 with SW_p^p for $p \in [1, +\infty)$.

Determining Lipschitz Constants. The first difficulty lies in showing that the functions $w_\theta^{(p)} := (X,Y) \longmapsto \mathrm{W}_p^p(P_\theta \# \gamma_X, P_\theta \# \gamma_Y)$ still have a locally Lipschitz regularity similar to Proposition A.III.1 (this proposition is only shown for $p=2$ in Chapter A.II). We generalise their result in the following proposition.

Proposition A.III.9. Let $K_w^{(p)}(r,X,Y) := pn(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}$, for $X,Y \in \mathbb{R}^{n\times d_y}$ and $r>0$. Then $w_\theta^{(p)}(\cdot,Y)$ is $K_w^{(p)}(r,X,Y)$-Lipschitz in the neighbourhood $B_{\|\cdot\|_{\infty,2}}(X,r)$:

$$\forall Y',Y'' \in B_{\|\cdot\|_{\infty,2}}(X,r),\ \forall \theta \in \mathbb{S}^{d_y-1},\ |w_\theta(Y',Y) - w_\theta(Y'',Y)| \leq K_w^{(p)}(r,X,Y)\|Y'-Y''\|_{\infty,2}.$$

Proof. Let $X,Y \in \mathbb{R}^{n\times d_y}, r>0$ and $Y',Y'' \in B_{\|\cdot\|_{\infty,2}}(X,r)$. By Lemma A.II.1, we have $|w_\theta^{(p)}(Y') - w_\theta^{(p)}(Y'')| \leq \|C'-C''\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm, and C' is a $n\times n$ matrix of entries $C'_{k,l} = |\theta^\top y'_k - \theta^\top y_l|^p$, with similarly $C''_{k,l} = |\theta^\top y''_k - \theta^\top y_l|^p$. Now consider the function

$$g_{y_l} := \left\{ \begin{array}{ccc} \mathbb{R}^{d_y} & \longrightarrow & \mathbb{R} \\ y & \longmapsto & |\theta^\top y - \theta^\top y_l|^p \end{array} \right. ,$$

which satisfies $C'_{k,l} = g_{y_l}(y'_k)$, and is differentiable almost-everywhere, with $\nabla g_{y_l}(y) = p|\theta^\top y - \theta^\top y_l|^{p-1}\theta$. For almost every $y \in B(x_k,r)$, we have

$$\begin{aligned} \|\nabla g_{y_l}(y)\|_2 &\leq p\|y-y_l\|_2^{p-1} = p\|y-x_k+x_k-y_l\|_2^{p-1} \\ &\leq p\left(\|y-x_k\|_2 + \|x_k\|_2 + \|y_l\|_2\right)^{p-1} \leq p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}. \end{aligned}$$

As a result, g_{y_l} is $p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}$-Lipschitz in $B(x_k,r)$. Now since $Y',y'' \in B_{\|\cdot\|_{\infty,2}}(X,r)$, we have $y'_k, y''_k \in B(x_k,r)$, thus

$$|[C']_{k,l} - [C'']_{k,l}| = |g_{y_l}(y'_k) - g_{y_l}(y''_k)| \leq p(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}\|y'_k - y''_k\|_2.$$

Then $\|C'-C''\|_F = \sqrt{\sum_{k,l}|[C']_{k,l}-[C'']_{k,l}|^2} \leq np(r+\|X\|_{\infty,2}+\|Y\|_{\infty,2})^{p-1}\|Y'-Y''\|_{\infty,2}$. □

Our results regarding the local Lipschitz property of f and F adapt immediately using the same method with the different constant $K_w^{(p)}(r,X,Y)$, we obtain the following constant for f (with L from Assumption A.III.2):

$$K_f^{(p)}(\varepsilon,u_0,X,Y) = pnL\left(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2}\right)^{p-1},$$

then the following constant for F:

$$K_F^{(p)}(\varepsilon,u_0) = pnL\int_{\mathcal{X}^n\times\mathcal{Y}^n}\left(\varepsilon L + \|T(u_0,X)\|_{\infty,2} + \|Y\|_{\infty,2}\right)^{p-1} \mathrm{d}\mu^{\otimes n}(X)\mathrm{d}\nu^{\otimes n}(Y).$$

In order to satisfy Condition A.III.2 item i) in the case $p \neq 2$, one needs to modify Assumption A.III.4 to require $\|T(u,x)\|_2 \leq g(x)^{1/(p-1)}(1+\|u\|_2)^{1/(p-1)}$, which in realistic cases is not much more expensive than asking for T to be bounded, which is a property of the class of NNs that we present in Section A.III.7.4.

Almost-Everywhere Gradient. A second difficulty lies in defining an almost-everywhere gradient f, since in our main text we rely on the formulation of an almost-everywhere gradient of $w_\theta^{(2)}(\cdot,Y)$ which was derived only for $p=2$ by [Bon+15a] in Chapter A.II. In fact, for θ,Y fixed $w_\theta^{(p)}(X,Y)$ is piecewise smooth, like $w_\theta^{(2)}(\cdot,Y)$ is piecewise quadratic. As a result, one may show that the following is an almost-everywhere gradient of $w_\theta^{(p)}(\cdot,Y)$:

$$\frac{\partial w_\theta^{(p)}}{\partial X}(X,Y) = \left(\frac{p}{n}\mathrm{sign}\left(\theta^\top x_k - \theta^\top y_{\sigma_\theta^{X,Y}(k)}\right)\left|\theta^\top x_k - \theta^\top y_{\sigma_\theta^{X,Y}(k)}\right|^{p-1}\theta\right)_{k\in[\![1,n]\!]} \in \mathbb{R}^{n\times d_y}.$$

The chain rule now yields the following almost-everywhere gradient for f:

$$\varphi(u,X,Y,\theta) = \sum_{k=1}^n \frac{p}{n}\mathrm{sign}\left(\theta^\top T(u,x_k) - \theta^\top y_{\sigma_\theta^{T(u,X),Y}(k)}\right)\left|\theta^\top T(u,x_k) - \theta^\top y_{\sigma_\theta^{T(u,X),Y}(k)}\right|^{p-1}\frac{\partial T}{\partial u}(u,x_k)\theta.$$

Adapting Proposition 4. Moving on to adapting Proposition A.III.4, the general case $p \neq 2$ makes things substantially more technical, but one may still show that the ψ functions are Lipschitz using restrictions on T its first and second-order derivatives (which can be formulated in a more technical version of Assumption A.III.5). In conclusion, Proposition A.III.4 can be adapted to apply to $p \in [1, +\infty)$, and it follows that Theorem A.III.1 also generalises to this case.

Path Differentiability. Regarding the results from Section A.III.4, the only substantial difference lies in showing that $T(\cdot, x)$ is path differentiable. The only missing link in the composition chain is the path differentiability of $\mathcal{E}^{(p)} := X \longmapsto \int_{\mathbb{S}^{d-1}} w_\theta^{(p)}(X, Y) \mathrm{d}\sigma(\theta)$. In the case $p = 2$, the difficulty of the integral can be circumvented by noticing that $\mathcal{E}$ is semi-concave (Proposition A.II.3), which implies path differentiability. This argument does not generalise to $p \in [1, +\infty)$ naturally, hence our Theorem A.III.2 only generalises to $p \in [1, +\infty)$ under the conjecture that $\mathcal{E}^{(p)}$ is indeed path differentiable.

A.IV

Differentiable Expectation-Maximisation and Applications to Gaussian Mixture Model Optimal Transport

A.IV.1 Introduction . . . 126
A.IV.2 Differentiation of the Expectation-Maximisation Algorithm . . . 127
A.IV.2.1 Main Ideas of the EM algorithm 127
A.IV.2.2 Fixed-Point Formulation and Differentiability 128
A.IV.2.3 Gradient Computation Methods 129
A.IV.3 Gaussian Mixture Model Optimal Transport . . . 131
A.IV.3.1 Reminders on GMM-OT 131
A.IV.3.2 Stability of MW_2^2 With Respect to GMM Parameters 132
A.IV.3.3 Minimisation of $\mathrm{EM} - \mathrm{MW}_2^2$: Local Optima and Weight Fixing 133
A.IV.3.4 Unbalanced GMM-OT 135
A.IV.4 Illustrations and Quantitative Study of Gradient Methods . . . 135
A.IV.4.1 Practical Implementation 135
A.IV.4.2 Flow of $\mathrm{EM} - \mathrm{MW}_2^2$ with Fixed Weights in 2D 136
A.IV.4.3 Flow of $\mathrm{EM} - \mathrm{MW}_2^2$ in 2D: Discussion on Uniform Weights 136
A.IV.4.4 Stochastic $\mathrm{EM} - \mathrm{MW}_2^2$ Flow with Fixed Weights 137
A.IV.4.5 Quantitative Study of EM Convergence and Gradients 137
A.IV.5 Applications of Differentiable EM . . . 138
A.IV.5.1 Barycentre Flow in 2D 138
A.IV.5.2 Colour Transfer 139
A.IV.5.3 Neural Style Transfer 140
A.IV.5.4 Image Generation 141
A.IV.5.5 Texture Synthesis 141
A.IV.6 Supplementary Material . . . 143
A.IV.6.1 Specific GMMs Used in Section A.IV.4.5 143
A.IV.6.2 Discussion on Gradient Ground Truths 144
A.IV.6.3 Explicit Differential Expressions 144
A.IV.6.4 Local Minima in (GMM)-OT 149
A.IV.6.5 Differentiating the Matrix Square Root 155
A.IV.6.6 Experimental Details and Additional Results 156

Abstract

The Expectation-Maximisation (EM) algorithm is a central tool in statistics and machine learning, widely used for latent-variable models such as Gaussian Mixture Models (GMMs). Despite its ubiquity, EM is typically treated as a non-differentiable black box, preventing its integration into modern learning pipelines where end-to-end gradient propagation is essential. In this work, we present and compare several differentiation strategies for EM, from full automatic differentiation to approximate

methods, assessing their accuracy and computational efficiency. As a key application, we leverage this differentiable EM in the computation of the Mixture Wasserstein distance MW_2 between GMMs, allowing MW_2 to be used as a differentiable loss in imaging and machine learning tasks. To complement our practical use of MW_2, we contribute a novel stability result which provides theoretical justification for the use of MW_2 with EM, and also introduce a novel unbalanced variant of MW_2. Numerical experiments on barycentre computation, colour and style transfer, image generation, and texture synthesis illustrate the versatility and effectiveness of the proposed approach in different settings. This chapter is based on the paper:

[Boï+25] Samuel Boïté*, Eloi Tanguy*, Julie Delon, Agnès Desolneux and Rémi Flamary.
"Differentiable Expectation-Maximisation
and Applications to Gaussian Mixture Model Optimal Transport".
arxiv preprint 2509.02109 (Sept. 2025). (*: equal contribution)

A.IV.1 Introduction

The Expectation-Maximisation (EM) algorithm [DLR77] is a ubiquitous tool in statistics to fit mixture models on data [BE94; End03; VH10; Ng13]. Numerous variants of the EM algorithm were proposed in the statistics and machine learning communities [DH97; Fri98; FH02; CJJ05; GTG07; VR08; CM09; Cap11; SCR12; GVS17; ZAC21; Kim22], and are the focus of various monographs [MK07; MP00]. From a theoretical standpoint, the EM algorithm is only known to converge under specific conditions [Wu83; Boy83; MK07; XHM16], and its behaviour is still not completely understood in full generality. In machine learning, Gaussian priors have been used for latent space representations [Ras03; KW14; RMW14; HJA20] with resounding success, while Gaussian Mixture Models (GMMs) are using more sparingly [NB06; VM19; Yua+20]. Beyond the difficulties of GMM estimation, the core challenge is that the EM algorithm is not easily integrated into end-to-end learning pipelines, as its differentiation with respect to the input data is not straightforward. In this chapter, we tackle the theory and practice of differentiating the EM algorithm, in the hope of sparking further research in this direction, beyond the various applications that we present.

One of the core contributions that sparked the Machine Learning wave is automatic differentiation [Wen64; Lin70; RHW86; GW08], which is a powerful method for complex optimisation problems. In the setting where the target objective is itself an optimisation problem, the problem is referred to as "bi-level". Numerous automatic differentiation methods for bi-level iterative minimisation were studied in [Gil92; Bec94; Sha+19; MO20; BPV22; BPV24b]. Another approach to bi-level optimisation involving fixed-point problems is the *implicit method* [Lui+18; BKK19; LVD20; Bol+21; Blo+22; ER24].

In order to illustrate the potential of differentiable EM, we apply it to imaging tasks which rely on the comparison of GMMs with Optimal Transport (OT) [Mon81; Kan42]. Specifically, we leverage a variant of the Wasserstein distance between GMMs, called the Mixture-Wasserstein distance MW_2 [DD20], which compares GMMs by matching their components using a small-scale discrete OT problem that can be solved efficiently [PC19b]. The Mixture-Wasserstein distance is one of many examples of recent advances in computational OT, where less costly surrogates of the Wasserstein distance such as regularised transport [Cut13] or sliced transport [Bon13] saw a wide range of success in machine learning [Bon+11; Cou+17; Kol+19b; KLA19; Fey+19]. Computing transport distances between empirical distributions remains challenging when the number of samples or the dimensionality of the space becomes too large, even though several solutions have been proposed in the literature [WB19; Gen+19; Chi+20].

The Mixture-Wasserstein distance has been used in texture synthesis [LDD23], for the evaluation of generative neural networks [Luz+23], in quantum chemistry [Dal+23], and for domain adaptation [MMS24b]. An efficient barycentre computation algorithm for this metric was also recently proposed in [TDG24]. When using MW_2 on discrete data, the space dimension d and

the number of samples n only appear in two stages of the whole computation: the GMM inference on the data, and the computation of Bures distances between the covariance matrices of the GMM components. This makes the approach highly versatile and robust to dimensionality in practice. Nevertheless, a current limitation of the MW_2 distance is the inference of the GMMs, which our work renders differentiable with EM, thus allowing the use of MW_2 as a differentiable loss function between datasets in machine learning tasks.

Objectives. Our goal in this chapter is to propose different ways to differentiate EM, allowing for applications in imaging or machine learning problems. As a focal application, we will pair differentiable EM with the Mixed-Wasserstein distance MW_2, which is not difficult to differentiate in practice (using classical results on the differentiation of discrete OT [PC19b], see also Proposition B.I.8). Differentiation of the Expectation-Maximisation algorithm is a more involved process. Surprisingly, while EM is well-known and extensively studied, the question of its differentiation seldom appears in the literature. To the best of our knowledge, the first work in this direction is [Kim22], which re-writes a Bayesian variant of EM which is related to the Optimal Transport Kernel Embedding [Mia+21]. In this chapter, we propose several approaches for this differentiation, exact with auto-differentiation[1] or approximated, and compare their performances on different applications, ranging from toy examples to larger-scale machine learning tasks.

Paper outline. In Section A.IV.2, we begin by recalling the EM algorithm and expressing it as a fixed point problem. We give precise mathematical meaning to the differentiation of a solution of the EM algorithm, and present numerous strategies to compute the differential of T steps of the method with respect to the input data. In Section A.IV.3, we provide a short reminder on the Mixture-Wasserstein distance MW_2 and show a stability result for the estimation of MW_2 between GMMs. We discuss practical difficulties in the differentiation of the MW_2 distance between GMMs estimated from data, and provide rationale for the importance of fixing EM weights. To circumvent the numerical difficulties incurred by weight optimisation and to ensure robustness to Gaussian outliers, we introduce an unbalanced variant of MW_2. In Section A.IV.4 we illustrate our methods on the flow of MW_2 composed with the EM algorithm, and perform a quantitative study on the convergence of the EM algorithm and the quality of the gradient approximations. In Section A.IV.5, we present several applications of differentiable EM: barycentre computation, colour transfer with inspiration from [Rab+12], style transfer in the spirit of [GEB15], image generation through MW_2-based Generative Adversarial Networks, and a novel texture synthesis method related to [GLR18; LDD23; Hou+23].

A.IV.2 Differentiation of the Expectation-Maximisation Algorithm

A.IV.2.1 Main Ideas of the EM algorithm

The Expectation-Maximisation (EM) algorithm [DLR77] attempts to fit a GMM to a dataset $X \in \mathbb{R}^{n\times d}$, with a fixed number of components K. We introduce the (hidden) quantities $Y \in [\![1, K]\!]^n$ which encode the component index of each sample x_i. The GMM parameters $\theta := (w, (m_k)_k, (\Sigma_k)_k)$, lie in the space $\Theta := \triangle_K \times (\mathbb{R}^d)^K \times (S_d^{++}(\mathbb{R}))^K$, where $\triangle_K$ is the K-simplex defined as

$$\triangle_K := \left\{ w \in (0,1)^K \mid \sum_{k=1}^{K} w_k = 1 \right\}, \tag{A.IV.1}$$

and $S_d^{++}(\mathbb{R})$ is the set of symmetric positive definite matrices. We will denote by $\mu(\theta)$ the GMM probability measure of parameters θ. Given $X \in \mathbb{R}^{n\times d}$ and $Y \in [\![1, K]\!]^n$, the complete likelihood

[1] barring numerical approximation, which is certain cases may cause substantial errors.

and its logarithm are respectively

$$L_\theta(X,Y) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(w_k g_{m_k,\Sigma_k}(x_i)\right)^{\mathbb{1}(Y_i=k)}, \quad \ell_\theta(X,Y) = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(Y_i=k)\log\left(w_k g_{m_k,\Sigma_k}(x_i)\right), \tag{A.IV.2}$$

where g_{m_k,Σ_k} is the Gaussian density with mean m_k and covariance Σ_k (recalled in Eq. (A.IV.23)). Note that ℓ_θ cannot be optimised in θ directly since we do not know the hidden variables Y. The EM algorithm [DLR77] maximises the log-likelihood by iterating two steps over θ_t, first computing the "responsibilities" $\gamma_{ik}(\theta_t)$ which are the posterior probabilities of the hidden quantities Y_i given the data X and the current parameters $\theta_t = (w^{(t)}, (m_k^{(t)})_k, (\Sigma_k^{(t)})_k) \in \Theta$:

$$\gamma_{ik}(\theta_t) = \mathbb{P}_{(\mathbf{X},\mathbf{Y})\sim\mu(\theta_t)^{\otimes n}}\left[\mathbf{Y}_i = k | \mathbf{X} = X\right] = \left[w_k^{(t)} g_{m_k^{(t)},\Sigma_k^{(t)}}(x_i)\right] \Big/ \left[\sum_{\ell=1}^{K} w_\ell^{(t)} g_{m_\ell^{(t)},\Sigma_\ell^{(t)}}(x_i)\right]. \tag{A.IV.3}$$

The next iteration θ_{t+1} corresponds to a maximisation of $\ell_\theta(X,Y)$ with the unknown quantities z replaced by the posterior probabilities $\gamma_{ik}(\theta_t)$, which leads to the following closed-form expressions for the parameters:

$$w_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ik}(\theta_t), \quad m_k^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)x_i}{\sum_{j=1}^{n}\gamma_{jk}(\theta_t)}, \quad \Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{ik}(\theta_t)(x_i - m_k^{(t+1)})(x_i - m_k^{(t+1)})^\top}{\sum_{j=1}^{n}\gamma_{jk}(\theta_t)}. \tag{A.IV.4}$$

It is a standard result (see [Moo96]) that the log-likelihood $\ell_\theta(X) = \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K} w_k g_{m_k,\Sigma_k}(x_i)\right)$ increases with respect to $\theta = (w, (m_k), (\Sigma_k))$ with each EM iteration. In practice, we shall see that it is sometimes preferable to tweak the standard EM algorithm by not updating the weights w and keeping the weight w_0 of the initialisation θ_0 (we refer to this as fixing the weights). We summarise the two algorithms in Algorithms A.IV.1 and A.IV.2, with the difference highlighted in red.

Alg. A.IV.1: EM Algorithm

Input: $\theta_0 \in \Theta$, $X \in \mathbb{R}^{n\times d}$, $T, K \in \mathbb{N}^*$.

for $t \in [\![0, T-1]\!]$ **do**

- **Expectation:** Compute the responsibilities $\gamma(\theta_t)$ using Eq. (A.IV.3);
- **Maximisation:** Update $\theta_{t+1} = (w^{(t+1)}, (m_k^{(t+1)})_k, (\Sigma_k^{(t+1)})_k)$ using Eq. (A.IV.4);

Alg. A.IV.2: Fixed-Weights EM

Input: $\theta_0 \in \Theta$, $X \in \mathbb{R}^{n\times d}$, $T, K \in \mathbb{N}^*$.

for $t \in [\![0, T-1]\!]$ **do**

- **Expectation:** Compute the responsibilities $\gamma(\theta_t)$ using Eq. (A.IV.3);
- **Maximisation:** Update $\theta_{t+1} = (w_0, (m_k^{(t+1)})_k, (\Sigma_k^{(t+1)})_k)$ using Eq. (A.IV.4);

For applications minimising a loss involving the output of the EM algorithm with respect to the data X, it is important to highlight that even in the fixed-weights version (Algorithm A.IV.2), the responsibilities γ evolve with the EM steps and with the updates of X. As a result, running EM algorithm along with X updates is paramount, and it is not sufficient to keep an initial responsibility assignment γ.

A.IV.2.2 Fixed-Point Formulation and Differentiability

The goal of this section is to express an EM step as a fixed-point operation. For this, a technical condition is required to ensure that the term $\Sigma_k^{(t+1)}$ is invertible (symmetry and non-negativity of the eigenvalues is immediate), which requires a slightly stronger condition than assuming that the points (x_i) span $\mathbb{R}^d$. Since the terms $\gamma_{i,k}$ are positive, we can write the condition as:

$$X = (x_1, \cdots, x_n) \in \mathcal{X} := \left\{(x_1, \cdots, x_n) \in (\mathbb{R}^d)^n \mid \forall y \in \mathbb{R}^d,\ \mathrm{Span}((x_i - y)_{i\in[\![1,n]\!]}) = \mathbb{R}^d\right\}. \tag{A.IV.5}$$

For $X \in \mathcal{X}$ and $\theta_0 \in \Theta$, the next iteration θ_1 obtained using Eq. (A.IV.4) verifies $\theta_1 \in \Theta$. Note that if ν is an absolutely continuous probability measure on $\mathbb{R}^d$, and if $n \geq d+1$, then Eq. (A.IV.5) is verified almost-surely for $\mathbf{X} \sim \nu^{\otimes n}$. This condition can be seen as a weaker variant of being in a "standard configuration". It can be shown that $\mathcal{X}$ is an open set of $(\mathbb{R}^d)^n \simeq \mathbb{R}^{n \times d}$.

We study the map $F : \Theta \times \mathcal{X} \longrightarrow \Theta$ that maps a parameter θ to the value of the next M-step using Eq. (A.IV.4). For convenience, we also write $F_X := F(\cdot, X)$ and F_X^t the t-th iteration $F_X \circ \cdots \circ F_X$ for $t \in \mathbb{N}$. An optimal solution to the EM algorithm can be seen as a fixed point of F, i.e. $\theta^* = F_X(\theta^*)$. Numerically, one takes a final iteration θ_T of the iteration scheme

$$\forall t \in [\![0, T-1]\!],\ \theta_{t+1} = F(\theta_t, X),$$

with an arbitrary initialisation $\theta_0 \in \Theta$. Due to the definition of Θ as the set of parameters with weights $w_k \in (0,1)$ and *positive* definite matrices Σ_k, the map F is of class $\mathcal{C}^\infty$ (jointly in (θ, X)), by virtue of the explicit expressions Eq. (A.IV.4).

We now aim to give meaning to the gradient with respect to the data of a "true" solution of the EM algorithm. To this end, we need to work under the assumption of convergence of EM to a non-degenerate fixed point:

Assumption A.IV.1. There exists $(\theta_0, X_0) \in \Theta \times \mathcal{X}$ such that $\theta^*(\theta_0, X_0) := \lim_{t \longrightarrow +\infty} F_{X_0}^t(\theta_0)$ exists in Θ, with $\frac{\partial F}{\partial \theta}(\theta^*(\theta_0, X_0), X_0) - I$ invertible.

While convergence is typically observed in practice numerically, from a theoretical standpoint, it is a delicate matter, see Chapter 3 [MK07] for a reference on this field of research. We are now ready to formulate Proposition A.IV.1, which shows that the gradient of an EM solution with respect to the data is well-defined, using the implicit function theorem.

Proposition A.IV.1. Under Assumption A.IV.1, there exists open vicinities $\Theta^*, \mathcal{X}_0$ such that $\theta^*(\theta_0, X_0) \in \Theta^* \subset \Theta$ and $X_0 \in \mathcal{X}_0 \subset \mathcal{X}$, where there exists $\theta^*(\theta_0, \cdot) \in \mathcal{C}^\infty(\mathcal{X}_0, \Theta^*)$ with:

$$\forall (\theta, X) \in \Theta^* \times \mathcal{X}_0,\ F(\theta, X) = \theta \Longleftrightarrow \theta = \theta^*(\theta_0, X). \tag{A.IV.6}$$

Proof. Let $G := \left\{ \begin{array}{ccc} \Theta \times \mathcal{X} & \longrightarrow & \Theta \\ (\theta, X) & \longmapsto & F(\theta, X) - \theta \end{array} \right.$. Thanks to the regularity of F, G is of class $\mathcal{C}^\infty$. Assumption A.IV.1 implies that $G(\theta^*(X_0), X_0) = 0$, with $\frac{\partial G}{\partial \theta}(\theta^*(X_0), X_0)$ invertible.

By the implicit function theorem, there exists open vicinities $\Theta^*, \mathcal{X}_0$ such that $\theta^*(\theta_0, X_0) \in \Theta^* \subset \Theta$ and $X_0 \in \mathcal{X}_0 \subset \mathcal{X}$, and there exists $g : \mathcal{X}_0 \longrightarrow \Theta^*$ of class $\mathcal{C}^\infty$ such that $g(X_0) = \theta^*(\theta_0, X_0)$, and:

$$\forall (\theta, X) \in \Theta^* \times \mathcal{X}_0,\ F(\theta, X) = \theta \Longleftrightarrow \theta = g(X).$$

For $X \in \mathcal{X}_0$, we define $\theta^*(\theta_0, X) := g(X)$. □

To alleviate notation, we will write simply $\theta^*(X) := \theta^*(\theta_0, X)$ and continue with Assumption A.IV.1 and the map θ^* from Proposition A.IV.1. For the sake of completion and to pave the way for future theoretical study, we provide the explicit expressions of the partial differentials of F in Section A.IV.6.3. Note that from a statistical viewpoint, the parameter θ^* is not a Maximum-Likelihood Estimator (which, in fact, does not exist for GMMs), it is only the output of the EM algorithm which is often a local maximum of the likelihood.

A.IV.2.3 Gradient Computation Methods

In this section, we are concerned with practical implementation of the computation of the gradient of the EM algorithm with respect to the data $\partial_X[F_X^T(\theta_0)]$, given some initialisation $\theta_0 \in \Theta$ and number of iterations $T \geq 1$. In addition to automatic-differentiation method, we present two alternative approximate strategies which rely only on the last parameter θ_T. These methods work under the assumption that $\theta_T(X) \approx \theta^*(X)$, where θ^* is defined in Proposition A.IV.1 and refers to the fixed point $\lim_{t \longrightarrow +\infty} F_X^t(\theta_0)$, assuming convergence.

Full Automatic Differentiation (AD). The most naïve approach consists in computing the gradient through all iterations using the backpropagation algorithm (for instance, using PyTorch's automatic differentiation [Pas+19]). In other words, the "full automatic gradient" corresponds to letting automatic differentiation compute $\partial_X[F_X^T(\theta_0)]$ directly, using a automatically differentiable implementation of the EM algorithm. For a large number of iterations T, AD can be considered as a natural baseline for computing the exact gradient, up to numerical precision. It is still an approximation, due to propagation of numerical errors, but is to the best of our knowledge the best available approximation of the true gradient, as discussed in Section A.IV.6.2. Nevertheless, we use AD as a natural baseline method for comparisons. The AD method may be costly (both in time and memory) if the number of iterations T is very large.

Approximate Implicit Gradient (AI). Our goal is to approximate the gradient $\frac{\partial \theta^*}{\partial X}(X)$ at a fixed point θ^* of $F(\cdot, X)$. Thanks to the differentiability property of Proposition A.IV.1 and under Assumption A.IV.1, we can differentiate with respect to X using the chain rule:

$$\frac{\partial \theta^*}{\partial X}(X) = \frac{\partial}{\partial X}[F(\theta^*, X)] = \frac{\partial F}{\partial \theta}(\theta^*, X)\frac{\partial \theta^*}{\partial X}(X) + \frac{\partial F}{\partial X}(\theta^*, X). \tag{A.IV.7}$$

We deduce the following equation on $\dfrac{\partial \theta^*}{\partial X}(X)$:

$$\left(I - \frac{\partial F}{\partial \theta}(\theta^*, X)\right)\frac{\partial \theta^*}{\partial X}(X) = \frac{\partial F}{\partial X}(\theta^*, X), \tag{A.IV.8}$$

using Assumption A.IV.1 again, we can invert the matrix on the left hand-side, yielding

$$\frac{\partial \theta^*}{\partial X}(X) = \left(I - \frac{\partial F}{\partial \theta}(\theta^*, X)\right)^{-1}\frac{\partial F}{\partial X}(\theta^*, X). \tag{A.IV.9}$$

We define the approximate implicit gradient by approximating $\theta^* \approx \theta_T$:

$$J_{\mathrm{AI}} := \left(I - \frac{\partial F}{\partial \theta}(\theta_T, X)\right)^{-1}\frac{\partial F}{\partial X}(\theta_T, X) \tag{A.IV.10}$$

The implicit approximation is theoretically exact (barring numerical imprecision in the inversion in particular) when $\theta_T = \theta^*$, however it requires additional costly computations, first of the differential $\partial_\theta F(\theta_T, X)$, and then solving a large linear system to compute $(I - \partial_\theta F(\theta^*, X))^{-1}\partial_X F(\theta^*, X)$.

One-Step Gradient Approximation (OS). The One-Step method (OS) studied in [BPV24b] (within a particular framework of bi-level optimisation), works under the following condition:

> **Condition A.IV.1.** $\left\|\frac{\partial F}{\partial \theta}(\theta, X)\right\|_{\mathrm{op}} \leq \rho \ll 1$ for any θ, X.

In the case of the EM algorithm, Condition A.IV.1 is not verified, since the eigenvalues of the partial differential $\frac{\partial F}{\partial \theta}(\theta, X)$ are commonly larger than 1, even in the vicinity of a fixed point. Under Condition A.IV.1, the OS approximation further neglects the term $(I - \partial_\theta F(\theta^*, X))^{-1}$ in Eq. (A.IV.10), yielding the following expression:

$$J_{\mathrm{OS}} := \frac{\partial F}{\partial X}(\theta_{T-1}, X) \approx \frac{\partial F}{\partial X}(\theta^*, X) \approx \left(I - \frac{\partial F}{\partial \theta}(\theta^*, X)\right)^{-1}\frac{\partial F}{\partial X}(\theta^*, X) = \frac{\partial \theta^*}{\partial X}(X). \tag{A.IV.11}$$

In practice, the OS method corresponds to computing the gradient of the EM output $\theta_T(X) = F_X^T(\theta_0)$ with respect to the data X only through the last iteration $\theta_T(X) = F_X(\theta_{T-1})$, neglecting the dependence of the penultimate iteration θ_{T-1} on X. Using automatic differentiation (for

example with PyTorch [Pas+19]), this is done conveniently by computing $\theta_{T-1} = F_X^{T-1}(\theta_0)$ without gradient computation (e.g. `with torch.no_grad()`), and performing the last step with gradient computation. Due to this computation method, the OS gradient is numerically inexpensive compared to the others, albeit at a the expense of precision. See [BPV24b] for a detailed presentation.

Method	Time	Memory
Full Automatic Differentiation (AD)	$\mathcal{O}(T(nKd^2 + Kd^3))$	$\mathcal{O}(TKd^2 + nd)$
Approximate Implicit gradient (AI)	$\mathcal{O}(TnKd^2 + K^3d^6 + nK^2d^5)$	$\mathcal{O}(nK^2d^4)$
One-Step gradient (OS)	$\mathcal{O}(T(nKd^2 + Kd^3))$	$\mathcal{O}(Kd^2 + nd)$

Table A.IV.1: Complexities of the `backward` passes (gradient computations) for T EM iterations on n points in $\mathbb{R}^d$ with K components.

Time complexity and memory footprint. The complexities of the gradient computation approaches are summarised in Table A.IV.1. As a baseline, the time complexity of the `forward` pass (EM algorithm without gradients) is $\mathcal{O}(T(nKd^2 + Kd^3))$, while its memory footprint is $\mathcal{O}(Kd^2+nd)$. The complexities are deduced from the differential expressions in Section A.IV.6.3. The $\mathcal{O}(Kd^3)$ factor corresponds to inverting the K covariance matrices of size $d \times d$ during each E-step. The $\mathcal{O}(nKd^2)$ factor comes from the M-step parameter updates (weights, means, covariances) and the differentiation of these updates with respect to X or θ.

The Warm-Start Method for Iteration of Differentiable EM. In many practical applications, we are interested in minimising a certain loss function $\mathcal{L}$ applied to the output $\theta_T(X)$ of the EM algorithm[2]. In this case, after a (small) gradient descent step X_{t+1} computed from X_t, the output of the EM algorithm with data X_t will often be a good initialisation for the EM algorithm with data X_{t+1}. As a result, we suggest operating the EM algorithm with only one iteration and using the output of the previous step as an initialisation. This leads to the following algorithm, which we refer to as the Warm-Start EM Flow of a loss $\mathcal{L} : \Theta \longrightarrow \mathbb{R}$.

Algorithm A.IV.3: Warm-Start EM Flow

Input: $\theta_0 \in \Theta$, $X_0 \in \mathcal{X}$, $T_{\mathrm{GD}} \in \mathbb{N}$, $\mathcal{L} : \Theta \longrightarrow \mathbb{R}$.

1 **for** $t \in [\![0, T_{\mathrm{GD}} - 1]\!]$ **do**

2 $\quad \theta_{t+1} = F(\theta_t, X_t)$;

3 $\quad X_{t+1} = X_t - \eta_t \frac{\partial \mathcal{L}}{\partial \theta}(\theta_t) \frac{\partial F}{\partial X}(\theta_t, X_t)$;

The gradient step at line 3 means that we perform automatic differentiation of the expression $\mathcal{L}(F(\theta_t, X_t))$ with respect to X at X_t, seeing θ_t as a constant and storing the value $\theta_{t+1} = F(\theta_t, X_t)$ for the next iteration. In essence, the Warm-Start method corresponds to an online OS gradient, which does not suffer from the approximation error of the OS method (since only one step is performed), yet benefits from the low memory footprint of the OS method.

A.IV.3 Gaussian Mixture Model Optimal Transport

A.IV.3.1 Reminders on GMM-OT

This section summarises the main results from [DD20]. The quadratic Wasserstein distance between two probability measures μ_0 and μ_1 on $\mathbb{R}^d$ with finite second moments is defined by

$$\mathrm{W}_2^2(\mu_0, \mu_1) := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y_0 - y_1\|_2^2 \mathrm{d}\pi(y_0, y_1), \tag{A.IV.12}$$

[2] for example $\mathcal{L}(\theta_T(X)) = \mathrm{MW}_2^2(\mu(F(\theta_T, X)), \nu)$, where ν is a target GMM, see Section A.IV.3.

where $\Pi(\mu_0, \mu_1)$ denotes the set of probability measures with finite second moments on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ_0 and μ_1. A solution π^* to Eq. (A.IV.12) is called an optimal transport plan between μ_0 and μ_1. This distance has been widely used over the past fifteen years for various applications in data science. Let GMM_d denote the set of probability measures that can be written as finite Gaussian Mixture Models (GMMs) on $\mathbb{R}^d$. Transport plans and barycentres between GMMs with respect to W_2 are generally not themselves GMMs, which is a limitation when such representations are used for data analysis or generation. For this reason, the authors of [DD20] propose to modify the W_2 formulation by restricting the couplings to be GMMs on $\mathbb{R}^d \times \mathbb{R}^d$. More precisely, given $\mu_0, \mu_1 \in \mathrm{GMM}_d$, one can define

$$\mathrm{MW}_2^2(\mu_0, \mu_1) := \inf_{\pi \in \Pi^{\mathrm{GMM}}(\mu_0, \mu_1)} \int_{\mathbb{R}^{2d}} \|y_0 - y_1\|_2^2 \mathrm{d}\pi(y_0, y_1), \tag{A.IV.13}$$

where $\Pi^{\mathrm{GMM}}(\mu_0, \mu_1)$ denotes the set of probability measures in GMM_{2d} with marginals μ_0 and μ_1. The problem is well-defined since this set contains the product measure $\mu_0 \otimes \mu_1$. The authors show that MW_2 defines a distance between elements of GMM_d. Moreover, if $\mu_0 = \sum_{k=1}^{K_0} w_k^{(0)} \mu_k^{(0)}$ and $\mu_1 = \sum_{\ell=1}^{K_1} w_\ell^{(1)} \mu_\ell^{(1)}$, where $(w_k^{(0)})_k \in \triangle_{K_0}$ and $(w_\ell^{(1)})_\ell \in \triangle_{K_1}$, and $\mu_k^{(0)}, \mu_\ell^{(1)}$ are Gaussian measures, then it can be shown ([DD20, Proposition 4]) that

$$\mathrm{MW}_2^2(\mu_0, \mu_1) = \min_{P \in \Pi(w_0, w_1)} \sum_{k,\ell} P_{kl} \mathrm{W}_2^2(\mu_k^{(0)}, \mu_\ell^{(1)}), \tag{A.IV.14}$$

where $\Pi(w_0, w_1)$ is the set of $K_0 \times K_1$ matrices with non-negative entries and marginals w_0 and w_1:

$$\Pi(w_0, w_1) = \left\{ P \in \mathcal{M}_{K_0, K_1}(\mathbb{R}^+);\ \forall k, \sum_j P_{kj} = w_k^{(0)} \text{ and } \forall j,\ \sum_k P_{kj} = w_j^{(1)} \right\}.$$

This discrete formulation makes MW_2 very easy to compute in practice, even in high dimensions. Indeed, the W_2 distance between two Gaussian measures $\mu = \mathcal{N}(m, \Sigma)$ and $\tilde{\mu} = \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ admits a closed-form expression:

$$\mathrm{W}_2^2(\mu, \tilde{\mu}) = \|m - \tilde{m}\|_2^2 + \mathrm{tr}\left(\Sigma + \tilde{\Sigma} - 2\left(\Sigma^{\frac{1}{2}} \tilde{\Sigma} \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right), \tag{A.IV.15}$$

where $M^{\frac{1}{2}}$ denotes the unique positive semi-definite square root of the positive semi-definite matrix M. If the parameters of the GMMs μ_0 and μ_1 are known, computing Eq. (A.IV.14) amounts to evaluating $K_0 \times K_1$ Wasserstein distances between Gaussians and solving a discrete transport problem of size $K_0 \times K_1$. It is also possible to define barycentres for MW_2 [DD20; TDG24], which leads to a similar discrete formulation. Given point clouds, [DD20] suggest to use EM to fit GMMs to the data, allowing comparison of the point clouds with EM-MW_2^2.

A.IV.3.2 Stability of MW_2^2 With Respect to GMM Parameters

To study the stability of the MW_2 distance with respect to the GMM parameters, we leverage a discrete OT stability result from Chapter A.II. To relate this problem to discrete OT stability, we see $\mathrm{MW}_2^2(\mu_0, \mu_1)$ as a particular discrete Kantorovich problem with cost matrix $C_{k_0,k_1} := \|m_{k_0}^{(0)} - m_{k_1}^{(1)}\|_2^2 + d_{\mathrm{BW}}^2(\Sigma_{k_0}^{(0)}, \Sigma_{k_1}^{(1)})$, where we recall the expression of the Bures-Wasserstein distance on $S_d^+(\mathbb{R})$:

$$\forall \Sigma, \Sigma' \in S_d^+(\mathbb{R}),\ d_{\mathrm{BW}}(\Sigma, \Sigma') := \sqrt{\mathrm{Tr}(\Sigma + \Sigma' - 2(\Sigma^{1/2}\Sigma'\Sigma^{1/2})^{1/2})}. \tag{A.IV.16}$$

The main idea of Proposition A.IV.2 is to say that if the GMM parameters are sufficiently close (thanks to EM convergence for instance), then the MW_2^2 costs will also be close thanks to the stability result from Lemma A.II.1. While general sample complexity results for the EM algorithm are not available, assuming a certain rate of convergence towards true parameters, this result shows that the precision obtained using EM translates into a precision on the MW_2^2 distance. This key observation is a first step towards guaranteeing the quality of MW_2^2 as a loss function with respect to the data.

Proposition A.IV.2. For $i \in \{0,1\}$, consider GMM parameters $(\hat{w}_i, \hat{m}_i, \hat{\Sigma}_i) \in \triangle_{K_i} \times \mathbb{R}^{K_i \times d} \times S_d^{++}(\mathbb{R})^{K_i}$ of a GMM $\hat{\mu}_i$ as estimators of $(w_i, m_i, \Sigma_i) \in \triangle_{K_i} \times \mathbb{R}^{K_i \times d} \times S_d^{++}(\mathbb{R})^{K_i}$ which are parameters of a target GMM μ_i. Assume that the means and covariances are bounded, namely that there exists $R_m > 0$, $R_\Sigma > 0$ such that:

$$\forall i \in \{0,1\},\ \forall k \in [\![1, K_i]\!],\ \|m_k^{(i)}\|_2 \leq R_m,\ \|\hat{m}_k^{(i)}\|_2 \leq R_m,\ \sqrt{\mathrm{Tr}\Sigma_k^{(i)}} \leq R_\Sigma,\ \sqrt{\mathrm{Tr}\hat{\Sigma}_k^{(i)}} \leq R_\Sigma.$$

Further assume "convergence rates" on the parameter estimations for $i \in \{0,1\}$ and $k \in [\![1, K_i]\!]$:

$$\forall i \in \{0,1\},\ \mathbb{E}\left[\|w_i - \hat{w}_i\|_1\right] \leq \rho_w,\ \mathbb{E}\left[\|m_k^{(i)} - \hat{m}_k^{(i)}\|_2\right] \leq \rho_m,\ \mathbb{E}\left[d_{\mathrm{BW}}\left(\Sigma_k^{(i)}, \hat{\Sigma}_k^{(i)}\right)\right] \leq \rho_\Sigma,$$

then the following stability bound holds:

$$\mathbb{E}\left[\left|\mathrm{MW}_2^2(\hat{\mu}_0, \hat{\mu}_1) - \mathrm{MW}_2^2(\mu_0, \mu_1)\right|\right] \leq 8R_m\rho_m + 8R_\Sigma\rho_\Sigma + 8(R_m^2 + R_\Sigma^2)\rho_w. \tag{A.IV.17}$$

Proof. Consider the cost matrices $C, \hat{C} \in \mathbb{R}_+^{K_0 \times K_1}$ defined by their entries at $(k_0, k_1) \in [\![1, K_0]\!] \times [\![1, K_1]\!]$:

$$C_{k_0,k_1} := \|m_{k_0}^{(0)} - m_{k_1}^{(1)}\|_2^2 + d_{\mathrm{BW}}^2(\Sigma_{k_0}^{(0)}, \Sigma_{k_1}^{(1)}),\ \hat{C}_{k_0,k_1} := \|\hat{m}_{k_0}^{(0)} - \hat{m}_{k_1}^{(1)}\|_2^2 + d_{\mathrm{BW}}^2(\hat{\Sigma}_{k_0}^{(0)}, \hat{\Sigma}_{k_1}^{(1)}),$$

and apply Lemma A.II.1 with weights $(w_i, \overline{w_i}) := (w_i, \hat{w}_i)$ for $i \in \{0,1\}$, and cost matrices $(M, \overline{M}) := (C, \hat{C})$: we obtain:

$$\left|\mathrm{MW}_2^2(\hat{\mu}_1, \hat{\mu}_2) - \mathrm{MW}_2^2(\mu_1, \mu_2)\right| \leq \|C - \hat{C}\|_\infty + \|C\|_\infty \left(\|w_0 - \hat{w}_0\|_1 + \|w_1 - \hat{w}_1\|_1\right).$$

First, noticing that $\forall \Sigma \in S_d^+(\mathbb{R}),\ d_{\mathrm{BW}}(0, \Sigma) = \sqrt{\mathrm{Tr}(\Sigma)}$, we apply the triangle inequality on $\|\cdot\|_2$ and d_{BW} to obtain $\|C\|_\infty \leq 4(R_m^2 + R_\Sigma^2)$. We now turn to upper-bounding $\|C - \hat{C}\|_\infty$. We will use the inequality $\forall (s,t) \in \mathbb{R}^2,\ |s^2 - t^2| \leq 2\max(|s|, |t|)|s - t|$. First, for $x, \hat{x}, y, \hat{y} \in B_{\mathbb{R}^d}(0, R_m)$, we have by the triangle inequality:

$$\left|\|x - y\|_2^2 + \|\hat{x} - \hat{y}\|_2^2\right| \leq 2\max(\|x-y\|_2, \|\hat{x}-\hat{y}\|_2)\|(x-y)-(\hat{x}-\hat{y})\|_2 \leq 4R_m\left(\|x - \hat{x}\|_2 + \|y - \hat{y}\|_2\right).$$

Similarly, for $A, \hat{A}, B, \hat{B} \in \{S \in S_d^+(\mathbb{R}) : \sqrt{\mathrm{Tr}(S)} \leq R_\sigma\}$, we compute:

$$\left|d_{\mathrm{BW}}(A, B) - d_{\mathrm{BW}}(\hat{A}, \hat{B})\right| \leq 4R_\Sigma\left(d_{\mathrm{BW}}(A, \hat{A}) + d_{\mathrm{BW}}(B, \hat{B})\right).$$

Applying our two inequalities to the entries of $C - \hat{C}$ and combining with the inequality on MW_2^2 gives Eq. (A.IV.17) in expectation. □

A.IV.3.3 Minimisation of $\mathrm{EM} - \mathrm{MW}_2^2$: Local Optima and Weight Fixing

In practice, the minimisation of the energy $X \longmapsto \mathrm{MW}_2^2(F_X^T(\theta_0), \nu)$ for some initialisation $\theta_0 \in \Theta$ and a target GMM ν comes with numerous challenges. The first hurdle is the "outer" minimisation of the MW_2^2 cost. To illustrate this difficulty, we begin with a study of the simpler energy $\mu \longmapsto \mathrm{W}_2^2(\mu, \nu)$ for a fixed (discrete) measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to the weights and support of the discrete measure $\mu = \sum_{i=1}^n a_i\delta_{x_i}$. This setting corresponds to the optimisation of MW_2^2 with known covariances, and thus highlights practical bottlenecks for the minimisation of the complete energy at stake, $\mathrm{EM} - \mathrm{MW}_2^2$. The objective of this section is to provide a theoretical rationale for fixing the mixture weights in practical applications, which is to say using Algorithm A.IV.2 instead of standard EM (Algorithm A.IV.1).

Local Minima of the Discrete 2-Wasserstein Distance. We focus on a particular instance of the minimisation of $\mu \longmapsto \mathrm{W}_2^2(\mu, \nu)$, and show the existence of a strict local minimum. We parametrise a discrete measure $\mu \in \mathcal{P}(\mathbb{R})$ with a support size of 3 as follows:

$$\forall \alpha \in [-\tfrac{1}{6}, \tfrac{1}{6}]^2,\ \forall \eta \in (-\tfrac{1}{2}, \tfrac{1}{2})^3,\ \mu_{\alpha,\eta} := (\tfrac{1}{6} + \alpha_1)\delta_{\eta_1} + (\tfrac{1}{6} + \alpha_2)\delta_{\eta_2} + (\tfrac{2}{3} - \alpha_1 - \alpha_2)\delta_{1+\eta_3},$$

and we fix a target measure $\nu := \frac{1}{3}(\delta_0 + \delta_{1-\varepsilon} + \delta_{1+\varepsilon})$ for a fixed $\varepsilon \in (0, \frac{1}{2})$. The energy to minimise is then:

$$\mathcal{E}_3 := \left\{ \begin{array}{ccc} [-\frac{1}{6}, \frac{1}{6}]^2 \times (-\frac{1}{2}, \frac{1}{2})^3 & \longrightarrow & \mathbb{R} \\ (\alpha, \eta) & \longmapsto & \mathrm{W}_2^2(\mu_{\alpha,\eta}, \nu) \end{array} \right. . \tag{A.IV.18}$$

Obviously, the energy $(\alpha, \eta) \longmapsto \mathrm{W}_2^2(\mu_{\alpha,\eta}, \nu)$ has a global minimum with value 0 at all (α, η) such that $\mu_{\alpha,\eta} = \nu$. However, on the region with $(\alpha, \eta) \in [-\frac{1}{6}, \frac{1}{6}]^2 \times (-\frac{1}{2}, \frac{1}{2})^3$, we show in Section A.IV.6.4.1 that the energy $\mathcal{E}_3$ has a minimum at $\alpha = 0_{\mathbb{R}^2}$ and $\eta = 0_{\mathbb{R}^3}$, with value $\mathcal{E}_3(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3}) > 0$. We represent our setup in Fig. A.IV.1. Note that for the case $n = 2$ we can show that there are unique local minima, see Section A.IV.6.4.2.

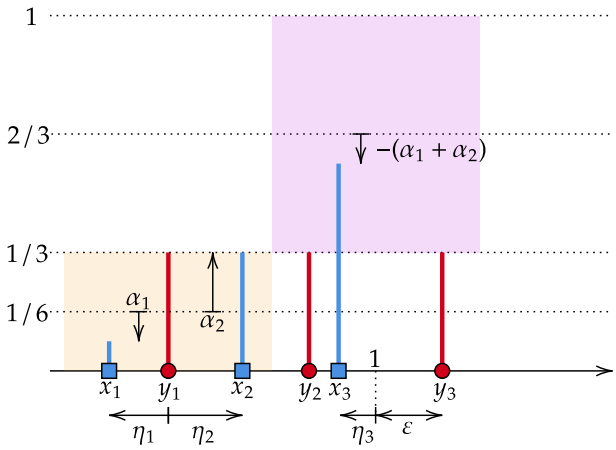


Figure A.IV.1: Setup for the local minimum of $\mathrm{W}_2^2(\mu_{\alpha,\eta}, \nu)$. The support of μ is represented with blue squares, and its weights with vertical blue lines. For the target ν, its support is red squares and its weights red lines. We consider a specific region where the points $x_1 = \eta_1$ and $x_2 = \eta_2$ stay within $(-\frac{1}{2}, \frac{1}{2})$ and the weights $a_1 = \frac{1}{6} + \alpha_1$ and $a_2 = \frac{1}{6} + \alpha_2$ stay within $[0, \frac{1}{3}]$, as represented by the orange rectangle. Likewise, the point $x_3 = 1 + \eta_3$ must stay within $(\frac{1}{2}, \frac{3}{2})$ and its weights $a_3 = \frac{2}{3} - (\alpha_1 + \alpha_2)$ must stay in $[\frac{1}{3}, 1]$, as shown with the purple rectangle.

Essential Stationary Points for the $\mathrm{EM} - \mathrm{MW}_2^2$ Loss. We have seen in the previous paragraph that optimising $\mu \longmapsto \mathrm{W}_2^2(\mu, \nu)$ with respect to the weights and support of μ can lead to local minima, which are not global minima. An additional difficulty arises when optimising the energy

$$\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2} := X \longmapsto \mathrm{MW}_2^2(\mu(F(\theta, X)), \nu), \tag{A.IV.19}$$

with one iteration F of the EM algorithm, due to the update on the weights. The arising issue is that at some problematic points X which are often converged to in practice, the gradient $\partial_X \mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}(X)$ becomes extremely small, in particular when the covariances are highly localised. This leads in practice to undesirable convergence to an (essential) local minimum, as illustrated by an example in Section A.IV.4.3. We provide a theoretical explanation in a simple case in Section A.IV.6.4.3. To avoid these numerical issues, we propose fixing the weights of the GMMs in the EM steps by using Algorithm A.IV.2.

Our theoretical observations suggest that considering GMMs with uniform weights and using fixed-weights EM (Algorithm A.IV.2) is a more stable alternative to standard EM (Algorithm A.IV.1). In practice, we believe it is also preferable to keep the same number of components K between the compared GMMs for additional stability. Note that an identifiability issue remains with GMMs: if the means and covariances of two modes coincide, then the GMM can also be written by fusing both components and adding their weights. At this stage, it remains

unclear whether this phenomenon has an impact on the optimisation behaviour (note that we never observed it in our experiments).

A.IV.3.4 Unbalanced GMM-OT

Starting from the discrete formulation of the MW_2 distance, we relax the constraints on the transport plan π, penalising the marginal conditions instead of enforcing them in the optimisation problem. The resulting optimisation problem defines an unbalanced GMM-OT distance on the set $\mathrm{GMM}_d^+(\infty)$ of GMMs with positive weights on $\mathbb{R}^d$, as in [LMS18]. Given two Gaussian mixtures

$$\mu = \sum_{k_0=1}^{K_0} w_{k_0}^{(0)} g_{k_0}, \quad \nu = \sum_{k_1=1}^{K_1} w_{k_1}^{(1)} g_{k_1} \ \in \mathrm{GMM}_d^+(\infty),$$

and regularisation parameters $(\lambda_0, \lambda_1) \in (0, +\infty)^2$, the unbalanced GMM-OT cost is defined as:

$$\mathrm{UMW}_2^2(\mu, \nu; \lambda_0, \lambda_1) := \min_{\pi \in \mathbb{R}_+^{K_0 \times K_1}} \sum_{k_0, k_1} \pi_{k_0,k_1} \mathrm{W}_2^2(g_{k_0}, g_{k_1}) + \lambda_0 \,\mathrm{KL}(\pi \mathbf{1} | w_0) + \lambda_1 \,\mathrm{KL}(\pi^\top \mathbf{1} | w_1), \tag{A.IV.20}$$

where we recall that for $a, b \in (0, +\infty)^K$, the Kullback-Leibler divergence is $\mathrm{KL}(a|b) := \sum_k a_k \log(\frac{a_k}{b_k})$. We have seen that MW_2 is a particular discrete Kantorovich problem, and likewise the unbalanced GMM-OT distance UMW_2 is simply a unbalanced discrete OT problem with a particular cost matrix.

Given the numerical challenges of optimising the weights in the balanced formulation (see the discussion in Section A.IV.3.3), we introduce this variant as a possibly more stable alternative. We suspect that the underlying geometry on the weights induced by unbalanced OT [LMS18; Chi+18b] is more amenable to optimisation. Furthermore, unbalanced OT has been shown by [Fat+21a] to be stable with respect to minibatch sampling, which is paramount for large-scale machine learning applications.

A.IV.4 Illustrations and Quantitative Study of Gradient Methods

A.IV.4.1 Practical Implementation

For practical implementation of the EM algorithm, some specific implementation strategies are required to ensure numerical stability, in particular when computing gradients. The first technique is applied in the E step, and consists in computing the responsibilities $\gamma_{ik}(\theta_t)$ in logarithmic space and using the so-called "log-sum-exp trick"[3] to compute the normalisation in Eq. (A.IV.3). Furthermore, to stabilise the (differentiable) expression of the Gaussian density $g_{m,\Sigma}(x)$, we leverage the Cholesky decomposition of the covariance matrix Σ, which uniquely decomposes $\Sigma = LL^\top$ where $L \in \mathbb{R}^{d \times d}$ is a lower triangular matrix. In particular, the computation of the inverse is simplified by solving triangular systems, and the determinant of Σ is simply $\det \Sigma = (\prod_a L_{aa})^2$ (which we compute in logarithmic space as well).

Another important implementation aspect concerns a differentiable implementation of the matrix square root of positive semi-definite symmetric matrices. This is required in the computation of the Bures distance Eq. (A.IV.16) for the MW_2 distance. Unfortunately, the naive implementation using the spectral decomposition suffers from numerical instability when the eigenvalues are too close[4]. Leveraging an explicit formula for the gradient of the matrix square root (detailed in Section A.IV.6.5), we circumvent these numerical issues by implementing our own differentiable square root function with an explicit gradient.

As is done in scikit-learn's implementation of the EM algorithm, we have an optional regularisation term $\varepsilon_r \geq 0$ for the covariance matrices Σ_k to ensure positive-definiteness. The idea is to replace the update $\Sigma_k^{(t+1)}$ with $\Sigma_k^{(t+1)} + \varepsilon_r I_d$ to enforce a minimum eigenvalue of ε_r. This regularisation was particularly crucial for numerical stability in higher-dimensional cases where

[3] for instance, see this blog post for an explanation of this well-known trick.

[4] as explained in the PyTorch documentation for `torch.linalg.eigh()`.

the covariances were almost singular, which led to gradient explosion. In the larger-scale examples from Section A.IV.5, we chose a heuristic which sets $\varepsilon_r := 10^{-4} \times s_{\max}$, where $s_{\max}$ is the largest eigenvalue of the covariances of a GMM fitted on the target data.

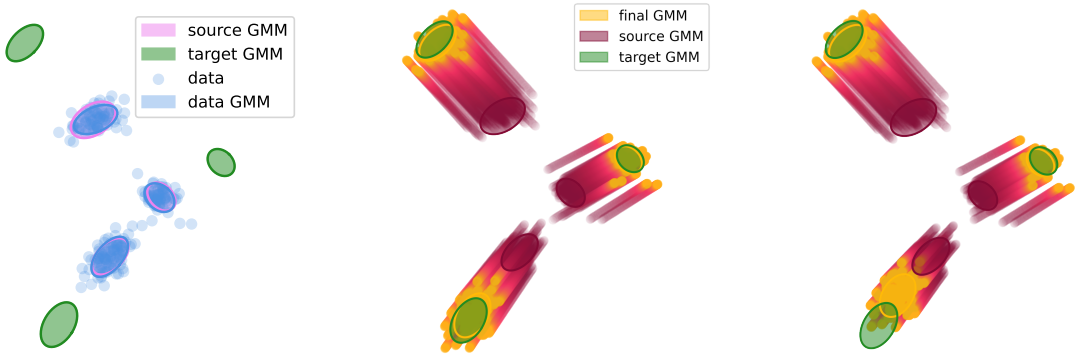
A.IV.4.2 Flow of $\mathrm{EM} - \mathrm{MW}_2^2$ with Fixed Weights in 2D

In this section, we illustrate the use of differentiable EM for OT by numerically computing the flow (i.e. gradient descent) of the following energy:

$$\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2} := X \in \mathbb{R}^{n\times 2} \longmapsto \mathrm{MW}_2^2(\mu(F_X^T(\theta_0)), \nu),$$

for a fixed target GMM ν, an initialisation θ_0 and a number of EM steps T. We use a variant of EM presented in Algorithm A.IV.2 that **fixes the mixture weights** in this experiment. We will compare three gradient computation methods to compute (or approximate) the gradient of $F_X^T(\theta_0)$ with respect to X, within the gradient descent of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$, performed using automatic differentiation. The setup is as follows: the initial dataset $X \in \mathbb{R}^{200\times 2}$ corresponds to samples of a GMM μ_0 with 3 components, and we want to displace this point cloud to match a target GMM ν with 3 components. We represent the setup in Fig. A.IV.2a and the flow for AD method in Fig. A.IV.2b.

The results for the AI method are both visually and quantitatively very close, however the experiment took six times longer to run for AI. We observe satisfactory convergence of the flow of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ towards the target GMM ν. In many applications involving fixed EM weights (Algorithm A.IV.2), we observe that particles follows rectilinear trajectories towards, which is a similar behaviour to Wasserstein flows of W_2^2 (see [CNR25, Section 5.3]). We interpret this phenomenon as a consequence of the fixed weights, which translate to a Lagrangian viewpoint on the GMMs. In simple cases, the MW_2^2-optimal plans between the GMMs may not change during the flow, and thus the particles are moved along the induced (rectilinear) trajectories between each GMM component (see [DD20, Proposition 4]). In Fig. A.IV.2c, we show the flow with the OS method, which converges slower and to an unsatisfactory stationary point. This is due to the fact that OS requires a contraction assumption that is not verified for EM. OS was comparable in computation time to AD. We also experimented with the Warm Start flow from Algorithm A.IV.3, which is a different minimisation method to minimise $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$, yet yielded almost identical results to the AD, with a 40% lower computation time.



(a) Experimental setup for the flow of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$.

(b) AD (Warm-Start and AI are almost identical).

(c) One Step method.

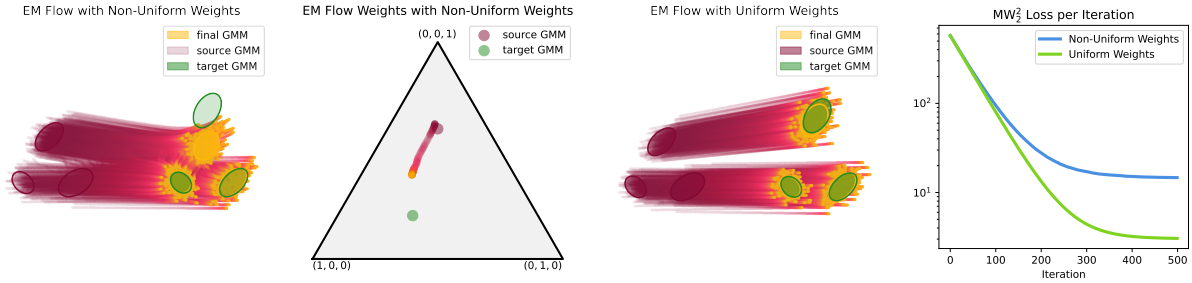
Figure A.IV.2: Comparison of experimental setup and flows of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ using different methods. The dark shades of purple correspond to earlier iterations, and the yellow shades to later iterations.

A.IV.4.3 Flow of $\mathrm{EM} - \mathrm{MW}_2^2$ in 2D: Discussion on Uniform Weights

We now consider a similar setting to Section A.IV.4.2, but without fixing the weights in the EM algorithm, i.e. using standard EM Algorithm A.IV.1. We compare two settings: the first

with non-uniform weights $w_0 := (\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$ for the initial GMM, and weights $w_1 := (\frac{1}{2}, \frac{3}{10}, \frac{1}{5})$ for the target GMM; and the second with uniform weights for both. In Fig. A.IV.3, we show the flow of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ with the AD method. We observe in Fig. A.IV.3a that the flow for non-uniform weights converges to an unsatisfactory local minimum, with a final GMM weight of `[0.41039274, 0.2030173, 0.38658995]` instead of the target `[0.5, 0.3, 0.2]`, as shown on the simplex in Fig. A.IV.3b. In contrast, the flow for uniform weights presented in Fig. A.IV.3c converges to the target GMM and achieves a substantially lower energy, as reported in Fig. A.IV.3d. The weights stay close to uniform, with a final GMM weight of `[0.33314218, 0.30985782, 0.357]`.

The optimisation failure in the non-uniform case is due to the essential stationary point problem illustrated in Section A.IV.3.3: intuitively, to change the weights of the current GMM, the particles need to change components, but this is not possible if the components are too distant. While our framework encompasses the case of non-uniform weights, as illustrated theoretically and experimentally, it appears that the non-uniform weight setting is impractical. As a result, we recommend using uniform weights, in particular using the fixed-weights EM approach (Algorithm A.IV.2) for speed and stability.



(a) Particle flow (NU). (b) Weight evolution on the simplex $\triangle_3$ (NU). (c) Particle flow (U). (d) Energy evolutions.

Figure A.IV.3: Flow of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ with the AD method and standard EM Algorithm A.IV.1. We compare two settings: one with non-uniform GMM weights (NU) and one with uniform weights (U).

A.IV.4.4 Stochastic EM $-$ MW_2^2 Flow with Fixed Weights

We consider a similar setting to Section A.IV.4.2 but introduce stochasticity in the flow at each step by performing EM only on a subsample of the optimised source point cloud and of the target point cloud. While we illustrate the technique on a toy example here, this "minibatch" stochastic gradient descent method is useful in practice when the dataset size is too large for simultaneous optimisation [Fat+20; Fat+21b; Fat+21a; Ton+24]. The same principle is applied to the image generation task in Section A.IV.5.4. We observe in Fig. A.IV.4 that the general trajectory remains similar to the deterministic case. Notice that in this setting, the components are sufficiently close together to interact, yielding non-rectilinear trajectories when points are influenced by multiple components. This is amplified by the stochasticity of the method.

A.IV.4.5 Quantitative Study of EM Convergence and Gradients

We study the impact of the number of points n, components K and EM iterations T on the convergence of EM iterations (to a fixed point of F), the local contractivity of F around the fixed point, and the gradient approximation methods introduced in Section A.IV.2.3.

The experimental setting is as follows: for each of the three parameters n, K, T separately (say n), we consider a range of values (say $n \in \{100, 200, 500, 1000, 1500, 2000\}$) with the others fixed. For each value of the parameter in this range, and for three different GMMs in $\mathrm{GMM}_d(K)$ with $d = 3$ (shown in Section A.IV.6.1), we sample the data from $X \sim \mu_0^{\otimes n}$ 60 times, then run the EM algorithm for T iterations. We run the experiments on three different GMMs taken with random parameters (adding a small term $10^{-14} I_d$ to the covariances to avoid vanishing

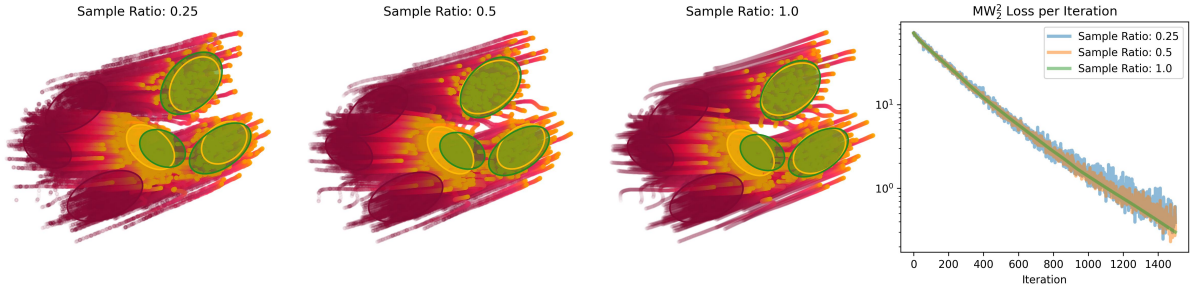


Figure A.IV.4: Stochastic Flow of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ for Algorithm A.IV.2 with the full automatic differentiation method. We vary the sub-sampling ratio $r \in (0, 1]$, which corresponds to performing EM on only $\lceil r \times n \rceil$ random points from the current point cloud at each step.

eigenvalues). We observe the Mean Square Error of the fixed point property $F(\theta_T, X) \approx \theta_T$ by evaluating $\frac{1}{p}\|\theta_T - F(\theta_T, X)\|_2^2$ with $p := K + Kd + Kd^2$, measuring the quality of convergence of the EM algorithm. To study the local contractivity of F, we compute the spectral norm $\|\partial_\theta F(\theta_T, X)\|_{\mathrm{op}}$: if this is close to 0, then locally the iterated function F_X has a tame behaviour and the OS method is expected to work well, while if it is close or larger to 1, the local landscape is difficult and the OS method is expected to fail. Finally, we compare the OS and AI gradients (from Eqs. (A.IV.10) and (A.IV.11)) to the reference AD gradient by computing the relative MSEs $\frac{1}{p}\|J_{\mathrm{OS}} - J_{\mathrm{AD}}\|_2^2 / (\frac{1}{p}\|J_{\mathrm{AD}}\|_2^2)$ and $\frac{1}{p}\|J_{\mathrm{AI}} - J_{\mathrm{AD}}\|_2^2 / (\frac{1}{p}\|J_{\mathrm{AD}}\|_2^2)$, where J_{AD} is the AD gradient, which serves as a baseline (see Section A.IV.6.2 for a discussion on this choice).

Impact of the number of samples n We begin by fixing $d = 3$, $K = 3$ and $T = 30$ and varying $n \in \{100, 200, 500, 1000, 1500, 2000\}$. The results are shown in row 1 of Fig. A.IV.5, and we observe that EM appears to converge to a fixed-point for all n, albeit with a large variance in the MSE depending on the sampled GMMs. The spectral norm of the Jacobian is often close to 0.6 and has no clear trend with n, hence we expect the OS method to be a very coarse approximation of the true gradient. The quality of the OS gradient is relatively poor, and substantially worse than the AI gradient, whose median MSE is much smaller, but suffers from very high variance (in log space). Comparing GMMs shows that a precise EM convergence leads to high precision for the AI gradient.

Impact of the number of EM iterations T Finally, we fix $n = 200, d = 3$ and $K = 3$, and vary $T \in \{1, 2, 5, 10, 15, 20, 30, 40\}$ in row 3 of Fig. A.IV.5. Reassuringly, increasing the number of iterations T leads to improved convergence of the EM algorithm to better-conditioned points. The convergence speed seems heavily dependent on the GMM, with an additional variance caused by the dataset sampling. In the favourable settings for larger T, the AI approximation substantially outperforms the OS approximation, but suffers from higher variance. Since the spectral norm of the Jacobian stabilises to values of the order of 0.5, the OS method plateaus at coarse MSEs, even for larger T.

Concerning the impact of the number of components K, we defer to Section A.IV.6.6.1, since the findings are less conclusive.

A.IV.5 Applications of Differentiable EM

A.IV.5.1 Barycentre Flow in 2D

Wasserstein barycentres [AC11] and their notoriously challenging computation [CD14; Álv+16; AB22; TDG24] are active fields of research. In this section, we illustrate the use of differentiable EM to flow a point cloud towards a barycentre of GMMs. Given M point clouds $Y_i \in \mathbb{R}^{n\times 2}$, our goal is to optimise a point cloud $X \in \mathbb{R}^{n\times 2}$, initialised as random normal noise, towards a barycentre (with uniform weights) of GMMs (ν_i) fitted from (Y_i). Specifically, we solve

$$\min_{X\in\mathbb{R}^{n\times 2}} \sum_{i=1}^{M} \mathrm{MW}_2^2\left(\mu(F_X^T(\theta_0)), \nu_i\right)$$

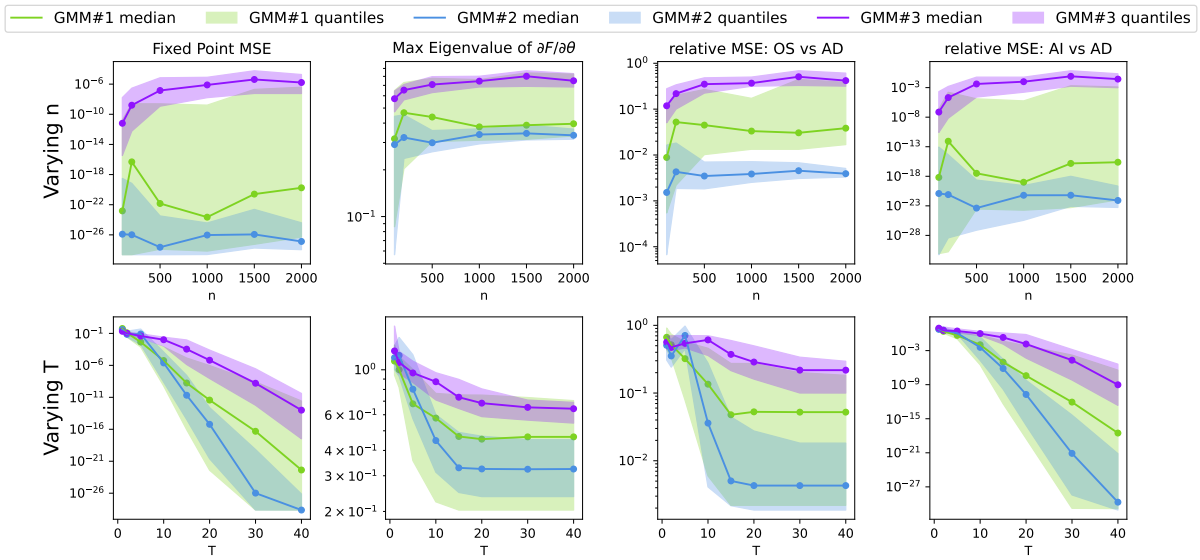


Figure A.IV.5: Varying the number of samples n and the number of iterations T, we study the convergence of EM, the local contractivity of F, and the MSEs of the OS and AI gradients against the AD gradient.

with respect to X. The GMMs ν_i are fitted beforehand, and $\mu(F_X^T(\theta_0))$ is the running EM estimation of the optimised cloud X. We illustrate the results in Fig. A.IV.6, where $M := 3$, $K := 2$ and $n := 500$. This method can be adapted to compute more general barycentres, as presented in Section A.IV.6.6.2.

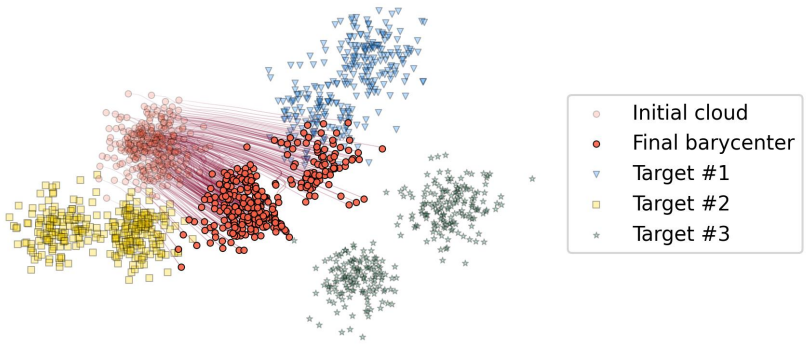
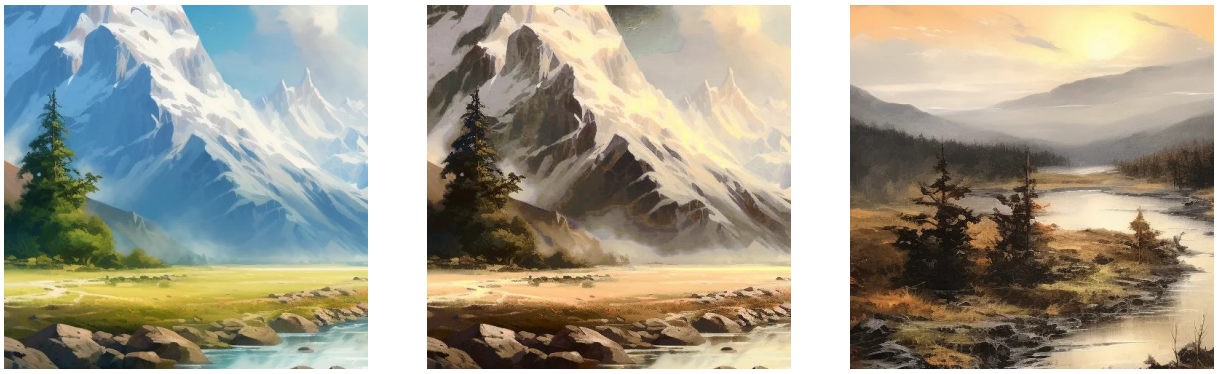


Figure A.IV.6: EM – MW_2^2 flow displacing particles in order to make their EM output approach a barycentre of three target GMMs.

A.IV.5.2 Colour Transfer

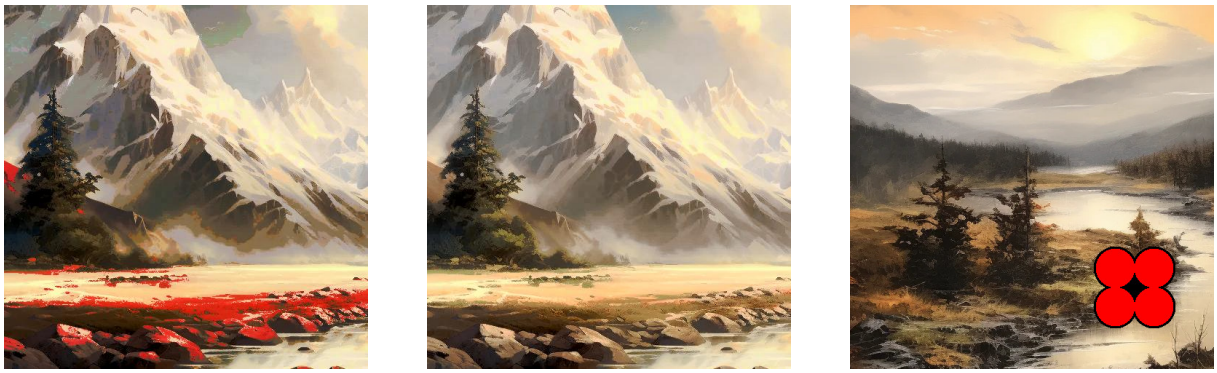
Colour transfer is a well-known imaging task where OT techniques have been used extensively [Rei+02; Del04; PK07; PKD07; PPC10; Rab+12; RFP14], it consists in transforming the RGB colour distribution of a source image to match that of a target image. We propose the following approach: we initialise an image X as the source image, and optimise it to minimise the MW_2^2 cost between a GMM fitted on X (seen as an RGB point cloud), and a target GMM fitted on the colour distribution of the target image. We use the *Warm-start EM* method from Algorithm A.IV.3, choose $K := 10$ components and use fixed uniform GMM weights (Algorithm A.IV.2 for EM) to avoid being trapped in a local minimum, as seen in Section A.IV.3.3. We present some results in Fig. A.IV.7, and provide additional discussions about the optimisation choices in Section A.IV.6.6.3. Even with only $K = 10$ components, the colour transfer preserves both details and global consistency, and in the resulting colour scheme, the source image appears to enjoy stronger contrast in this example.



(a) Source (b) Result (c) Target

Figure A.IV.7: Colour transfer using $\mathrm{EM} - \mathrm{MW}_2^2$ from the source image (a) to the target image (b).

Balanced OT methods may be sensitive to outliers in the distributions, leading to artifacts in the colour transfer results if colour aberrations are present in the target. Unbalanced OT methods [LMS18] can mitigate this issue [Chi+18a; Bon+24]. We now consider the unbalanced variant of the Mixture Loss defined in Section A.IV.3.4: by relaxing the marginals constraints, it can ignore outliers in the input distributions. Specifically, in Fig. A.IV.8, we consider an illustration with a corrupted target image where a patch of red has been added, and observe that the unbalanced approach is more robust to this aberration, displaying no leakage of the red artefact.



(a) Balanced result (b) Unbalanced result (c) Corrupted target

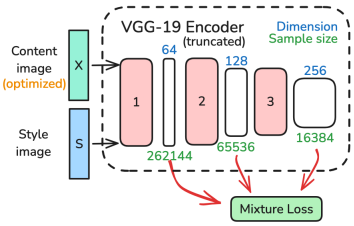
Figure A.IV.8: Unbalanced colour transfer with regularisations $\lambda_1 = 10$ (source) and $\lambda_2 = 0.1$ (target).

A.IV.5.3 Neural Style Transfer

We apply our distance to Gatys et al.'s neural style transfer [GEB16]. The goal is to generate an image that combines the content of one image X_0 with the artistic style of another image Y. We rely on a pre-trained VGG-19 network [SZ15] (see Fig. A.IV.9a) to encode our image and extract relevant features from three specific layers ℓ. Given an image X in $\mathbb{R}^{3\times H\times W}$, we note these features $\mathrm{VGG}_{1\dots\ell}(X)$. Starting with the content image as X_0, we optimise X such that its features progressively match those of the style image Y. The target mixtures $\overline{\mu}_\ell^{\text{style}}$ are fitted at the beginning of the procedure on style features $\mathrm{VGG}_{1\dots\ell}(Y)$. Notably, the target style is encoded as a low-dimensional GMM, and the reference image is not needed during training, unlike in [GEB16]. Our objective is a weighted sum of Mixture Losses between the optimised features $\mathrm{VGG}_{1\dots\ell}(X)$ and the target $\overline{\mu}_\ell^{\text{style}}$, for each layer ℓ in $\{1, 2, 3\}$: we solve

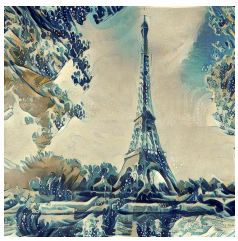
$$\min_{X\in\mathbb{R}^{3\times H\times W}} \sum_{l=1}^{3} \lambda_\ell \,\mathrm{MW}_2^2\left(F(\theta_{\text{init}}, \mathrm{VGG}_{1\dots\ell}(X)), \overline{\mu}_\ell^{\text{style}}\right). \tag{A.IV.21}$$

The weights λ_ℓ follow Gatys et al.'s scheme ([GEB15]) of $1/d_\ell^2$ where d_ℓ is the dimension of features in layer ℓ. We fit $K = 3$ Gaussian components at each layer. The chosen procedure for fitting Gaussian mixtures is still *Warm-Start EM*. We optimise using 100 iterations of Adam with learning rate 0.01, which takes approximately 20 seconds using CUDA on a RTX 4000. The example results shown in Fig. A.IV.9 illustrate the ability of GMM to encode (and store) an image style which has to the best of our knowledge never been shown. The experimental setup is further detailed in Section A.IV.6.6.4, along with additional examples.

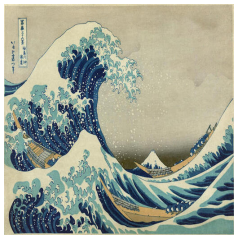


(a) Model architecture

(b) Source image



(c) Generated image



(d) Style image

Figure A.IV.9: Style transfer method inspired by [GEB15]: setup and example result.

A.IV.5.4 Image Generation

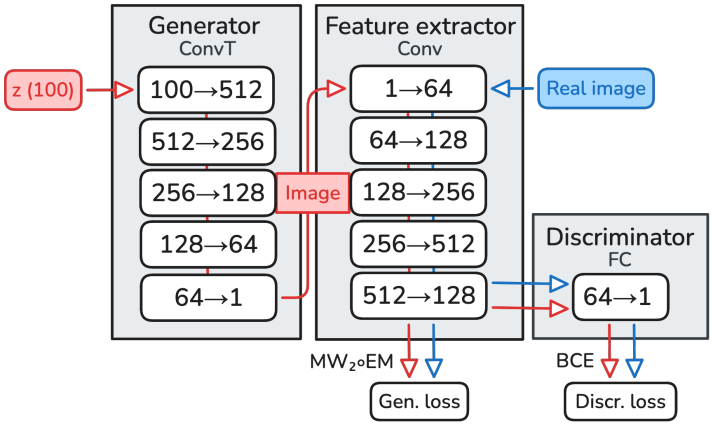
As a proof of concept, we train a Generative Adversarial Network (GAN) [Goo+14; ACB17; DZS18] using the Mixture–Wasserstein distance as a regularisation. The idea is to encourage the generator to produce images which have similar features to real images: while this does not suffice to produce realistic images, it can guide the training optimisation out of spurious local minima induced by the notoriously unstable GAN training ([ACB17]). Given a batch size b and a latent dimension ℓ, we independently sample $x_1, \cdots, x_b \sim \mathcal{N}(0_\ell, I_\ell)$, and we draw $y_1, \cdots, y_b$ from our dataset of real images. The aim of our *generator* $\mathtt{G} : \mathbb{R}^\ell \longrightarrow \mathbb{R}^{C\times H\times W}$ is to map these Gaussian samples to real images of height H, width W and C channels. We also introduce a *feature extractor* $\mathtt{F} : \mathbb{R}^{C\times H\times W} \longrightarrow \mathbb{R}^f$ who learns to map images to relevant features of dimension f, and a *discriminator* $\mathtt{D} : \mathbb{R}^f \longrightarrow \mathbb{R}$ whose goal is to discriminate real and fake images by producing different vectors of features. The full adversarial objective on a noise batch X and a real image batch Y is then defined as follows:

$$\min_{\mathtt{G}} \left[\mathrm{MW}_2^2 \left(F^T(\theta_0, \mathtt{F} \circ \mathtt{G}(X)), F^T(\theta_0', \mathtt{F}(Y)) \right) + \max_{\mathtt{F},\mathtt{D}} \left(\sum_{i=1}^{b} \log \mathtt{D} \circ \mathtt{F}\,(y_i) + \log\left(1 - \mathtt{D} \circ \mathtt{F} \circ \mathtt{G}(x_i)\right) \right) \right], \tag{A.IV.22}$$

where θ_0 and θ_0' are chosen using *k-means* initialisation. Note that the feature extractor $\mathtt{F}$ intervenes in the MW_2^2 regularisation, but this regularisation is not used for the optimisation $\mathtt{F}$. We optimise these losses using Stochastic Gradient Descent and *Full Automatic Differentiation* (see Section A.IV.2.3). For every sampled batch, we alternate one step on $\mathtt{F}$ and $\mathtt{D}$, then one step on $\mathtt{G}$. The generator is encouraged to produce images with similar features to the real ones, and aims to fool the discriminator. The feature extractor $\mathtt{F}$ attempts to extract features such that $\mathtt{D}$ is able to discriminate real from fake images. The full network architecture is described in Section A.IV.5.4. Note that as in other GANs [Goo+14; ACB17; DZS18], we do not optimise over the full dataset and rather optimise by sampling mini-batches at each step, incurring a seldom-discussed bias [Fat+20; Fat+21b; Fat+21a; Ton+24] in the favour of computational efficiency.

A.IV.5.5 Texture Synthesis

We perform texture synthesis using a novel method inspired by [GLR18; LDD23; Hou+23]. We initialise the synthesised texture using a stationary Gaussian field of same mean and covariance as the target texture. We then optimise a weighted sum of Mixture Losses over different scales

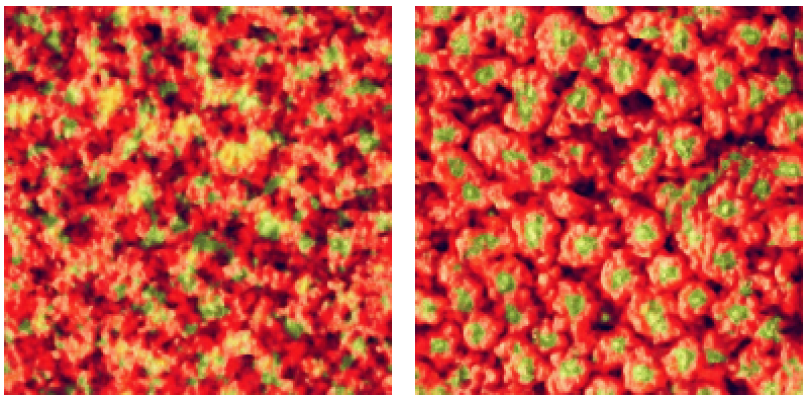


(a) GAN architecture (adapted from DCGAN [RMC16])

(b) MNIST generation with $K = 10$.

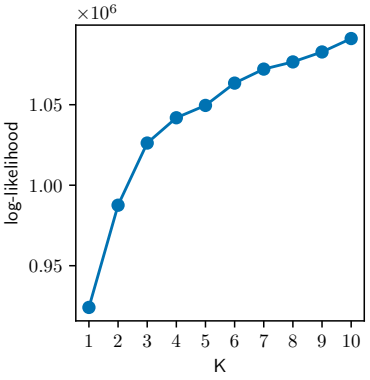
Figure A.IV.10: Image generation with the MW_2^2-GAN from Eq. (A.IV.22).

in the patch space (we refer the reader to Section A.IV.6.6.5 for a full explanation). When doing multi-scale synthesis, we simply choose to downscale the images by a factor 2^s for s between 0 and S, so that the image downscaled by a factor 2^S has size at least 16×16. In our experiments, we choose to fit $K = 4$ Gaussian components in our mixtures. As illustrated in Fig. A.IV.11, this corresponds (roughly) to the elbow of the model's log-likelihood and gives more convincing results. As for patch sizes, notice that choosing a patch size of 1 amounts to optimising the colour intensities directly, i.e. performing standard colour transfer. In Fig. A.IV.12, we compare the results of taking 4×4 and 8×8 patches.



(a) $K = 1$ components

(b) $K = 4$ components



(c) Log-likelihood of fit w.r.t. K

Figure A.IV.11: Choosing the number of components K for texture synthesis.

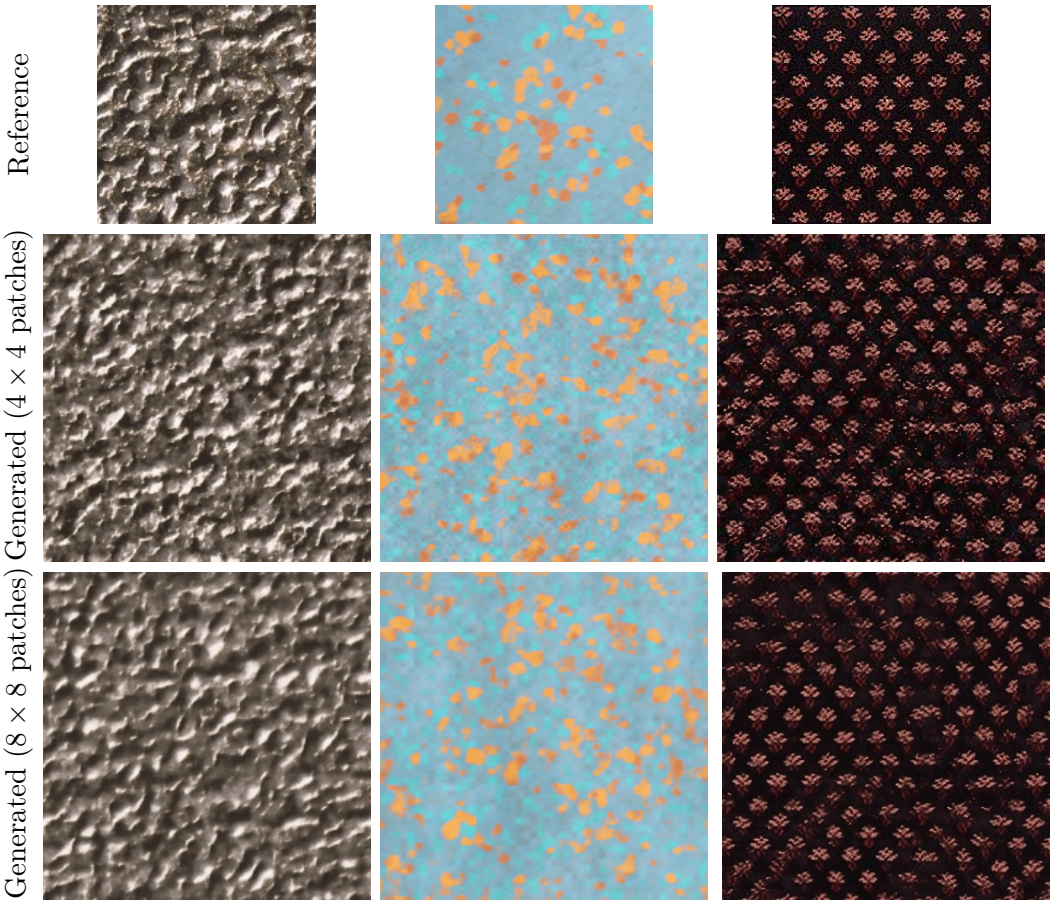


Figure A.IV.12: Multi-scale texture synthesis with $K = 4$ components for different patch sizes.

Acknowledgements

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

A.IV.6 Supplementary Material

A.IV.6.1 Specific GMMs Used in Section A.IV.4.5

In Fig. A.IV.13, we show the three different GMMs used in our experiments, which were sampled randomly with fixed seeds. We observe that GMM #1 has relatively large variances, allowing for less numerical difficulty, while GMM #2 and (to an even larger extent) GMM #3 have some almost-degenerate covariances, leading to increased numerical instability.

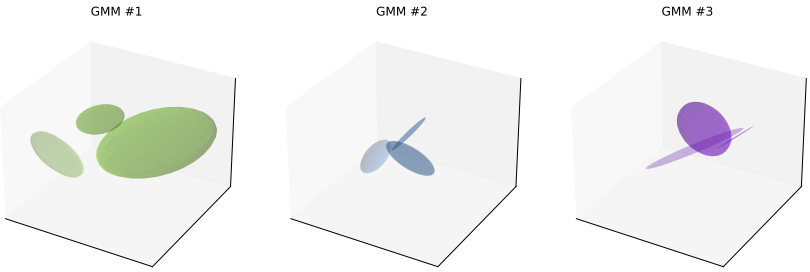


Figure A.IV.13: 3D representation of the three GMMs used in Section A.IV.4.5 to compare EM gradient methods.

A.IV.6.2 Discussion on Gradient Ground Truths

As discussed in Section A.IV.2.3, the most natural baseline for the computation of a gradient of EM is the full automatic differentiation method (AD). Unfortunately, this strategy is only theoretically sound under strong assumptions [Gil92; Bec94; MO20], and may be prone to the propagation of numerical errors across iterations. Due to this latter issue in particular, one may not claim that the AD gradient is exact. In order to discuss its use as a baseline (in particular in Section A.IV.4.5), we compare it to the gradient computed using the finite differences approximation (FD) with a step size ε_{FD}, implemented using `torch.autograd.gradcheck._get_numerical_jacobian` in Pytorch [Pas+19]. The experimental setting is as in Section A.IV.4.5: in Fig. A.IV.14, we observe the relative MSEs $\|J_{\text{FD}} - J_{\text{AD}}\|_2^2 / \|J_{\text{AD}}\|_2^2$ varying the FD step size $\varepsilon_{\text{FD}} \in \{10^{-5}, 10^{-6}, 10^{-7}\}$ on three different GMMs, taking each time 60 samples for the data X with $n = 300$, $d = 3$ and running EM with $K = 3$ and $T = 30$. Note that the FD approximation is extremely intensive numerically, prohibiting its use in practice and even for extensive testing. First, we observe that the variability between data samples is very high, indicating sensitivity of the methods to the algorithm initialisation (and in turn, convergence). Furthermore, while AD is close to FD at numerical precision for GMMs #1 and #2, the approximation is substantially coarser with a relative error of the order of 10^{-5} for GMM #3, indicating numerical instability. Given that the FD method itself is a sensitive approximation that depends strongly on ε_{FD}, we conclude that there appears to be no reliable ground truth for an exact gradient, and we resort to AD as a baseline in our experiments. We leave the challenging question of quantifying the arising numerical errors for future work.

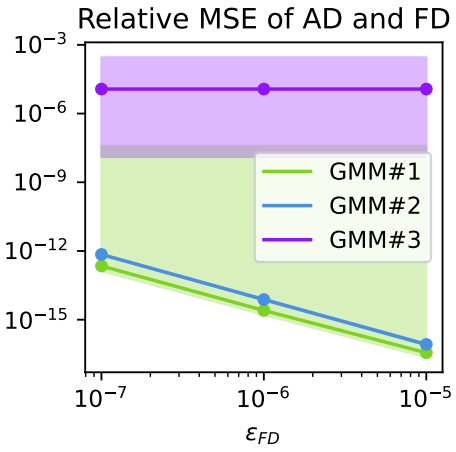


Figure A.IV.14: Relative MSE of the full automatic differentiation gradient (AD) against the finite differences approximation (FD) for three different GMMs and varying the FD step size ε_{FD}.

A.IV.6.3 Explicit Differential Expressions

A.IV.6.3.1 Differential of Gaussian Density

First, we compute the explicit differential of

$$g : \begin{cases} \mathbb{R}^d \times S_d^{++}(\mathbb{R}) \times \mathcal{X} & \longrightarrow & \mathbb{R}^n \\ (m, \Sigma, X) & \longmapsto & \left(\dfrac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\dfrac{1}{2}(x_i - m)^\top \Sigma^{-1}(x_i - m) \right) \right)_{i=1}^n \end{cases}, \tag{A.IV.23}$$

where we wrote $X = (x_1, \cdots, x_n)$. We first compute the differential with respect to m:

$$\frac{\partial g}{\partial m}(m, \Sigma, X) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n) \simeq \mathbb{R}^{n \times d} : \left[\frac{\partial g}{\partial m}(m, \Sigma, X) \right]_{i,\cdot} = \Sigma^{-1}(x_i - m) g_{m,\Sigma}(x_i). \tag{A.IV.24}$$

For the differential with respect to Σ, we remark that $T_\Sigma S_d^{++}(\mathbb{R}) = S_d(\mathbb{R})$, and use [PP08, Equation (49)] stating that $\partial_A \det A = (\det A)A^{-\top}$ and [PP08, Equation (61)] which states $\partial_A a^\top A^{-1} b = -A^{-\top} a b^\top A^{-\top}$, which yields:

$$\frac{\partial g}{\partial \Sigma}(m, \Sigma, X) \in \mathcal{L}(S_d(\mathbb{R}), \mathbb{R}^n) \hookrightarrow \mathbb{R}^{n \times d \times d} : \tag{A.IV.25}$$

$$\left[\frac{\partial g}{\partial \Sigma}(m, \Sigma, X)\right]_{i,\cdot,\cdot} = \frac{g_{m,\Sigma}(x_i)}{2}\left(-\Sigma^{-1} + \Sigma^{-1}(x_i - m)(x_i - m)^\top \Sigma^{-1}\right). \tag{A.IV.26}$$

Finally, the differential with respect to the data X reads

$$\frac{\partial g}{\partial X}(m, \Sigma, X) \in \mathcal{L}(\mathbb{R}^{n \times d}, \mathbb{R}^n) \simeq \mathbb{R}^{n \times n \times d} : \left[\frac{\partial g}{\partial \Sigma}(m, \Sigma, X)\right]_{i,j,\cdot} = -\delta_{i,j} g_{m,\Sigma}(x_i)\Sigma^{-1}(x_i - m). \tag{A.IV.27}$$

A.IV.6.3.2 Differential of γ

We now compute the differential of the function

$$\gamma : \begin{cases} \underbrace{\triangle_K \times (\mathbb{R}^d)^K \times (S_d^{++}(\mathbb{R}))^K}_{\Theta} \times \mathcal{X} & \longrightarrow & \mathbb{R}^{n \times K} \\ (\underbrace{w, (m_k), (\Sigma_k)}_{\theta}, X) & \longmapsto & \left(\dfrac{w_k g_{m_k,\Sigma_k}(x_i)}{\sum_l w_l g_{m_l,\Sigma_l}(x_i)}\right)_{i,k} \end{cases}. \tag{A.IV.28}$$

For notational convenience, we introduce

$$\forall (i,k) \in [\![1,n]\!] \times [\![1,K]\!],\ g_{i,k} := g_{m_k,\Sigma_k}(x_i),\ Z_i := \sum_{k'} w_{k'} g_{i,k'}. \tag{A.IV.29}$$

First, the tangent space of the simplex is the same everywhere: $T_w \triangle_K = \triangle_K^0 := (\mathbb{1}_K)^\perp$, and the differential with respect to w is

$$\frac{\partial \gamma}{\partial w}(\theta, X) \in \mathcal{L}\left(\triangle_K^0, \mathbb{R}^{n \times K}\right) \hookrightarrow \mathbb{R}^{n \times K \times K} : \left[\frac{\partial \gamma}{\partial w}(\theta, X)\right]_{i,k,l} = \frac{g_{i,k}}{Z_i}\left(\delta_{k,l} - \frac{w_k g_{i,l}}{Z_i}\right). \tag{A.IV.30}$$

Using Eq. (A.IV.24), we compute the differential of γ w.r.t. $\mathbf{m} = (m_k)$:

$$\frac{\partial \gamma}{\partial \mathbf{m}}(\theta, X) \in \mathcal{L}\left((\mathbb{R}^d)^K, \mathbb{R}^{n \times K}\right) \simeq \mathbb{R}^{n \times K \times K \times d} : \tag{A.IV.31}$$

$$\left[\frac{\partial \gamma}{\partial \mathbf{m}}(\theta, X)\right]_{i,k,l,\cdot} = \frac{w_k g_{i,k}}{Z_i}\left(\delta_{k,l} - \frac{w_l g_{i,l}}{Z_i}\right)\Sigma_l^{-1}(x_i - m_l). \tag{A.IV.32}$$

Similarly, using Eq. (A.IV.26), we compute the differential of γ w.r.t. $\mathbf{\Sigma} = (\Sigma_k)$:

$$\frac{\partial \gamma}{\partial \mathbf{\Sigma}}(\theta, X) \in \mathcal{L}\left((S_d(\mathbb{R}))^K, \mathbb{R}^{n \times K}\right) \hookrightarrow \mathbb{R}^{n \times K \times K \times d \times d} : \tag{A.IV.33}$$

$$\left[\frac{\partial \gamma}{\partial \mathbf{\Sigma}}(\theta, X)\right]_{i,k,l,\cdot,\cdot} = \frac{w_k g_{i,k}}{2Z_i}\left(\delta_{k,l} - \frac{w_l g_{i,l}}{Z_i}\right)\left(-\Sigma_l^{-1} + \Sigma_l^{-1}(x_i - m_l)(x_i - m_l)^\top \Sigma_l^{-1}\right). \tag{A.IV.34}$$

Finally, Eq. (A.IV.27) allows the computation of the differential of γ w.r.t.X:

$$\frac{\partial \gamma}{\partial X}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{n \times d}, \mathbb{R}^{n \times K}\right) \simeq \mathbb{R}^{n \times K \times n \times d} : \tag{A.IV.35}$$

$$\left[\frac{\partial \gamma}{\partial X}(\theta, X)\right]_{i,k,j,\cdot} = \frac{\delta_{i,j} w_k g_{i,k}}{Z_i}\left(-\Sigma_k^{-1}(x_i - m_k) + \frac{1}{Z_i}\sum_{l=1}^K w_l g_{i,l}\Sigma_l^{-1}(x_i - m_l)\right). \tag{A.IV.36}$$

A.IV.6.3.3 Differential of F

We introduce the coordinate notation

$$F(\theta, X) = (F_w(\theta, X), F_{\mathbf{m}}(\theta, X), F_{\mathbf{\Sigma}}(\theta, X)) \in \triangle_K \times (\mathbb{R}^d)^K \times (S_d^{++}(\mathbb{R}))^K,$$

hence the differentials of F with respect to θ and X can be written "block-wise" as follows:

$$\frac{\partial F}{\partial \theta} = \begin{pmatrix} \frac{\partial F_w}{\partial w} & \frac{\partial F_w}{\partial \mathbf{m}} & \frac{\partial F_w}{\partial \mathbf{\Sigma}} \\ \frac{\partial F_{\mathbf{m}}}{\partial w} & \frac{\partial F_{\mathbf{m}}}{\partial \mathbf{m}} & \frac{\partial F_{\mathbf{m}}}{\partial \mathbf{\Sigma}} \\ \frac{\partial F_{\mathbf{\Sigma}}}{\partial w} & \frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{m}} & \frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{\Sigma}} \end{pmatrix}, \quad \frac{\partial F}{\partial X} = \begin{pmatrix} \frac{\partial F_w}{\partial X} \\ \frac{\partial F_{\mathbf{m}}}{\partial X} \\ \frac{\partial F_{\mathbf{\Sigma}}}{\partial X} \end{pmatrix}.$$

Differentials of F_w. We now compute the differentials of F_w whose expression we remind from Eq. (A.IV.4):

$$F_w : \begin{cases} \Theta \times \mathcal{X} & \longrightarrow & \triangle_K \\ (\theta, X) & \longmapsto & \left(\frac{1}{n} \sum_{i=1}^n \gamma_{i,k}(\theta, X) \right)_{k=1}^K \end{cases}. \tag{A.IV.37}$$

By Eq. (A.IV.30), we have

$$\frac{\partial F_w}{\partial w}(\theta, X) \in \mathcal{L}\left(\triangle_K^0, \triangle_K^0\right) \hookrightarrow \mathbb{R}^{K \times K} : \left[\frac{\partial F_w}{\partial w}(\theta, X)\right]_{k,l} = \frac{1}{n} \sum_{i=1}^n \frac{g_{i,k}}{Z_i} \left(\delta_{k,l} - \frac{w_k g_{i,l}}{Z_i}\right). \tag{A.IV.38}$$

Likewise, Eq. (A.IV.32) yields

$$\frac{\partial F_w}{\partial \mathbf{m}}(\theta, X) \in \mathcal{L}\left((\mathbb{R}^d)^K, \triangle_K^0\right) \hookrightarrow \mathbb{R}^{K \times K \times d} : \tag{A.IV.39}$$

$$\left[\frac{\partial F_w}{\partial \mathbf{m}}(\theta, X)\right]_{k,l,\cdot} = \frac{1}{n} \sum_{i=1}^n \frac{w_k g_{i,k}}{Z_i} \left(\delta_{k,l} - \frac{w_l g_{i,l}}{Z_i}\right) \Sigma_l^{-1}(x_i - m_l). \tag{A.IV.40}$$

Eq. (A.IV.34) gives

$$\frac{\partial F_w}{\partial \mathbf{\Sigma}}(\theta, X) \in \mathcal{L}\left((S_d(\mathbb{R}))^K, \triangle_K^0\right) \hookrightarrow \mathbb{R}^{K \times K \times d \times d} : \tag{A.IV.41}$$

$$\left[\frac{\partial F_w}{\partial \mathbf{\Sigma}}(\theta, X)\right]_{k,l,\cdot,\cdot} = \frac{1}{n} \sum_{i=1}^n \frac{w_k g_{i,k}}{2 Z_i} \left(\delta_{k,l} - \frac{w_l g_{i,l}}{Z_i}\right) \left(-\Sigma_l^{-1} + \Sigma_l^{-1}(x_i - m_l)(x_i - m_l)^\top \Sigma_l^{-1}\right). \tag{A.IV.42}$$

Finally, with Eq. (A.IV.36), we obtain

$$\frac{\partial F_w}{\partial X}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{n \times d}, \triangle_K^0\right) \hookrightarrow \mathbb{R}^{K \times n \times d} : \tag{A.IV.43}$$

$$\left[\frac{\partial F_w}{\partial X}(\theta, X)\right]_{k,i,\cdot} = \frac{w_k g_{i,k}}{n Z_i} \left(-\Sigma_k^{-1}(x_i - m_k) + \frac{1}{Z_i} \sum_{l=1}^K w_l g_{i,l} \Sigma_l^{-1}(x_i - m_l)\right). \tag{A.IV.44}$$

Differentials of $F_{\mathbf{m}}$. We now turn to $F_{\mathbf{m}}$:

$$F_{\mathbf{m}} : \begin{cases} \Theta \times \mathcal{X} & \longrightarrow & \mathbb{R}^{K \times d} \\ (\theta, X) & \longmapsto & \left(\frac{\sum_{i=1}^n \gamma_{i,k}(\theta, X) x_i}{\sum_{j=1}^n \gamma_{j,k}(\theta, X)} \right)_{k=1}^K \end{cases}. \tag{A.IV.45}$$

For convenience we introduce $\Gamma_k := \sum_{i=1}^n \gamma_{i,k}$. The chain rule gives

$$\frac{\partial F_{\mathbf{m}}}{\partial w}(\theta, X) \in \mathcal{L}\left(\triangle_K^0, \mathbb{R}^{K\times d}\right) \hookrightarrow \mathbb{R}^{K\times d\times K} : \tag{A.IV.46}$$

$$\left[\frac{\partial F_{\mathbf{m}}}{\partial w}(\theta, X)\right]_{k,\cdot,l} = \frac{1}{\Gamma_k}\left(\sum_{i=1}^n \left[\frac{\partial \gamma}{\partial w}\right]_{i,k,l} x_i - \frac{1}{\Gamma_k}\sum_{j=1}^n \left[\frac{\partial \gamma}{\partial w}\right]_{j,k,l} \sum_{i=1}^n \gamma_{i,k} x_i\right). \tag{A.IV.47}$$

We continue with the differential with respect to $\mathbf{m}$:

$$\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{m}}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{K\times d}, \mathbb{R}^{K\times d}\right) \simeq \mathbb{R}^{K\times d\times K\times d} : \tag{A.IV.48}$$

$$\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{m}}(\theta, X)\right]_{k,\cdot,l,\cdot} = \frac{1}{\Gamma_k}\left(\sum_{i=1}^n x_i \left[\frac{\partial \gamma}{\partial \mathbf{m}}\right]^\top_{i,k,l,\cdot} - \frac{1}{\Gamma_k}\left(\sum_{i=1}^n \gamma_{i,k} x_i\right)\sum_{j=1}^n \left[\frac{\partial \gamma}{\partial \mathbf{m}}\right]^\top_{j,k,l,\cdot}\right). \tag{A.IV.49}$$

For the differential with respect to $\mathbf{\Sigma}$, we require the notion of *outer product* between two tensors, which we define below:

$$\forall A \in \mathbb{R}^{n_1\times\cdots\times n_N},\ B \in \mathbb{R}^{m_1\times\cdots\times m_M},\ A\otimes B := (A_{i_1,\cdots,i_N} B_{j_1,\cdots,j_M}) \in \mathbb{R}^{n_1\times\cdots\times n_N\times m_1\times\cdots\times m_M}. \tag{A.IV.50}$$

Note that in Eq. (A.IV.49), we could have written $x_i \otimes \left[\frac{\partial \gamma}{\partial \mathbf{m}}\right]_{i,k,l,\cdot}$. We will need the following differentiation rule for the outer product:

Lemma A.IV.1. Let $A : \mathbb{R}^{p_1\times\cdots\times p_P} \longrightarrow \mathbb{R}^{n_1\times\cdots\times n_N}$ and $B : \mathbb{R}^{k_1\times\cdots\times k_P} \longrightarrow \mathbb{R}^{m_1\times\cdots\times m_M}$ be differentiable. Then the differential of $C := A\otimes B$ is:

$$\frac{\partial C}{\partial X}(X) \in \mathcal{L}(\mathbb{R}^{p_1\times\cdots\times p_P}, \mathbb{R}^{n_1\times\cdots\times n_N\times m_1\times\cdots\times m_M}) \simeq \mathbb{R}^{n_1\times\cdots\times n_N\times m_1\times\cdots\times m_M\times p_1\times\cdots\times p_P} : \tag{A.IV.51}$$

$$\begin{aligned}\left[\frac{\partial C}{\partial X}(X)\right]_{i_1,\cdots,i_N,j_1,\cdots,j_M,k_1,\cdots,k_P} &= \left[\frac{\partial A}{\partial X}(X)\right]_{i_1,\cdots,i_N,k_1,\cdots,k_P} B_{j_1,\cdots,j_M}(X) \\ &\quad + A_{i_1,\cdots,i_N}\left[\frac{\partial B}{\partial X}(X)\right]_{j_1,\cdots,j_M,k_1,\cdots,k_P},\end{aligned} \tag{A.IV.52}$$

$$\frac{\partial C}{\partial X}(X) = \tau\left(\frac{\partial A}{\partial X}(X)\otimes B(X)\right) + A(X)\otimes\frac{\partial B}{\partial X}(X), \tag{A.IV.53}$$

where τ is the transposition $(\tau(T))_{i_1,\cdots,i_N,j_1,\cdots,j_m,k_1,\cdots,k_P} = T_{i_1,\cdots,i_N,k_1,\cdots,k_P,j_1,\cdots,j_M}$. If $B = A$, then the formula can be written as

$$\frac{\partial C}{\partial X}(X) = \tau\left(A(X)\otimes\frac{\partial A}{\partial X}(X)\right) + A(X)\otimes\frac{\partial A}{\partial X}(X). \tag{A.IV.54}$$

Note that in particular, the intuitive formula $\partial(A\otimes B) = (\partial A)\otimes B + A\otimes(\partial B)$ does not hold. We now compute the differential with respect to $\mathbf{\Sigma}$:

$$\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{\Sigma}}(\theta, X) \in \mathcal{L}\left((S_d(\mathbb{R}))^K, \mathbb{R}^{K\times d}\right) \hookrightarrow \mathbb{R}^{K\times d\times K\times d\times d} : \tag{A.IV.55}$$

$$\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{\Sigma}}(\theta, X)\right]_{k,\cdot,l,\cdot,\cdot} = \frac{1}{\Gamma_k}\left(\sum_{i=1}^n x_i \otimes \left[\frac{\partial \gamma}{\partial \mathbf{\Sigma}}\right]_{i,k,l,\cdot,\cdot} - \frac{1}{\Gamma_k}\left(\sum_{i=1}^n \gamma_{i,k} x_i\right)\otimes\sum_{j=1}^n \left[\frac{\partial \gamma}{\partial \mathbf{\Sigma}}\right]_{j,k,l,\cdot,\cdot}\right). \tag{A.IV.56}$$

The differential with respect to X is slightly different due to the product with x_i in Eq. (A.IV.45):

$$\frac{\partial F_{\mathbf{m}}}{\partial X}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{n\times d}, \mathbb{R}^{K\times d}\right) \simeq \mathbb{R}^{K\times d\times n\times d} : \tag{A.IV.57}$$

$$\left[\frac{\partial F_{\mathbf{m}}}{\partial X}(\theta, X)\right]_{k,\cdot,i,\cdot} = \frac{1}{\Gamma_k}\left(\gamma_{i,k} I_d + \sum_{h=1}^{n} x_h \left[\frac{\partial \gamma}{\partial X}\right]_{h,k,i,\cdot}^{\top} - \frac{1}{\Gamma_k}\left(\sum_{h=1}^{n}\gamma_{h,k}x_h\right)\sum_{j=1}^{n}\left[\frac{\partial \gamma}{\partial X}\right]_{j,k,i,\cdot}^{\top}\right). \tag{A.IV.58}$$

Differential of $F_{\mathbf{\Sigma}}$. We finish with the computation of the differentials of

$$F_{\mathbf{\Sigma}} : \begin{cases} \Theta \times \mathcal{X} & \longrightarrow & (S_d^{++}(\mathbb{R}))^K \\ (\theta, X) & \longmapsto & \left(\frac{\sum_{i=1}^n \gamma_{i,k}(\theta, X)(x_i - F_{\mathbf{m}}(\theta, X)_k)(x_i - F_{\mathbf{m}}(\theta, X)_k)^\top}{\sum_{j=1}^n \gamma_{j,k}(\theta, X)}\right)_{k=1}^{K} \end{cases}. \tag{A.IV.59}$$

For convenience, let $F_{m_k} := [F_{\mathbf{m}}(\theta, X)]_{k,\cdot}$. We first compute the differential with respect to w:

$$\frac{\partial F_{\mathbf{\Sigma}}}{\partial w}(\theta, X) \in \mathcal{L}\left(\triangle_K^0, (S_d(\mathbb{R}))^K\right) \hookrightarrow \mathbb{R}^{K\times d\times d\times K} : \tag{A.IV.60}$$

$$\begin{aligned}\left[\frac{\partial F_{\mathbf{\Sigma}}}{\partial w}(\theta, X)\right]_{k,\cdot,\cdot,l} &= \frac{1}{\Gamma_k}\sum_{i=1}^{n}\left[\frac{\partial \gamma}{\partial w}\right]_{i,k,l}(x_i - F_{m_k})(x_i - F_{m_k})^\top \\ &\quad - \frac{1}{\Gamma_k^2}\sum_{j=1}^{n}\left[\frac{\partial \gamma}{\partial w}\right]_{j,k,l}\sum_{i=1}^{n}\gamma_{i,k}(x_i - F_{m_k})(x_i - F_{m_k})^\top \\ &\quad - \frac{1}{\Gamma_k}\sum_{i=1}^{n}\gamma_{i,k}\left(\left[\frac{\partial F_{\mathbf{m}}}{\partial w}\right]_{k,\cdot,l}(x_i - F_{m_k})^\top + (x_i - F_{m_k})\left[\frac{\partial F_{\mathbf{m}}}{\partial w}\right]_{k,\cdot,l}^{\top}\right).\end{aligned} \tag{A.IV.61}$$

For the differential with respect to $\mathbf{m}$, we will use Lemma A.IV.1:

$$\frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{m}}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{K\times d}, (S_d(\mathbb{R}))^K\right) \hookrightarrow \mathbb{R}^{K\times d\times d\times K\times d} : \tag{A.IV.62}$$

$$\begin{aligned}\left[\frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{m}}(\theta, X)\right]_{k,\cdot,\cdot,l,\cdot} &= \frac{1}{\Gamma_k}\sum_{i=1}^{n}(x_i - F_{m_k})(x_i - F_{m_k})^\top \otimes \left[\frac{\partial \gamma}{\partial \mathbf{m}}\right]_{i,k,l,\cdot} \\ &\quad - \frac{1}{\Gamma_k^2}\left(\sum_{i=1}^{n}\gamma_{i,k}(x_i - F_{m_k})(x_i - F_{m_k})^\top\right)\otimes\sum_{j=1}^{n}\left[\frac{\partial \gamma}{\partial \mathbf{m}}\right]_{j,k,l,\cdot} \\ &\quad - \frac{1}{\Gamma_k}\sum_{i=1}^{n}\gamma_{i,k}\left\{\tau_{1,2}\left((x_i - F_{m_k})\otimes\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{m}}\right]_{k,\cdot,l,\cdot}\right) + (x_i - F_{m_k})\otimes\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{m}}\right]_{k,\cdot,l,\cdot}\right\}.\end{aligned} \tag{A.IV.63}$$

In Eq. (A.IV.63), the transposed term comes from the differential of the outer product Eq. (A.IV.54), where $\tau_{1,2}(A)_{i,j,\dots} = A_{j,i,\dots}$. For the differential with respect to $\mathbf{\Sigma}$, we use the same method as in Eq. (A.IV.63):

$$\frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{\Sigma}}(\theta, X) \in \mathcal{L}\left((S_d(\mathbb{R}))^K, (S_d(\mathbb{R}))^K\right) \hookrightarrow \mathbb{R}^{K\times d\times d\times K\times d\times d} : \tag{A.IV.64}$$

$$\begin{aligned}\left[\frac{\partial F_{\mathbf{\Sigma}}}{\partial \mathbf{\Sigma}}(\theta, X)\right]_{k,\cdot,\cdot,l,\cdot,\cdot} &= \frac{1}{\Gamma_k}\sum_{i=1}^{n}(x_i - F_{m_k})(x_i - F_{m_k})^\top \otimes \left[\frac{\partial \gamma}{\partial \mathbf{\Sigma}}\right]_{i,k,l,\cdot,\cdot} \\ &\quad - \frac{1}{\Gamma_k^2}\left(\sum_{i=1}^{n}\gamma_{i,k}(x_i - F_{m_k})(x_i - F_{m_k})^\top\right)\otimes\sum_{j=1}^{n}\left[\frac{\partial \gamma}{\partial \mathbf{\Sigma}}\right]_{j,k,l,\cdot,\cdot} \\ &\quad - \frac{1}{\Gamma_k}\sum_{i=1}^{n}\gamma_{i,k}\left\{\tau_{1,2}\left((x_i - F_{m_k})\otimes\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{\Sigma}}\right]_{k,\cdot,l,\cdot,\cdot}\right) + (x_i - F_{m_k})\otimes\left[\frac{\partial F_{\mathbf{m}}}{\partial \mathbf{\Sigma}}\right]_{k,\cdot,l,\cdot,\cdot}\right\}.\end{aligned} \tag{A.IV.65}$$

Finally, for the differential with respect to X, the method is almost identical, with an additional term due to the presence of x_i in $(x_i - F_{m_k})(x_i - F_{m_k})^\top$.

$$\frac{\partial F_{\mathbf{\Sigma}}}{\partial X}(\theta, X) \in \mathcal{L}\left(\mathbb{R}^{n\times d}, (S_d(\mathbb{R}))^K\right) \hookrightarrow \mathbb{R}^{K\times d\times d\times n\times d} : \tag{A.IV.66}$$

$$\begin{aligned}\left[\frac{\partial F_{\mathbf{\Sigma}}}{\partial X}(\theta, X)\right]_{k,\cdot,\cdot,i,\cdot} &= \frac{1}{\Gamma_k}\sum_{h=1}^n (x_h - F_{m_k})(x_h - F_{m_k})^\top \otimes \left[\frac{\partial \gamma}{\partial X}\right]_{h,k,i,\cdot} \\ &\quad - \frac{1}{\Gamma_k^2}\left(\sum_{h=1}^n \gamma_{h,k}(x_h - F_{m_k})(x_h - F_{m_k})^\top\right) \otimes \sum_{j=1}^n \left[\frac{\partial \gamma}{\partial X}\right]_{j,k,i,\cdot} \\ &\quad - \frac{1}{\Gamma_k}\sum_{h=1}^n \gamma_{h,k}\left\{\tau_{1,2}\left((x_h - F_{m_k}) \otimes \left[\frac{\partial F_{\mathbf{m}}}{\partial X}\right]_{k,\cdot,i,\cdot}\right) + (x_h - F_{m_k}) \otimes \left[\frac{\partial F_{\mathbf{m}}}{\partial X}\right]_{k,\cdot,i,\cdot}\right\} \\ &\quad + \frac{\gamma_{i,k}}{\Gamma_k}\left\{\tau_{1,2}\Big((x_i - F_{m_k}) \otimes I_d\Big) + (x_i - F_{m_k}) \otimes I_d\right\}. \end{aligned} \tag{A.IV.67}$$

A.IV.6.4 Local Minima in (GMM)-OT

A.IV.6.4.1 Computation of the Local Minima of the Discrete 2-Wasserstein Distance

We provide an explicit local minimum of the energy $\mathcal{E}_3$ defined in Eq. (A.IV.18). To compute the energy $\mathcal{E}_3(\alpha, \eta)$ at $[-\frac{1}{6}, \frac{1}{6}]^2 \times (-\frac{1}{2}, \frac{1}{2})^3$, we must distinguish the two cases $\alpha_1 + \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 < 0$, which yield two different OT plans between $\mu_{\alpha,\eta}$ and ν. We will show separately that the energy $\mathcal{E}_3$ has a strict local minimum at $(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3})$ in both sub-regions. To break the symmetry in $(\alpha_1, \alpha_2, \eta_1, \eta_2) \leftarrow (\alpha_2, \alpha_1, \eta_2, \eta_1)$, we will impose the constraint $\eta_1 \leq \eta_2$ in the optimisation problem. To summarise, we split the optimisation problem in two as follows:

$$\min_{\substack{\alpha\in[-\frac{1}{6},\frac{1}{6}]^2\\ \eta\in(-\frac{1}{2},\frac{1}{2})^3}} \mathcal{E}_3(\alpha,\eta) = \min(\mathcal{E}_3^+, \mathcal{E}_3^-), \quad \mathcal{E}_3^+ := \min_{\substack{\alpha\in[-\frac{1}{6},\frac{1}{6}]^2\\ \eta\in(-\frac{1}{2},\frac{1}{2})^3\\ \alpha_1+\alpha_2\geq 0\\ \eta_1\leq\eta_2}} \mathcal{E}_3(\alpha,\eta), \quad \mathcal{E}_3^- := \min_{\substack{\alpha\in[-\frac{1}{6},\frac{1}{6}]^2\\ \eta\in(-\frac{1}{2},\frac{1}{2})^3\\ \alpha_1+\alpha_2\leq 0\\ \eta_1\leq\eta_2}} \mathcal{E}_3(\alpha,\eta). \tag{A.IV.68}$$

Case $\alpha_1 + \alpha_2 \geq 0$. For any $(\alpha, \eta) \in [-\frac{1}{6}, \frac{1}{6}]^2 \times (-\frac{1}{2}, \frac{1}{2})^3$ such that $\alpha_1 + \alpha_2 \geq 0$ and $\eta_1 \leq \eta_2$, the optimal plan between $\mu_{\alpha,\eta}$ and ν is given by:

$$\pi^+(\alpha, \eta) = \begin{pmatrix} \frac{1}{6} + \alpha_1 & 0 & 0 \\ \frac{1}{6} - \alpha_1 & \alpha_1 + \alpha_2 & 0 \\ 0 & \frac{1}{3} - \alpha_1 - \alpha_2 & \frac{1}{3} \end{pmatrix},$$

and thus the energy $\mathcal{E}_3(\alpha, \eta)$ is given by:

$$\mathcal{E}_3(\alpha,\eta) = \left(\tfrac{1}{6} + \alpha_1\right)\eta_1^2 + \left(\tfrac{1}{6} - \alpha_1\right)\eta_2^2 + (\alpha_1 + \alpha_2)\,(1 - \varepsilon - \eta_2)^2 + \left(\tfrac{1}{3} - \alpha_1 - \alpha_2\right)(\eta_3 + \varepsilon)^2 + \tfrac{1}{3}(\eta_3 - \varepsilon)^2. \tag{A.IV.69}$$

Taking dual variables $\lambda_1, \lambda_2 \in \mathbb{R}$, the Lagrangian of the problem $\mathcal{E}_3^+$ defined in Eq. (A.IV.68) is given by:

$$\mathcal{L}^+(\alpha, \eta, \lambda_1, \lambda_2) = \mathcal{E}_3(\alpha, \eta) + \lambda_1(-\alpha_1 - \alpha_2) + \lambda_2(\eta_1 - \eta_2). \tag{A.IV.70}$$

To find solutions of the problem $\mathcal{E}_3^+$, we study the necessary KKT system (see [BV04, Section 5.5.3], or [NW06, Theorem 12.1]). We begin with the stationarity condition, expressed at the

point $(\alpha,\eta)\in[-\frac{1}{6},\frac{1}{6}]^2\times(-\frac{1}{2},\frac{1}{2})^3$:

$$\begin{aligned}
0&=\frac{\partial\mathcal{L}^+}{\partial\alpha_1}=\eta_1^2-\eta_2^2+(1-\varepsilon-\eta_2)^2-(\eta_3+\varepsilon)^2-\lambda_1;\\
0&=\frac{\partial\mathcal{L}^+}{\partial\alpha_2}=(1-\varepsilon-\eta_2)^2-(\eta_3+\varepsilon)^2-\lambda_1;\\
0&=\frac{\partial\mathcal{L}^+}{\partial\eta_1}=2\left(\tfrac{1}{6}+\alpha_1\right)\eta_1+\lambda_2;\\
0&=\frac{\partial\mathcal{L}^+}{\partial\eta_2}=2\left(\tfrac{1}{6}-\alpha_1\right)\eta_2-2(\alpha_1+\alpha_2)(1-\varepsilon-\eta_2)-\lambda_2;\\
0&=\frac{\partial\mathcal{L}^+}{\partial\eta_3}=2\left(\tfrac{1}{3}-\alpha_1-\alpha_2\right)(\eta_3+\varepsilon)+\tfrac{2}{3}(\eta_3-\varepsilon).
\end{aligned}$$

The remaining KKT conditions are the primal / dual feasibility conditions, which are given by:

$$\alpha_1+\alpha_2\geq 0,\ \eta_1\leq\eta_2,\ \lambda_1\geq 0,\ \lambda_2\geq 0,$$

and the complementary slackness conditions:

$$\lambda_1(-\alpha_1-\alpha_2)=0,\ \lambda_2(\eta_1-\eta_2)=0.$$

It is easy to see that the point $(\alpha,\eta,\lambda_1,\lambda_2)=(0_{\mathbb{R}^2},0_{\mathbb{R}^3},1-2\varepsilon,0)$ is a solution of the KKT system (note that $\varepsilon<\frac{1}{2}$).

We now check that our critical point is a local minimum. For $\alpha\in[-\frac{1}{6},\frac{1}{6}]^2$, $\eta\in(-\frac{1}{2},\frac{1}{2})^3$ verifying $\alpha_1+\alpha_2\geq 0$, we compute the Hessian of $\mathcal{E}_3$:

$$H^+(\alpha,\eta)=\begin{pmatrix}
0 & 0 & 2\eta_1 & 2\varepsilon-2 & -2\eta_3-2\varepsilon\\
0 & 0 & 0 & 2\eta_2+2\varepsilon-2 & -2\eta_3-2\varepsilon\\
2\eta_1 & 0 & 2\alpha_1+\frac{1}{3} & 0 & 0\\
2\varepsilon-2 & 2\eta_2+2\varepsilon-2 & 0 & 2\alpha_2+\frac{1}{3} & 0\\
-2\eta_3-2\varepsilon & -2\eta_3-2\varepsilon & 0 & 0 & -2\alpha_1-2\alpha_2+\frac{4}{3}
\end{pmatrix}. \tag{A.IV.71}$$

Using numerical solvers, we obtain that at $(\alpha,\eta)=(0_{\mathbb{R}^2},0_{\mathbb{R}^3})$, the Hessian verifies the following property:

$$H^+(0_{\mathbb{R}^2},0_{\mathbb{R}^3})v=0,\ v:=(1,-1,0,0,0),\ \forall w\in v^\perp,\ w^\top H^+(0_{\mathbb{R}^2},0_{\mathbb{R}^3})w>0. \tag{A.IV.72}$$

Adding the vector tv to (α,η) for $|t|$ small enough corresponds to adding t to α_1 and subtracting t from α_2, and we notice that at $(\alpha,\eta)=(0_{\mathbb{R}^2},0_{\mathbb{R}^3})$, since $\eta_1=\eta_2$, this operation does not change the cost: $\mathcal{E}_3((t,-t),0_{\mathbb{R}^3})=\mathcal{E}_3(0_{\mathbb{R}^2},0_{\mathbb{R}^3})$. We conclude that $(0_{\mathbb{R}^2},0_{\mathbb{R}^3})$ is a local minimum for the problem $\mathcal{E}_3^+$ defined in Eq. (A.IV.68).

Case $\alpha_1+\alpha_2\leq 0$. For any $(\alpha,\eta)\in[-\frac{1}{6},\frac{1}{6}]^2\times(-\frac{1}{2},\frac{1}{2})^3$ such that $\alpha_1+\alpha_2\leq 0$ and $\eta_1\leq\eta_2$, the optimal plan between $\mu_{\alpha,\eta}$ and ν and the associated energy $\mathcal{E}_3$ are given by:

$$\pi^-(\alpha,\eta)=\begin{pmatrix}
\frac{1}{6}+\alpha_1 & 0 & 0\\
\frac{1}{6}+\alpha_2 & 0 & 0\\
-\alpha_1-\alpha_2 & \frac{1}{3} & \frac{1}{3}
\end{pmatrix},$$

$$\mathcal{E}_3(\alpha,\eta)=\left(\tfrac{1}{6}+\alpha_1\right)\eta_1^2+\left(\tfrac{1}{6}+\alpha_2\right)\eta_2^2-(\alpha_1+\alpha_2)\,(1+\eta_3)^2+\tfrac{1}{3}\,(\eta_3+\varepsilon)^2+\tfrac{1}{3}(\eta_3-\varepsilon)^2.$$

We deduce the expression of the Lagrangian of the problem $\mathcal{E}_3^-$ defined in Eq. (A.IV.68):

$$\mathcal{L}^-(\alpha,\eta,\lambda_1,\lambda_2)=\mathcal{E}_3(\alpha,\eta)+\lambda_1(\alpha_1+\alpha_2)+\lambda_2(\eta_1-\eta_2). \tag{A.IV.73}$$

The KKT stationarity condition writes then:

$$
\begin{aligned}
0 &= \frac{\partial \mathcal{L}^-}{\partial \alpha_1} = \eta_1^2 - (1+\eta_3)^2 + \lambda_1; \\
0 &= \frac{\partial \mathcal{L}^-}{\partial \alpha_2} = \eta_2^2 - (1+\eta_3)^2 + \lambda_1; \\
0 &= \frac{\partial \mathcal{L}^-}{\partial \eta_1} = 2\left(\tfrac{1}{6} + \alpha_1\right)\eta_1 + \lambda_2; \\
0 &= \frac{\partial \mathcal{L}^-}{\partial \eta_2} = 2\left(\tfrac{1}{6} - \alpha_1\right)\eta_2 - \lambda_2; \\
0 &= \frac{\partial \mathcal{L}^-}{\partial \eta_3} = -2\left(\alpha_1 + \alpha_2\right)(1+\eta_3) + \tfrac{4}{3}\eta_3.
\end{aligned}
$$

Again the primal / dual feasibility conditions read:

$$
\alpha_1 + \alpha_2 \leq 0,\ \eta_1 \leq \eta_2,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0,
$$

and the complementary slackness conditions:

$$
\lambda_1(\alpha_1 + \alpha_2) = 0,\ \lambda_2(\eta_1 - \eta_2) = 0.
$$

Likewise, the point $(\alpha, \eta, \lambda_1, \lambda_2) = (0_{\mathbb{R}^2}, 0_{\mathbb{R}^3}, 1, 0)$ is a solution of the KKT system.

To prove local minimality, we compute the Hessian of $\mathcal{E}_3$ at $\alpha \in [-\frac{1}{6}, \frac{1}{6}]^2$, $\eta \in (-\frac{1}{2}, \frac{1}{2})^3$ verifying $\alpha_1 + \alpha_2 \leq 0$:

$$
H^-(\alpha, \eta) = \begin{pmatrix}
0 & 0 & 2\eta_1 & 0 & -2\eta_3 - 2 \\
0 & 0 & 0 & 2\eta_2 & -2\eta_3 - 2 \\
2\eta_1 & 0 & 2\alpha_1 + \frac{1}{3} & 0 & 0 \\
0 & 2\eta_2 & 0 & 2\alpha_2 + \frac{1}{3} & 0 \\
-2\eta_3 - 2 & -2\eta_3 - 2 & 0 & 0 & -2\alpha_1 - 2\alpha_2 + \frac{4}{3}
\end{pmatrix}.
$$

The property of $H^+(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3})$ from Eq. (A.IV.72) is also satisfied by $H^-(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3})$, and we conclude likewise that $(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3})$ is a local minimum for the problem $\mathcal{E}_3^-$ defined in Eq. (A.IV.68).

We conclude that $(\alpha, \eta) = (0_{\mathbb{R}^2}, 0_{\mathbb{R}^3})$ is a minimum of $\mathcal{E}_3$ on the set $[-\frac{1}{6}, \frac{1}{6}]^2 \times (-\frac{1}{2}, \frac{1}{2})^3$, with value $\mathcal{E}_3(0_{\mathbb{R}^2}, 0_{\mathbb{R}^3}) = \frac{2}{3}\varepsilon^2 > 0$ (use Eq. (A.IV.69) for example).

A.IV.6.4.2 Discrete 2-Wasserstein Distance for $n = 2$: Proof of Unique Local Minimum

Unique Local Minimum of the Discrete 2-Wasserstein Distance for $n = 2$. We consider the minimisation of the quadratic Wasserstein cost $\mu \longmapsto \mathrm{W}_2^2(\mu, \nu)$ when μ is restricted to two Dirac atoms,

$$
\mu_{x,\alpha} = \alpha\delta_{x_1} + (1-\alpha)\delta_{x_2}, \qquad (x, \alpha) \in \mathbb{R}^{2d} \times [0,1],
$$

against the fixed target $\nu = \gamma\delta_{y_1} + (1-\gamma)\delta_{y_2}$ with $\gamma \in (0,1)$ and $y_1 \neq y_2$. Denoting

$$
\mathcal{E}_2(x, \alpha) := \mathrm{W}_2^2(\mu_{x,\alpha}, \nu), \tag{A.IV.74}
$$

we show that inside the domain $\mathcal{D} := \{(x, \alpha) : x_1 \neq x_2,\ \alpha \in (0,1)\}$ every local minimiser of $\mathcal{E}_2$ is such that $\mu_{x,\alpha} = \nu$, and is thus a global minimiser. We show this result more generally, for costs $c(x, y) := \phi(x - y)$ where ϕ is convex, $\mathcal{C}^1$, $\phi(v) = 0 \iff v = 0$ and $\nabla\phi(v) = 0 \iff v = 0$. Additionally, boundary stationary points can occur when atoms merge, i.e. $\alpha \in \{0, 1\}$ or $x_1 = x_2$; they are local but not global minima. When these weights arise from EM, we can prevent such degeneracies by fixing weights or imposing a hard lower bound on them throughout the iterations.

We now determine all the local minimisers of the energy $\mathcal{E}_2$ defined in Eq. (A.IV.74) in the domain $\mathcal{D} := \{(x, \alpha) : x_1 \neq x_2,\ \alpha \in (0,1)\}$. Fix $\gamma \in (0,1)$, and let x_1, x_2, y_1 and y_2 in $\mathbb{R}^d$, with $y_1 \neq y_2$. For $\alpha \in (0,1)$, we consider the measures:

$$
\mu(x_1, x_2, \alpha) := \alpha\delta_{x_1} + (1-\alpha)\delta_{x_2}, \quad \nu := \gamma\delta_{y_1} + (1-\gamma)\delta_{y_2}.
$$

Consider the ground cost $c := (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \longmapsto \phi(x - y)$, with:

- (H1) $\phi : \mathbb{R}^d \to \mathbb{R}_+$ convex and $\mathcal{C}^1$,
- (H2) $\phi(v) = 0 \iff v = 0$,
- (H3) $\nabla\phi(v) = 0 \iff v = 0$.

For instance, ϕ could be $\|\cdot\|_p^p$ with $p > 1$. Every feasible coupling between $\mu(x_1, x_2, \alpha)$ and ν can be written under the form:

$$\pi(\alpha, t) := \begin{pmatrix} \alpha - t & t \\ \gamma - \alpha + t & 1 - \gamma - t \end{pmatrix}, \ t \in [t_-(\alpha), t_+(\alpha)], \ t_-(\alpha) := \max(0, \alpha - \gamma), \ t_+(\alpha) := \min(\alpha, 1 - \gamma).$$

The interval $[t_-(\alpha), t_+(\alpha)]$ is non-empty for all $\alpha \in (0, 1)$, since $\gamma \in (0, 1)$. The transport cost associated to a plan $\pi(\alpha, t)$ writes:

$$\mathcal{F}(x, \alpha, t) := (\alpha - t)c(x_1, y_2) + (\gamma - \alpha + t)c(x_2, y_2) + tc(x_1, y_2) + (1 - \gamma - t)c(x_2, y_2). \quad \text{(A.IV.75)}$$

Since $\mathcal{E}_2(x_1, x_2, \alpha) = \min_{t \in [t_-(\alpha), t_+(\alpha)]} \mathcal{F}(x_1, x_2, \alpha, t)$ and that $\mathcal{F}$ is linear in t, we obtain:

$$\begin{aligned} \mathcal{E}_2(x_1, x_2, \alpha) &= \min\left(\mathcal{E}^-(x_1, x_2, \alpha), \mathcal{E}^+(x_1, x_2, \alpha)\right), \\ \mathcal{E}^-(x_1, x_2, \alpha) := \mathcal{F}(x_1, x_2, \alpha, t_-(\alpha)), \ \mathcal{E}^+(x_1, x_2, \alpha) &:= \mathcal{F}(x_1, x_2, \alpha, t_+(\alpha)). \end{aligned} \quad \text{(A.IV.76)}$$

In particular, any local optimum of $\mathcal{E}_2$ in $\mathcal{D} := \{(x_1, x_2, \alpha) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, 1) : \ x_1 \neq x_2\}$ must be a local optimum of $\mathcal{E}^-$ or $\mathcal{E}^+$. Straightforward computation yields the following symmetrical expression:

$$\forall x_1, x_2 \in \mathbb{R}^d, \ \forall \alpha \in (0, 1), \ \forall t \in \mathbb{R}, \ \mathcal{F}(x_1, x_2, \alpha, t) = \mathcal{F}(x_2, x_1, 1 - \alpha, 1 - \gamma - t),$$

which can be understood as exchanging the roles of x_1 and x_2. For any $\alpha \in (0, 1)$, we have $1 - \gamma - t_-(\alpha) = t_+(\alpha)$, concluding a symmetrical relationship between $\mathcal{E}^-$ and $\mathcal{E}^+$:

$$\forall x_1, x_2 \in \mathbb{R}^d, \ \forall \alpha \in (0, 1), \ \mathcal{E}^+(x_2, x_1, 1 - \alpha) = \mathcal{E}^-(x_1, x_2, \alpha). \quad \text{(A.IV.77)}$$

As a consequence, any local optimum $(x_1, x_2, \alpha) \in \mathcal{D}$ of $\mathcal{E}^+$ is such that $(x_2, x_1, 1 - \alpha)$ is a local optimum of $\mathcal{E}^-$, and conversely. To determine the local minima of $\mathcal{E}_2$ in $\mathcal{D}$, we can focus on the local minima of $\mathcal{E}^-$ in $\mathcal{D}$. We split $\mathcal{D}$ intro three sub-regions wherein $\mathcal{E}^-$ has an explicit expression: consider

$$\mathcal{R}_< := \{0 < \alpha < \gamma\} \cap \mathcal{D}, \quad \mathcal{R}_= := \{\alpha = \gamma\} \cap \mathcal{D}, \quad \mathcal{R}_> := \{\gamma < \alpha < 1\} \cap \mathcal{D},$$

we have the following expressions for $\mathcal{E}^-$ at $(x_1, x_2, \alpha) \in \mathcal{D}$:

$$\mathcal{E}^-(x_1, x_2, \alpha) = \begin{cases} \alpha c(x_1, y_1) + (\gamma - \alpha)c(x_2, y_1) + (1 - \gamma)c(x_2, y_2) & \text{if } (x_1, x_2, \alpha) \in \mathcal{R}_< \cup \mathcal{R}_= \\ \gamma c(x_1, y_1) + (\alpha - \gamma)c(x_1, y_2) + (1 - \alpha)c(x_2, y_2) & \text{if } (x_1, x_2, \alpha) \in \mathcal{R}_= \cup \mathcal{R}_> \end{cases}. \quad \text{(A.IV.78)}$$

No local minimum in $\mathcal{R}_<$. For $(x_1, x_2, \alpha) \in \mathcal{R}_<$, consider

$$\mathcal{E}^-_<(x_1, x_2, \alpha) := \alpha c(x_1, y_1) + (\gamma - \alpha)c(x_2, y_1) + (1 - \gamma)c(x_2, y_2).$$

A local optimum $(x_1, x_2, \alpha) \in \mathcal{R}_<$ of $\mathcal{E}^-$ must satisfy the stationarity conditions:

$$0 = \frac{\partial \mathcal{E}^-_<}{\partial x_1} = \alpha \nabla\phi(x_1 - y_1) \quad \underset{\text{(H3)}}{\Longrightarrow} \quad x_1 = y_1,$$

$$0 = \frac{\partial \mathcal{E}^-_<}{\partial \alpha} = \underbrace{c(x_1, y_1)}_{=0 \text{ by (H2)}} - c(x_2, y_1) \quad \underset{\text{(H2)}}{\Longrightarrow} \quad x_2 = y_1.$$

$$0 = \frac{\partial \mathcal{E}^-_<}{\partial x_2} = (\gamma - \alpha)\underbrace{\nabla\phi(x_2 - y_1)}_{=0 \text{ by (H2)}} + (1 - \gamma)\nabla\phi(\underbrace{x_2}_{=y_1} - y_2) = (1 - \gamma)\nabla\phi(y_1 - y_2).$$

By (H3), since $y_1 \neq y_2$, we have $\nabla c(y_1, y_2) \neq 0$, hence $\mathcal{E}_-$ has no local minimum in $\mathcal{R}_<$.

No local minimum in $\mathcal{R}_>$. For $(x_1, x_2, \alpha) \in \mathcal{R}_>$, consider

$$\mathcal{E}^-_>(x_1, x_2, \alpha) := \gamma c(x_1, y_1) + (\alpha - \gamma)c(x_1, y_2) + (1 - \alpha)c(x_2, y_2).$$

Likewise, a local minimum $(x_1, x_2, \alpha) \in \mathcal{R}_>$ of $\mathcal{E}^-$ must verify:

$$0 = \frac{\partial \mathcal{E}^-_>}{\partial x_1} = (1 - \alpha)\nabla\phi(x_2 - y_2) \qquad \underset{\text{(H3)}}{\Longrightarrow} \qquad x_2 = y_2,$$

$$0 = \frac{\partial \mathcal{E}^-_>}{\partial \alpha} = c(x_1, y_2) - \underbrace{c(x_2, y_2)}_{=0 \text{ by (H2)}} \qquad \underset{\text{(H2)}}{\Longrightarrow} \qquad x_1 = y_2.$$

$$0 = \frac{\partial \mathcal{E}^-_>}{\partial x_2} = \gamma\nabla\phi(\underbrace{x_1}_{=y_2} - y_1) + (\alpha - \gamma)\underbrace{\nabla\phi(x_2 - y_2)}_{=0 \text{ by (H2)}} = \gamma\nabla\phi(y_2 - y_1) \neq 0.$$

As before, this system cannot hold, and thus $\mathcal{E}_-$ has no local minimum in $\mathcal{R}_>$.

Analysis on $\mathcal{R}_=$. For $(x_1, x_2, \alpha) \in \mathcal{R}_=$, we have:

$$\mathcal{E}^-(x_1, x_2, \alpha) = \mathcal{E}^-_=(x_1, x_2) := \gamma c(x_1, y_1) + (1 - \gamma)c(x_2, y_2),$$

thus a local minimum $(x_1, x_2, \alpha) \in \mathcal{R}_=$ of $\mathcal{E}^-$ must satisfy the stationarity conditions:

$$0 = \frac{\partial \mathcal{E}^-_=}{\partial x_1} = \gamma\nabla\phi(x_1 - y_1) \qquad \underset{\text{(H3)}}{\Longrightarrow} \qquad x_1 = y_1,$$

$$0 = \frac{\partial \mathcal{E}^-_=}{\partial x_2} = (1 - \gamma)\nabla\phi(x_2 - y_2) \qquad \underset{\text{(H3)}}{\Longrightarrow} \qquad x_2 = y_2,$$

Since $(x_1, x_2, \alpha) = (y_1, y_2, \gamma)$ is a global minimum of $\mathcal{E}^-_=$, we conclude that the only local minimum of $\mathcal{E}^-$ in $\mathcal{D}$ is (y_1, y_2, γ).

Using Eq. (A.IV.77), we deduce that the only local minimum of $\mathcal{E}^+$ in $\mathcal{D}$ is $(y_2, y_1, 1 - \gamma)$, which is a global minimum of $\mathcal{E}_2$ in $\mathcal{D}$. Finally, there are two local minima of $\mathcal{E}_2$ in $\mathcal{D}$, which are the two global minima corresponding to $\mu(x_1, x_2, \alpha) = \nu$.

Local minimum for a single-Dirac source, and how to avoid it. The constraint $(x_1, x_2, \alpha) \in \mathcal{D}$ imposes in particular that $\mu(x_1, x_2, \alpha)$ is composed of two distinct Dirac masses. We now focus on the pathological case where $\alpha = 0$, showing the existence of a local optimum. For simplicity, we consider $d = 1$ and the cost $c(x, y) := |x - y|^2$. Set

$$z^\star := (x_1^\star, x_2^\star, \alpha^\star) := (-1, 1 - \gamma, 0),\ y_1 := 0,\ y_2 := 1,\ \gamma \in (0, 1).$$

For (x_1, x_2, α) in an open vicinity of $z^\star$, the cost simplifies to

$$\mathcal{E}_2(x_1, x_2, \alpha) = \alpha x_1^2 + (\gamma - \alpha)x_2^2 + (1 - \gamma)(x_2 - 1)^2.$$

To show that $z^\star$ is a local minimum, we compute the gradient of $\mathcal{E}_2$ at $z^\star$:

$$\frac{\partial \mathcal{E}_2}{\partial x_1}(z^\star) = 2\alpha^\star x_1^\star = 0,$$

$$\frac{\partial \mathcal{E}_2}{\partial x_2}(z^\star) = 2\left(x_2^\star - (1 - \gamma)\right) = 0,$$

$$\frac{\partial \mathcal{E}_2}{\partial \alpha}(z^\star) = x_1^2 - x_2^2 = 1 - (1 - \gamma)^2 > 0.$$

Consider a perturbation $h := (h_1, h_2, h_3) \in \mathbb{R}^3$ with $h_3 > 0$ (as $\alpha \geq 0$). For $z := z^\star + h$, we have:

$$\mathcal{E}_2(z) = \mathcal{E}_2(z^\star) + \underbrace{\frac{\partial \mathcal{E}_2}{\partial \alpha}(z^\star)h_3}_{>0} + \mathcal{O}(\|h\|^2),$$

so every feasible perturbation increases the objective. Thus $z^\star$ is a strict local minimiser, but is not global. When α arises as a mixture weight in EM, two remedies are possible: fixing uniform weights, or enforcing $\alpha \geq \varepsilon > 0$: both methods sidestep the degenerate local minimum above.

A.IV.6.4.3 Essential Stationary Points for the MW2-EM Loss: Computations

We illustrate a phenomenon of points where the gradients are infinitesimally small on a simple example with two Gaussian components, and studying a variant of the EM algorithm that does not update the covariances for simplicity. We fix $\varepsilon > 0$ and the following parametrised input GMM:

$$\forall \alpha \in (0,1),\ \forall m_1, m_2 \in \mathbb{R},\ \mu(\alpha, m_1, m_2) := \alpha\mathcal{N}(m_1, \varepsilon^2) + (1-\alpha)\mathcal{N}(m_2, \varepsilon^2),$$

and for $w := \frac{2}{3}$, the target GMM: $\nu := w\mathcal{N}(0, \varepsilon^2) + (1-w)\mathcal{N}(1, \varepsilon^2)$. We consider the particular dataset $X_\varepsilon \in \mathbb{R}^{6\times 1}$ defined by:

$$X_\varepsilon := (x_1, \cdots, x_6),\ x_1 := -\varepsilon,\ x_2 := \varepsilon,\ x_3 := x_5 := m^\star - \varepsilon,\ x_4 := x_6 := m^\star + \varepsilon.$$

We define the GMM $\mu^\star := (1-w)\mathcal{N}(0, \varepsilon^2) + w\mathcal{N}(m^\star, \varepsilon^2)$ associated to the vanishing gradient point, and introduce $\theta^\star := ((1-w), 0, m^\star)$ its parameters. For any $X \in \mathbb{R}^{6\times 1}$, we denote by $\theta(X)$ the parameters of the GMM fitted by the EM algorithm in one iteration on X with initialisation $\theta^\star$: $\theta(X) := F(\theta^\star, X)$ (we remind that in this section, we consider a simplified EM that does not update covariances). We shall first see that $\theta(X_\varepsilon) \approx \theta^\star$, then that the energy

$$\mathcal{E}_{\mathrm{EM-MW}_2^2} := X \longmapsto \mathrm{MW}_2^2(\mu(\theta(X)), \nu), \quad \mu(\theta(X)) := \alpha(X)\mathcal{N}(m_1(X), \varepsilon^2) + (1-\alpha(X))\mathcal{N}(m_2(X), \varepsilon^2),$$

verifies $\partial_X \mathcal{E}_{\mathrm{EM-MW}_2^2}(X_\varepsilon) \approx 0$. We summarise our setting in Fig. A.IV.15.

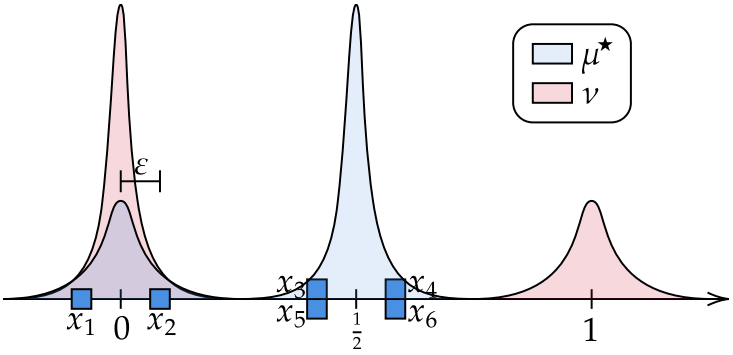


Figure A.IV.15: With the data $X_\varepsilon := (x_1, \cdots, x_6)$, one iteration of the EM algorithm initialised at $\theta^\star$ (corresponding to the GMM $\mu^\star$) yields approximately the same parameters $\theta^\star$. We shall see that the gradient of the energy $\mathcal{E}_{\mathrm{EM-MW}_2^2}$ at X_ε is approximately zero, illustrating the vanishing gradient phenomenon.

Showing that $\theta(X_\varepsilon) \approx \theta^\star$. Using the expressions of the EM update from Eq. (A.IV.3), we obtain that responsibilities $\gamma(X_\varepsilon)$ computed with initialisation $\theta^\star$ verify:

$$\forall i \in \{1,2\},\ \gamma_{i,1}(X_\varepsilon) = 1 + \mathcal{O}(e^{-1/\varepsilon^2}),\ \forall i \in \{3,4,5,6\},\ \gamma_{i,2}(X_\varepsilon) = 1 + \mathcal{O}(e^{-1/\varepsilon^2}),$$

as $\varepsilon \longrightarrow 0^+$. This means that the two points x_1, x_2 are considered as belonging to the first component $\mathcal{N}(0, \varepsilon^2)$ of $\mu^\star$, and the four points x_3, x_4, x_5, x_6 to the second component $\mathcal{N}(m^\star, \varepsilon^2)$ of $\mu^\star$. As for the weight $\alpha(X_\varepsilon)$ and means $m(X_\varepsilon)$ of the GMM $F(\theta^\star, X^\star)$, we use Eq. (A.IV.4) to deduce that:

$$\alpha(X_\varepsilon) = (1-w) + \mathcal{O}(e^{-1/\varepsilon^2}),\ m_1(X_\varepsilon) = 0 + \mathcal{O}(e^{-1/\varepsilon^2}),\ m_2(X_\varepsilon) = m^\star + \mathcal{O}(e^{-1/\varepsilon^2}). \quad \text{(A.IV.79)}$$

In this sense, we have $\theta(X_\varepsilon) \approx \theta^\star$ as $\varepsilon \longrightarrow 0^+$.

Vanishing gradient of the energy $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ at X_ε. For $\varepsilon > 0$ sufficiently small and X in a sufficiently small open vicinity of X_ε, it follows by regularity of F that $\theta(X) = (\alpha(X), m_1(X), m_2(X))$ will be sufficiently close to $\theta(X_\varepsilon)$ to ensure that $\alpha(X) < w$ and that $m_1(X) < m_2(X)$, which yields (by property on one-dimensional OT) the following expression for the energy:

$$\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}(X) := \mathrm{MW}_2^2(\mu(\theta(X)), \nu) = \mathcal{F}(\alpha(X), m_1(X), m_2(X)),$$

where $\mathcal{F} : (0,1) \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ is the function defined by:

$$\forall (\alpha, m_1, m_2) \in (0,1) \times \mathbb{R} \times \mathbb{R},\ \mathcal{F}(\alpha, m_1, m_2) := \alpha m_1^2 + (w - \alpha) m_2^2 + (1 - w)(m_2 - 1)^2.$$

To determine the gradient of $\mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}$ at X_ε, we use the chain rule:

$$\nabla \mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}(X_\varepsilon) = \frac{\partial \mathcal{F}}{\partial \alpha}(\theta(X_\varepsilon)) \frac{\partial \alpha}{\partial X}(X_\varepsilon) + \frac{\partial \mathcal{F}}{\partial m_1}(\theta(X_\varepsilon)) \frac{\partial m_1}{\partial X}(X_\varepsilon) + \frac{\partial \mathcal{F}}{\partial m_2}(\theta(X_\varepsilon)) \frac{\partial m_2}{\partial X}(X_\varepsilon).$$

Differentiating $\mathcal{F}$ and evaluating at $\theta(X_\varepsilon)$ which verifies the properties given in Eq. (A.IV.79) yields:

$$\frac{\partial \mathcal{F}}{\partial \alpha}(\theta(X_\varepsilon)) = -m^\star + \mathcal{O}(e^{-1/\varepsilon^2}) = \mathcal{O}(1),\ \frac{\partial \mathcal{F}}{\partial m_1}(\theta(X_\varepsilon)) = \mathcal{O}(e^{-1/\varepsilon^2}),\ \frac{\partial \mathcal{F}}{\partial m_2}(\theta(X_\varepsilon)) = \mathcal{O}(e^{-1/\varepsilon^2}).$$

Using the expression of $\frac{\partial \alpha}{\partial X}$ from Eq. (A.IV.44) with $X := X_\varepsilon$ and $\theta := \theta^\star$ yields $\frac{\partial \alpha}{\partial X}(X_\varepsilon) = \mathcal{O}(\varepsilon^{-2} e^{-1/\varepsilon^2})$.

With the expressions of $\frac{\partial m_1}{\partial X}$ and $\frac{\partial m_2}{\partial X}$ computed in Eq. (A.IV.58), we obtain $\frac{\partial m_1}{\partial X}(X_\varepsilon) = \mathcal{O}(1)$ and $\frac{\partial m_2}{\partial X}(X_\varepsilon) = \mathcal{O}(1)$. Putting everything together, we conclude that the gradient vanishes:

$$\nabla \mathcal{E}_{\mathrm{EM}-\mathrm{MW}_2^2}(X_\varepsilon) = \mathcal{O}(\varepsilon^{-2} e^{-1/\varepsilon^2}).$$

A.IV.6.5 Differentiating the Matrix Square Root

We consider the (differentiable) matrix square root function:

$$R := \left\{ \begin{array}{ccc} S_d^{++}(\mathbb{R}) & \longrightarrow & S_d^{++}(\mathbb{R}) \\ A & \longmapsto & \sqrt{A} \end{array} \right. ,$$

and provide a formula for its differential which is useful for numerical automatic differentiation.

Proposition A.IV.3. Let $A \in S_d^{++}(\mathbb{R})$ and $H \in S_d(\mathbb{R})$. Then the differential of the matrix square root at A in the direction H is given by the following matrix in $S_d(\mathbb{R})$:

$$\mathrm{d}_A R(H) = P G P^\top, \quad [G]_{i,j} := \frac{[P^\top H P]_{i,j}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}},$$

where the orthonormal decomposition of A is given by $A = P \operatorname{diag}(\lambda_1, \cdots, \lambda_d) P^\top$.

Proof. For any $A \in S_d^{++}(\mathbb{R})$, we have by definition $R(A)R(A) = A$, and differentiating this identity at A in the direction H yields that $\mathrm{d}_A R(H)$ is a solution of the following Sylvester equation:

$$R(A)X + XR(A) = H. \tag{A.IV.80}$$

By [Bha13, Theorem VII.2.1], Eq. (A.IV.80) has a unique solution in $\mathbb{R}^{d \times d}$, which is therefore $\mathrm{d}_A R(H)$. Using the notation of the result statement, we consider the symmetric matrix $X := PGP^\top$ and notice that by definition of G, we have:

$$\forall i, j \in [\![1, d]\!],\ \sqrt{\lambda_i} G_{i,j} + G_{i,j} \sqrt{\lambda_j} = [P^\top H P]_{i,j},$$

introducing $D := \operatorname{diag}(\lambda_1, \cdots, \lambda_d)$, we deduce that $R(D)G + GR(D) = P^\top H P$ and then that X is indeed a solution of Eq. (A.IV.80), hence by uniqueness $X = \mathrm{d}_A R(H)$. □

Below, we provide a PyTorch [Pas+19] implementation of this gradient computation, allowing for automatic gradient propagation.

```
import torch

class MatrixSquareRoot(torch.autograd.Function):
    @staticmethod
    def forward(ctx, A):
        A_sym = .5 * (A + A.transpose(-2, -1))
        L, P = torch.linalg.eigh(A_sym)
        S = L.clamp_min(0).sqrt()
        R = (P * S.unsqueeze(-2)) @ P.transpose(-2,-1)
        ctx.save_for_backward(P, S)
        return R

    @staticmethod
    def backward(ctx, H):
        P, S = ctx.saved_tensors
        H_sym = .5 * (H + H.transpose(-2, -1))
        D = S.unsqueeze(-1) + S.unsqueeze(-2)
        G = (P.transpose(-2,-1) @ H_sym @ P) / D
        G = G.masked_fill(D == 0, 0)
        return P @ G @ P.transpose(-2, -1)
```

A.IV.6.6 Experimental Details and Additional Results

A.IV.6.6.1 Impact of the Number of Components for the Gradient Methods

In this section, we present the impact of the parameter K in the setting of Section A.IV.4.5. We fix $n = 2000, d = 3$ and $T = 30$, varying $K \in \{3, 5, 8, 10, 12, 15\}$ in Fig. A.IV.16. The results suggest that the number of components K greatly impacts the difficulty of the EM algorithm to converge to a "stable" fixed point, incurring worsening gradient approximations. In certain settings, the spectral norm of the Jacobian can be orders of magnitude larger than 1, completely invalidating the OS method. In this experiment, we required a regularisation of $10^{-8} I_d$ for the covariances for numerical stability.

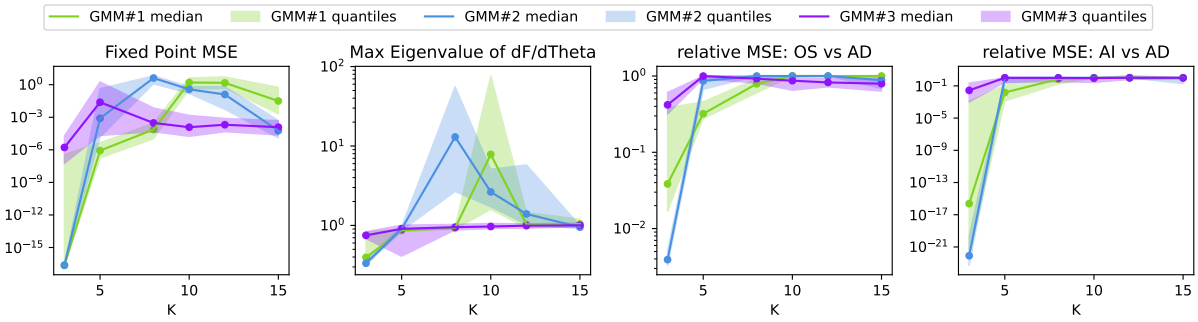


Figure A.IV.16: Varying the number of components K, we study the convergence of EM, the local contractivity of F, and the MSEs of the OS and AI gradients.

A.IV.6.6.2 Barycentres

We consider a barycentre problem similar to Section A.IV.5.1, with more complex datasets: we take three two-dimensional images I_1, I_2 and I_3, we randomly sample $n = 500$ points $Y_i \in \mathbb{R}^{n\times 2}$ from each. In Fig. A.IV.17, we flow a point cloud initialised as random normal noise, towards a barycentre of $K = 15$ GMMs fitted from (Y_i).

We can also compute a generalised barycentre X in $\mathbb{R}^{n\times 3}$, such that every projection $P_i(X) \in \mathbb{R}^{n\times 2}$ orthogonal to the canonical direction e_i coincides with Y_i. Specifically, we solve

$$\min_{X\in\mathbb{R}^{n\times 2}} \sum_{i=1}^{3} \mathrm{MW}_2^2 \left(P_i \# \mu(F_X^T(\theta_0)), \nu_i\right).$$

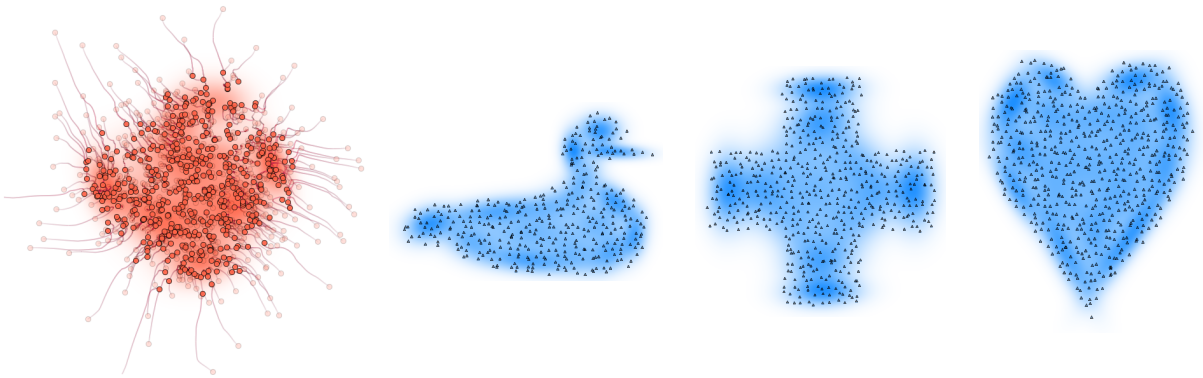


Figure A.IV.17: Left: EM – MW_2^2-Barycentre flow; right: input point clouds.

The GMMs $(\nu_i) \in \mathrm{GMM}_2(40)^3$ are fitted beforehand with the point clouds (Y_i), and μ_X is the running EM estimation of optimised cloud X. The projected GMM $P_i \# \mu(F_X^T(\theta_0))$ is defined by projecting the means and covariances of the three-dimensional GMM $\mu(F_X^T(\theta_0))$. We fit $K = 40$ Gaussian components in each cloud. Example results are shown in Fig. A.IV.18. We obtained very similar results with an alternative loss which estimates three GMMs in $\mathbb{R}^2$ instead of one in $\mathbb{R}^3$. In all our barycenter experiments, we set $\varepsilon_r = 10^{-3}$.

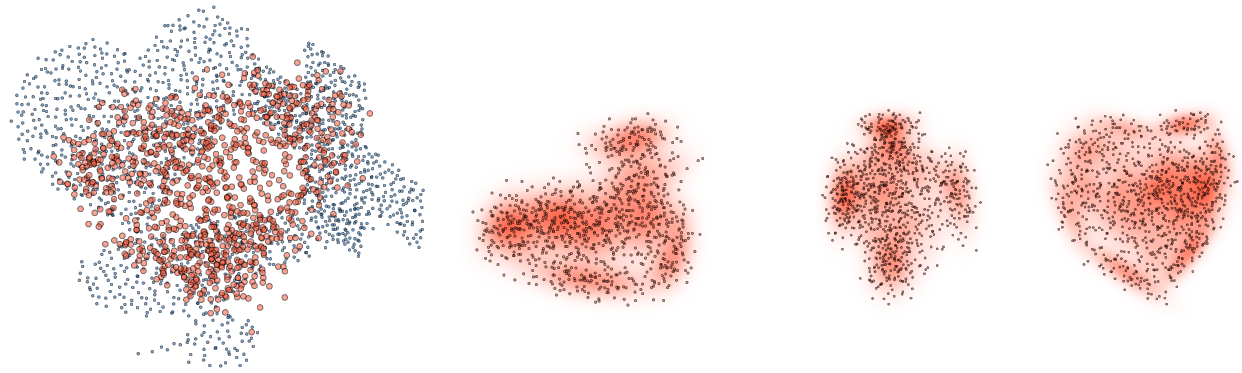


Figure A.IV.18: 3D barycentre (left) and its projections (right).

A.IV.6.6.3 Colour Transfer

Optimiser choice. We use gradient descent with fixed step size. Indeed, optimisers such as Adam have per-parameter learning rates that are adapted dynamically. As a result, points are treated differently and move at different speeds. For instance, imagine that we independently sample points $x_1, \cdots, x_n$ in $\mathbb{R}^d$ following the law $\mathcal{N}(0_d, I_d)$ and want to match them to $\mathcal{N}(0_d, 2I_d)$ using the Mixture Loss. Optimisers like Adam might focus on moving the points on the boundary to make the cloud's variance larger. Fixed-step gradient descent, on the contrary, will treat each point equally, and will rescale the cloud as a whole.

Fixing mixture weights. As discussed in Section A.IV.3.3, fixing the mixture weights to be uniform (using Algorithm A.IV.2) avoids local minima, as illustrated in Fig. A.IV.19. For a colour transfer task with $K = 6$ components, when weights are allowed to vary, some components of the optimised mixture (in red) are trapped between two target components (in blue).

Choice of the number of components K. We choose $K = 10$ Gaussian components in our colour transfer experiments. Taking $K = 1$, i.e. one Gaussian per source and target, is equivalent to applying $T : x \in \mathbb{R}^d \longmapsto m_t + A(x - m_s)$ to source pixels, where m_s and m_t are the empirical

means of source and target colour distributions, and where $A = \Sigma_s^{-1/2} \left(\Sigma_s^{1/2} \Sigma_t \Sigma_s^{1/2}\right)^{1/2} \Sigma_s^{-1/2}$. This map performs a coarser colour transfer than the optimisation with $K = 10$, as compared in Fig. A.IV.20: the contrast is better preserved with higher values of K. We always set $\varepsilon_r = 10^{-3}$ in this experiment.

A.IV.6.6.4 Neural Style Transfer

We note that the optimiser choice and EM variant (between Algorithms A.IV.1 and A.IV.2) do not change the results qualitatively. In our experiments, we use Adam and use standard EM Algorithm A.IV.1.

Choice of the number of components K. We choose $K = 3$ in our experiments. When using $K = 1$ Gaussian component, the style transfer objective in Eq. (A.IV.21) simplifies to

$$\min_{X \in \mathbb{R}^{3\times H\times W}} \sum_{\ell=1}^{3} \lambda_\ell \left(\|m_{s,\ell}(X) - m_{t,\ell}\|_2^2 + d_{\mathrm{BW}}^2 \left(\Sigma_{s,\ell}(X), \Sigma_{t,\ell}\right)\right),$$

where $m_{s,\ell}(X)$ and $\Sigma_{s,\ell}(X)$ are the empirical means and covariances of source features $\mathrm{VGG}_{1\cdots\ell}(X)$. We similarly define $m_{t,\ell}$ and $\Sigma_{t,\ell}$ for the target image Y. As there are only one source and target Gaussian, there is no need for an EM algorithm, we simply estimate the means and covariance and apply Gaussian OT. To evaluate the influence of K, we take two content images (the Eiffel tower and Gatys' [GEB15] picture of Tuebingen), and two style images (*The Great Wave* and *The Starry Night*). We compute their features corresponding to first layers $1, \cdots, \ell$ of VGG, for $\ell \in \{1, \cdots, 3\}$. We fit Gaussian mixtures on these features with varying number of components K, and we evaluate corresponding log-likelihoods as a measure of model quality. Results are presented in Fig. A.IV.21: most of them elbow around $K = 3$, the value we retain in our experiments. We set $\varepsilon_r = 0.1$ (as the dominant eigenvalue of the target covariance matrices empirically lies between 10^2 and 10^3).

See Fig. A.IV.22 for a comparison of style transfer with $K = 1$ and $K = 3$. Taking higher values of K does not yield significant improvement in the results. Yet, for Fig. A.IV.22a, the sky is less uniform with $K = 1$ and the bottom-left corner is more blurry. For Fig. A.IV.22c, the sky differs a bit between $K = 1$ and $K = 3$.

A.IV.6.6.5 Texture Synthesis

Consider a (space-periodic) target texture $u \in [0,1]^{h\times w\times C}$. Our objective is to produce a (space-periodic) texture $x \in [0,1]^{H\times W\times C}$. To initialise, we sample a stationary Gaussian field $Z \in \mathbb{R}^{H\times W\times C}$ of i.i.d. entries of law $\mathcal{N}(0, I_C)$. The initialisation is then a stationary Gaussian field with the same statistics as u, defined as $\mathbb{T}_{H,W}$ by $x_0 := m + u \star Z$, where $m \in \mathbb{R}^C$ is the mean of u and $\star$ denotes the discrete convolution with periodic boundary conditions.

Given a texture $v \in \mathbb{R}^{H\times W\times C}$, we consider $P_p(v) \subset [0,1]^{C\times p^2}$ the set of its $p \times p$ patches (with periodic boundary conditions and indexing from the top-left corner). We also define the down-sampling operator D_{s_i} that shrinks the image by 2^{s_i}. The patch distribution of a texture v is noted $\mu_p^v = \frac{1}{|P_p(v)|} \sum_{y\in P_p(v)} \delta_y$.

For a given list of scales $\mathcal{S} = \{(p_i, s_i)\}_{1\le i\le L}$, we aim to make the patch distributions $\mu_p(x)$ of the optimised texture x match the targets $\mu_p(v)$. Each μ_p^v is approximated by a Gaussian mixture $\hat{\mu}_p^v$. The target mixture $\hat{\mu}_p^u$ is fitted once by EM and fixed. We use the *Warm-Start EM* (see Algorithm A.IV.3) variant during optimisation, and take $\varepsilon_r = 10^{-3}$. We solve the problem

$$\min_x \sum_{i=1}^{L} 2^{2s_i} \mathrm{MW}_2^2 \left(\hat{\mu}_{p_i}^{D_{s_i}x}, \hat{\mu}_{p_i}^{D_{s_i}u}\right).$$

At the end, we perform a nearest neighbour projection on the largest scale: each patch in the generated x is matched to its nearest patch in the target u. Then, each pixel is reconstructed by averaging the corresponding patches.

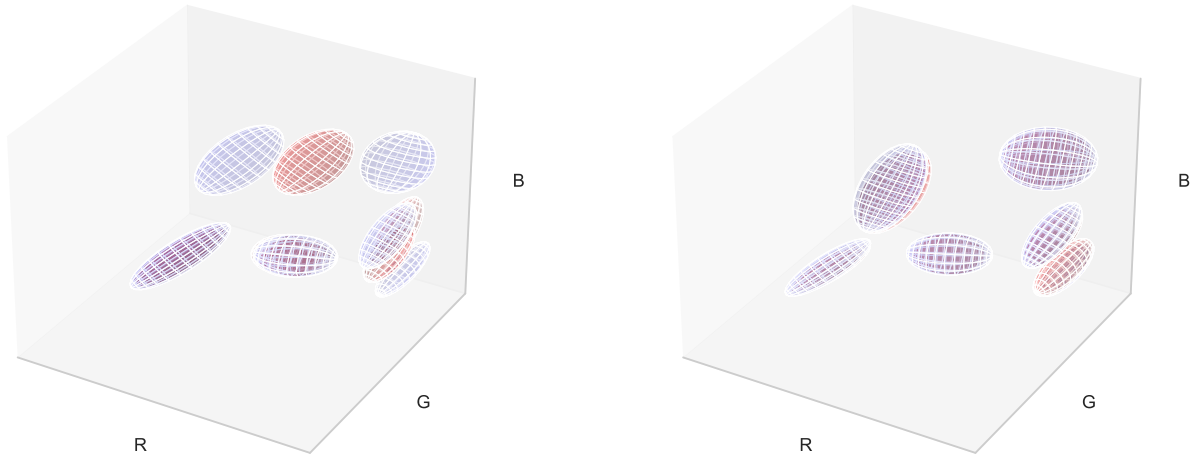
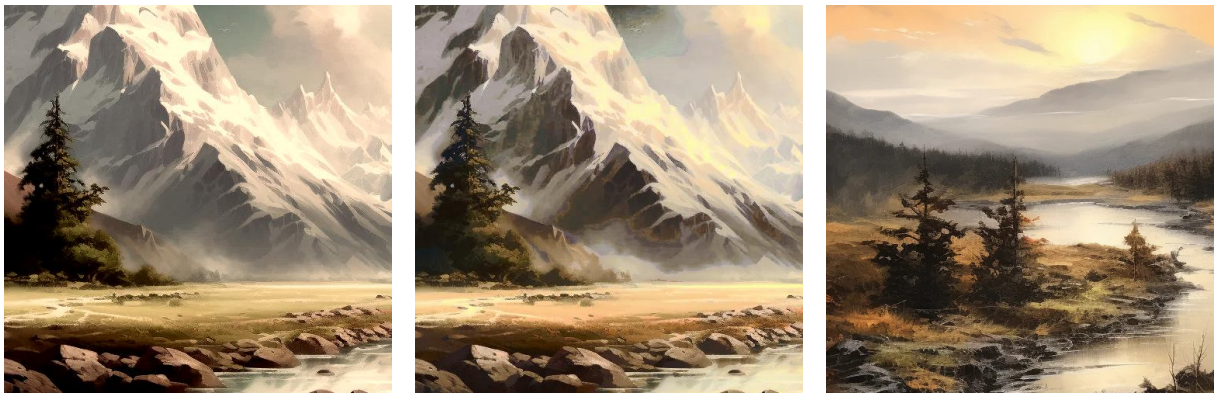
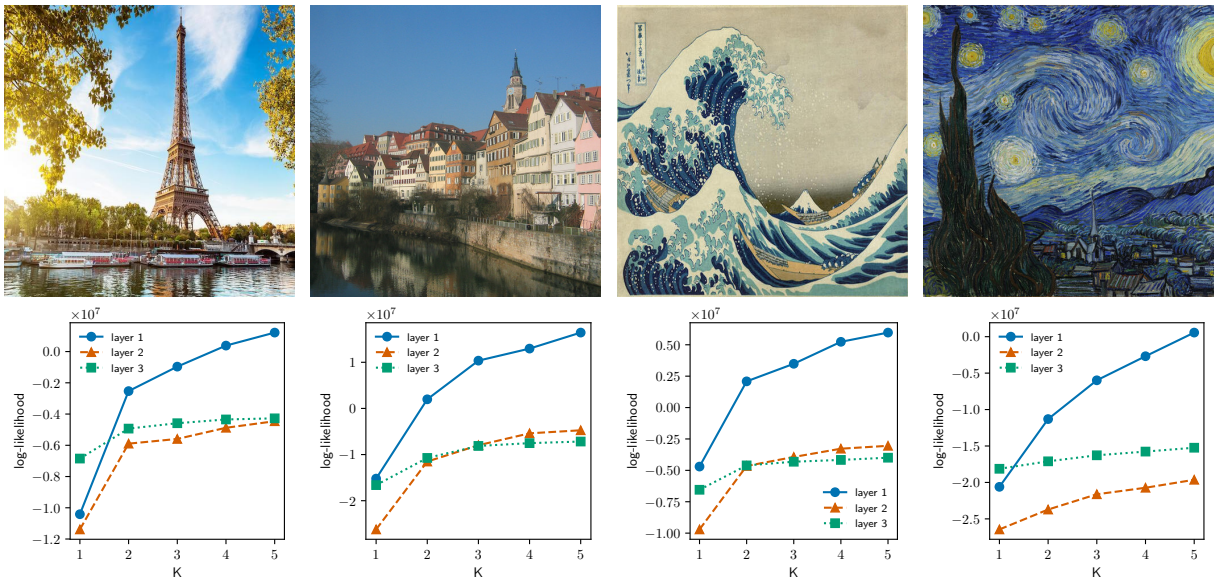


Figure A.IV.19: Final GMMs in RGB space with (left) variable weights and (right) fixed weights. The target is in blue and the optimised mixture is in red. On the left we are stuck in a local minimum, while on the right we converged.



(a) Result with $K = 1$ (b) Result with $K = 10$ (c) Target

Figure A.IV.20: Colour transfer with (a) $K = 1$ components and (b) $K = 10$ components.



(a) Eiffel Tower (b) Tuebingen (c) The Great Wave (d) The Starry Night

Figure A.IV.21: Four images and their corresponding log-likelihood vs. K plots, for VGG layers $\ell \in \{1, \cdots, 3\}$.

(a) Eiffel → The Great Wave (b) Tuebingen → Starry Night (c) Tuebingen → The Scream

Figure A.IV.22: Taking $K = 1$ (top) gives results comparable to $K = 3$ (bottom).

Part B

Variants of Optimal Transport Maps and Plans

Chapter B.I explores a variant of the Monge problem which minimises an OT cost between $g\#\mu$ and ν within a class of functions G. This chapter is based on the paper:

[TDD25] Eloi Tanguy, Agnès Desolneux, and Julie Delon.
"Constrained Approximate Optimal Transport Maps".
ESAIM: Control, Optimisation and Calculus of Variations. (Aug. 2025)

Chapter B.II investigates the theory behind transportation plans based on a novel Pivot Sliced Discrepancy based on [Mah+23] and on Expected Sliced Plans introduced by [Liu+24]. This chapter is based on the paper:

[TCD25] Eloi Tanguy, Laetitia Chapel and Julie Delon.
"Sliced Optimal Transport Plans".
arxiv preprint 2508.01243 (Aug. 2025).

Chapter B.III presents known algorithms for the resolution of the Gromov Wasserstein problem, and introduces a novel method using the "sliced plans" from Chapter B.II. At the time of writing, this method appears to be only cautiously promising. This chapter is based on a joint project with Laetitia Chapel, Julie Delon and Nicolas Courty.

B.I

Constrained Approximate Optimal Transport Maps

B.I.1 Introduction . . . 164
B.I.2 A Constrained Approximate Transport Map Problem . . . 167
B.I.2.1 Problem Definition 167
B.I.2.2 Existence of a Solution 168
B.I.2.3 Function Class Example: Gradients of Convex Functions 171
B.I.2.4 Function Class Example: Neural Networks 173
B.I.2.5 On the Necessity of the Lipschitz Constraint for Existence 173
B.I.2.6 Discussion on Uniqueness 175
B.I.2.7 The Plan Approximation Problem 177
B.I.3 Alternate Minimisation in the Squared Euclidean Case . . . 179
B.I.3.1 Projection of the Barycentric Map 180
B.I.3.2 Equivalence to a Constrained Barycentric Projection in Dimension 1 182
B.I.3.3 Non-Equivalence to Constrained Barycentric Projection in Dimension 2 183
B.I.4 Discrete Measures and Numerical Methods . . . 184
B.I.4.1 Regularity of Discrete Optimal Transport Costs 184
B.I.4.2 Numerical Method for Gradients of Convex Functions 185
B.I.4.3 Numerical Method for Maps in a RKHS 188
B.I.4.4 Stochastic Gradient Descent for Neural Networks 191
B.I.4.5 Illustrative Application to Colour Transfer 192
B.I.5 Conclusion and Outlook . . . 194
B.I.6 Appendix . . . 195
B.I.6.1 Continuous-to-Discrete Case: Semi-discrete OT 195
B.I.6.2 Lemmas on Pseudo-inverses and Quantile Functions 196
B.I.6.3 Reminder on Reduction in RKHS methods 197
B.I.6.4 Extending Kantorovich Potentials to Maps 198

Abstract

We investigate finding a map g within a function class G that minimises an Optimal Transport (OT) cost between a target measure ν and the image by g of a source measure μ. This is relevant when an OT map from μ to ν does not exist or does not satisfy the desired constraints of G. We address existence and uniqueness for generic subclasses of L-Lipschitz functions, including gradients of (strongly) convex functions and typical Neural Networks. We explore a variant that approaches a transport plan, showing equivalence to a map problem in some cases. For the squared Euclidean cost, we propose alternating minimisation over a transport plan π and map g, with the optimisation over g being the L^2 projection on G of the barycentric mapping $\overline{\pi}$. In dimension one, this global problem equates the L^2 projection of $\overline{\pi^*}$ onto G for an OT plan π^* between μ and ν, but this does not extend to higher dimensions. We introduce a simple kernel method to find g within a Reproducing Kernel Hilbert

Space in the discrete case. We present numerical methods for L-Lipschitz gradients of ℓ-strongly convex potentials, and study the convergence of Stochastic Gradient Descent methods for Neural Networks. We finish with an illustration on colour transfer, applying learned maps on new images, and showcasing outlier robustness. This chapter is based on the paper:

[TDD25] Eloi Tanguy, Agnès Desolneux, and Julie Delon.
"Constrained Approximate Optimal Transport Maps".
ESAIM: Control, Optimisation and Calculus of Variations. (Aug. 2025)

B.I.1 Introduction

Let μ and ν denote two probability distributions on two (potentially different) measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Many problems in applied fields can be written under the form

$$\inf_{g \in G} \mathcal{D}(g\#\mu, \nu), \tag{B.I.1}$$

where $\#$ denotes the *push-forward* operation[1], $\mathcal{D}$ is a non-negative discrepancy (such as a distance metric or a ϕ-divergence) measuring the similarity between $g\#\mu$ and ν, and G is a set of acceptable functions from $\mathcal{X}$ to $\mathcal{Y}$. Under appropriate assumptions on $\mathcal{D}$, this problem can be interpreted as a projection of ν on the set $G\#\mu := \{g\#\mu, g \in G\}$ for the discrepancy $\mathcal{D}$. In this chapter, we focus on cases where ν cannot be written as $g\#\mu$ for $g \in G$.[2]

In the highly popular field of generative modelling, the target distribution is usually an empirical distribution composed of m samples, $\nu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$, μ is an easy-to-sample latent distribution (for instance a Gaussian distribution), and the set $G = \{g_\theta,\ \theta \in \Theta\}$ generally denotes functions represented by a specific neural network architecture. The goal is to find the parameter θ such that $\mu_\theta := g_\theta\#\mu$ fits ν as well as possible. Models taking this form are often called push-forward generative models [Sal+22], and include Variational Auto-Encoders (VAEs) [KW14], Generative Adversarial Networks (GANs) [Goo+14], Normalising flows [RM15] and even Diffusion Models [Son+20], which can be reinterpreted as indirect push-forward generative models [Sal+22]. In these works, the discrepancy $\mathcal{D}$ is often chosen as the Kullback-Leibler divergence, as it is the case for traditional GANs and VAEs, or as the Wasserstein distance, like in Wasserstein-GANs [ACB17]. The discrepancy $\mathcal{D}(g_\theta\#\mu, \nu)$ is minimised in θ, for instance by using sophisticated versions of stochastic gradient descent. In such problems, it is clear that $g_\theta\#\mu$ does not target exactly ν, since it would mean that the model has only learned to reproduce existing samples, and not to create new ones. This is possible because the expressivity of neural networks is limited, but also because the training steps usually impose regularity properties on g_θ and constrain its Lipschitz constant in order to increase its robustness [VS18; Faz+19] or stabilise its training [Miy+18]. It is therefore natural to wonder to what extent the optimisation of such discrepancies with regularity constraints on the set of functions G is well-posed, depending on the choice of $\mathcal{D}$, and what this means in practice.

In an Euclidean setting, another example of Eq. (B.I.1) appears when we need to compare two distributions μ and ν potentially living in spaces of different dimensions, or when invariance to geometric transformations is required (for problems such as shape matching or word embedding). In such cases, it is usual to choose G as a well chosen set of linear or affine embeddings (such as matrices in the Stiefel manifold if the space dimension is different between $\mathcal{X}$ and $\mathcal{Y}$). For instance, this idea underpins several sets of works introducing global invariances in optimal transport [AJJ19; SDD23].

In both of the previous examples, G is parametrised by a set Θ of parameters which is potentially extremely large (for neural networks) but of finite dimension. Alternatively, the

[1] The image measure $g\#\mu$ is defined as the law of $g(X)$ for X a random variable of law μ, or more abstractly by $g\#\mu(B) = \mu(g^{-1}(B))$ for any Borel set $B \subset \mathcal{Y}$.

[2] Obviously, if ν belongs to $G\#\mu$, the problem is trivial (from a theoretical standpoint) and the infimum in Eq. (B.I.4) is 0.

set of functions G can be much more complex and characterised by regularity or convexity assumptions, the problem becoming non-parametric. This is typically the case in the field of optimal transport [Vil09; San15]. Given μ and ν probability measures on respective Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, Monge's Optimal Transport consists in finding a Transport map T such that $T\#\mu = \nu$ and which minimises a given displacement cost. From a theoretical standpoint, the existence of (unconstrained) Monge maps has been widely studied [Bre91; GM96; Pra07], under regularity assumption for μ.

The regularity of an Optimal Transport map, when it exists, has garnered substantial attention. The first studies date to Cafarelli [Caf00] through the study of the *Monge-Ampère* equation, which (by the change-of-variables formula), is satisfied by any smooth and injective transport map from $\mu = f\mathrm{d}x$ to $\nu = g\mathrm{d}x$:

$$g(T(x))|\det(\partial T(x))| = f(x), \text{ for almost-every } x \in \mathbb{R}^d. \tag{B.I.2}$$

The study of the Monge-Ampère equation (in more general settings) and its ties to regularity in Optimal Transport is an active field of research, we refer to the surveys by Figalli [Fig09], De Philippis and Figalli [DF14], as well as a monograph by Figalli [Fig17].

When there is no map T such that $T\#\mu = \nu$ (for example, if μ is discrete and if ν is not), or when the map solution does not meet the regularity requirements for some given practical setting, it makes sense instead to solve problems of the form of Eq. (B.I.1), with $\mathcal{D}$ a Wasserstein distance and G a set of functions with acceptable regularity. For instance, as studied in [PdC20], G can be composed of functions $g = \nabla\phi$ with ϕ ℓ-strongly-convex with an L-Lipschitz gradient. For cases where μ is discrete, this formulation also overcomes a classic shortcoming of numerical optimal transport approaches, which usually compute solutions which are only defined on the support of μ. If a machine learning algorithm requires the computation of the transport of new inputs, the map must be either recomputed, or an approximation of the previous map must be defined outside of the support of μ. Several solutions have been proposed in the literature to solve this problem [LGL21; BYF20; Man+24; PdC20; PN24; Seg+18; Per+16], and some of them [Man+24; PdC20] consists in solving Eq. (B.I.1) with an appropriate set of functions G. Consistency and asymptotic properties of such estimators are also the subject of several of these works [LGL21; Man+24; HR21].

For the sake of legibility and to avoid excessive technicality, we focus on the case where the target space is $\mathbb{R}^d$, however it is possible to extend our considerations to a target space $\mathcal{Y}$ which is a Polish space verifying the Heine-Borel property (i.e. that any bounded and closed set is a compact set), which in particular allows the case where $\mathcal{Y}$ is a connected and complete Riemannian manifold (in which case the Heine-Borel property follows from the Hopf-Rinow Theorem, see [DF92, Theorem 2.8]). Similarly, the problem naturally extends to the case where the codomain of the maps g and the target measure ν are different spaces $\mathcal{Y}, \mathcal{Y}'$.

OT discrepancies. In this chapter, we focus on problems of the form Eq. (B.I.1) when $\mathcal{D}$ is chosen as an optimal transport discrepancy for a general ground cost c. We recall that if $\mathcal{X}$ and $\mathcal{Y}$ are two Polish spaces, the Optimal Transport cost between two measures $\nu_1 \in \mathcal{P}(\mathcal{X})$ and $\nu_2 \in \mathcal{P}(\mathcal{Y})$ for a ground cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is defined by the following optimisation problem

$$\mathcal{T}_c(\nu_1, \nu_2) = \min_{\pi \in \Pi(\nu_1, \nu_2)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y), \tag{B.I.3}$$

where $\Pi(\nu_1, \nu_2)$ is the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ whose first marginal is ν_1 and second marginal is ν_2[3]. Given this method of quantifying the discrepancy between $g\#\mu$ and ν, Eq. (B.I.1) becomes

$$\inf_{g \in G} \mathcal{T}_c(g\#\mu, \nu). \tag{B.I.4}$$

[3]The fact that the minimum is attained is a consequence of the direct method of calculus of variations (see [San15, Theorem 1.7]). The value of $\mathcal{T}_c(\nu_1, \nu_2)$ may be $+\infty$, but a sufficient condition for $\mathcal{T}_c(\nu_1, \nu_2) < +\infty$ ([Vil09, Remark 5.14]) is that

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\nu_1(x)\mathrm{d}\nu_2(y) < +\infty.$$

In the case where the source measure is discrete and the target measure is absolutely continuous, the Optimal Transport problem in Eq. (B.I.4) is said to be semi-discrete, and has a slightly more explicit expression (see [MT21] for a course on the matter). If we suppose in addition that $c(x, y) = \|x-y\|_2^2$ and that the source measure weights are uniform ($a_i = 1/n$), then Eq. (B.I.4) is a constrained version the Optimal Uniform Quantization problem studied thoroughly in [MSS21].

Existence of minimisers. An important question regarding this optimisation problem concerns the existence of minimisers, depending on the ground cost c and the set of functions G. While numerous works in the literature have focused on the convergence of optimisation algorithms (such as stochastic gradient descent) to critical points for this kind of problem [Fat+21b], the existence of minimisers has surprisingly been little studied. We derive in Theorems B.I.1 and B.I.2 generic conditions to ensure existence of such minimisers in G, and show counter-examples when these conditions are not met. We also show that these conditions are satisfied for two classes of functions, namely classes of L-Lipschitz functions which can be written as gradient of l-strongly convex functions (recovering a result shown in [PdC20] as a particular case of Theorem B.I.2), and classes of neural networks with Lipschitz activation functions. We also discuss uniqueness of the solutions, which is usually not satisfied, and remains a difficult question without strong assumptions on the set of functions G.

Approximating a coupling. In the field of optimal transport, a particular setting where Eq. (B.I.4) is interesting is when we have access to a non deterministic coupling π solution of a regularised version of a optimal transport between two probability measures μ and ν. For instance, the entropic optimal transport [PC19b], or the mixture Wasserstein formulation [DD20] both yield optimal plans π which cannot be trivially written as optimal maps between μ and ν. For some applications, it can be interesting to approximate π by another transport plan supported by the graph of a function with possible additional regularity assumptions. This can be done by approximating π by $(I, g)\#\mu$, with specific regularity properties on g, which is a particular case of Eq. (B.I.4), replacing ν by π and G by the set $H := \{(I, g), g \in G\}$. In this specific setting, we show in Section B.I.2.7 under which conditions on the ground cost c the solutions of this problem between plans are equivalent to solutions of the original Eq. (B.I.4) when $\pi \in \Pi(\mu, \nu)$. Numerical approaches seeking maps that approach barycentric projections have been studied in [Seg+18; Per+16].

Alternate minimisation. Under appropriate assumptions, Eq. (B.I.4) can be rewritten as a minimisation problem over $\pi \in \Pi(\mu, \nu)$ and $g \in G$:

$$\min_{g \in G} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(g(x), y) \mathrm{d}\pi(x, y). \tag{B.I.5}$$

This naturally leads to consider Eq. (B.I.4) as an alternate minimisation problem, that we study in Section B.I.3 in the Euclidean case when $c(x, y) = \|x - y\|_2^2$. More precisely, we show that Eq. (B.I.4) is strongly linked to the barycentric projection problem: when π is fixed, the solution g minimising Eq. (B.I.5) can be reinterpreted as the L^2-projection of the barycentric projection of π on the set G. In the one-dimensional case, when G is a subclass of increasing functions, this yields an explicit solution to the problem (as it was shown in [PdC20] in a more specific case), and we show that this explicit solution does not hold in dimension larger than 1 by presenting a counter-example.

Extending Discrete Kantorovich Potentials to Maps A particular advantage of solving the constrained map problem is that it defines a map g even outside the support of the source measure μ. In Section B.I.6.4, we consider the case where both measures are discrete, and study a known yet scarcely documented method which extends discrete Kantorovich potentials to the whole space, thereby extending a discrete OT map to the whole space. We prove some properties of this extended map, and compare it to our approximate map method.

Outline of the chapter. In this work, we address problem Eq. (B.I.4) for large classes of functions G. In Section B.I.2, we define the problem and establish general conditions for the existence of solutions, exploring examples involving gradients of convex functions and neural networks. Section B.I.3 examines the link between Eq. (B.I.4) and a constrained barycentric projection problem, demonstrating an explicit solution in one dimension and providing a counterexample in higher dimension. Section B.I.4 focuses on practical numerical methods to solve

the optimisation problems for Lipschitz gradients of strongly convex potentials, kernel methods and Neural Networks. We conclude with an illustration on colour transfer.

B.I.2 A Constrained Approximate Transport Map Problem

B.I.2.1 Problem Definition

We consider $(\mathcal{X}, d_{\mathcal{X}})$ a locally compact Polish space, and $\mu \in \mathcal{P}(\mathcal{X})$ a probability measure on $\mathcal{X}$. Our objective is to find a map $g : \mathcal{X} \longrightarrow \mathbb{R}^d$ verifying the constraint $g \in G$ for some class of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$, such that the image measure $g\#\mu$ is "close" to a fixed probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$, in the sense of Eq. (B.I.4). Applying the definition of $\mathcal{T}_c$ directly (Eq. (B.I.3)) yields the following expression for Eq. (B.I.4):

$$\inf_{g \in G} \min_{\pi \in \Pi(g\#\mu, \nu)} \int_{\mathcal{X} \times \mathbb{R}^d} c(x, y) \mathrm{d}\pi(x, y). \tag{B.I.6}$$

The optimisation variable g acts on the set of constraints of the Optimal Transport problem, however thanks to a well-known "change of variables" result (see [DLV24] for a reference), we will be able to reformulate Eq. (B.I.6). In the following, we shall denote by $\Pi_c^*(\nu_1, \nu_2)$ the set of minimisers of the optimal transport problem Eq. (B.I.3) between two measures ν_1 and ν_2.

Lemma B.I.1. ([DLV24, Lemmas 2.6 and 2.7],) Let $\mathcal{X}, \mathcal{Y}, \mathcal{X}', \mathcal{Y}'$ be Polish spaces. Let $g : \mathcal{X} \longrightarrow \mathcal{X}'$ and $h : \mathcal{Y} \longrightarrow \mathcal{Y}'$ two measurable maps and let $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$. Consider two costs $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ and $c' : \mathcal{X}' \times \mathcal{Y}' \longrightarrow \mathbb{R}$ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, $c(x, y) = c'(g(x), h(y))$.

- For any $\gamma' \in \Pi(g\#\mu, h\#\nu)$, there exists $\gamma \in \Pi(\mu, \nu)$ such that $\gamma' = (g, h)\#\gamma$.
- We have $\Pi_{c'}^*(g\#\mu, h\#\nu) = (g, h)\#\Pi_c^*(\mu, \nu)$.

Using Lemma B.I.1, the energy of the map problem Eq. (B.I.4) can be written as follows:

$$\mathcal{T}_c(g\#\mu, \gamma) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathbb{R}^d} c(g(x), y) \mathrm{d}\pi(x, y). \tag{B.I.7}$$

In our study of the map problem Eq. (B.I.4), we will consider classes G that are a subset of the L-Lipschitz functions. The first reason is that with unbounded Lipschitz constants, the problem may not have a solution, as we shall see in Section B.I.2.5. Moreover, there are multiple practical considerations that lead to choosing functions with an upper-bounded Lipschitz constant. To begin with, numerous practical models enforce this condition, such as Wasserstein GANs [ACB17], and diffusion models [Son+20] (see also [Sal+22] Appendix S2), furthermore most neural networks are Lipschitz (since typical non-linearities are chosen as Lipschitz), and the control of the Lipschitz constant is often used as a regularisation method [VS18].

From a theoretical standpoint, a Lipschitz function g has the convenient property of conserving the moment conditions of a measure μ through its image measure, as we show in Lemma B.I.2, which automatically ensures the finiteness of the transport cost $\mathcal{T}_c(g\#\mu, \nu)$ for measures admitting p-moments and $c(x, y) = d_{\mathcal{X}}(x, y)^p$.

Lemma B.I.2. Let $(\mathcal{X}, d_{\mathcal{X}})$ a Polish space and μ a probability measure on $\mathcal{X}$ with a finite moment of order $p \geq 1$: $\int_{\mathcal{X}} d_{\mathcal{X}}(x_0, \cdot)^p \mathrm{d}\mu < +\infty$ (for any or all $x_0 \in \mathcal{X}$). Then for an L-Lipschitz function $g : \mathcal{X} \longrightarrow \mathcal{Y}$, with $(\mathcal{Y}, d_{\mathcal{Y}})$ a Polish space, the push-forward measure $g\#\mu$ also has a finite moment of order p.

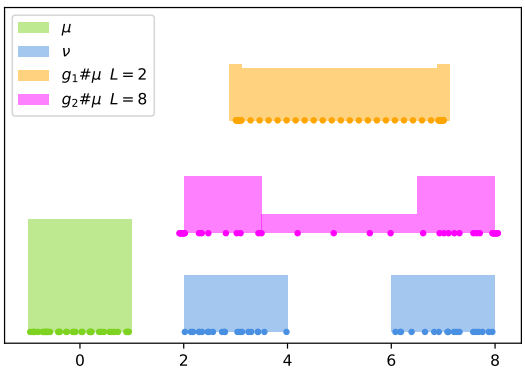
Proof. Choose $y_0 \in \mathcal{Y}$ and $x_0 \in \mathrm{supp}(\mu)$. We have $\int_{\mathcal{Y}} d_{\mathcal{Y}}(y_0, y)^p \mathrm{d}g\#\mu(y) = \int_{\mathcal{X}} d_{\mathcal{Y}}(y_0, g(x))^p \mathrm{d}\mu(x)$,

then given $x \in \mathcal{X}$, write

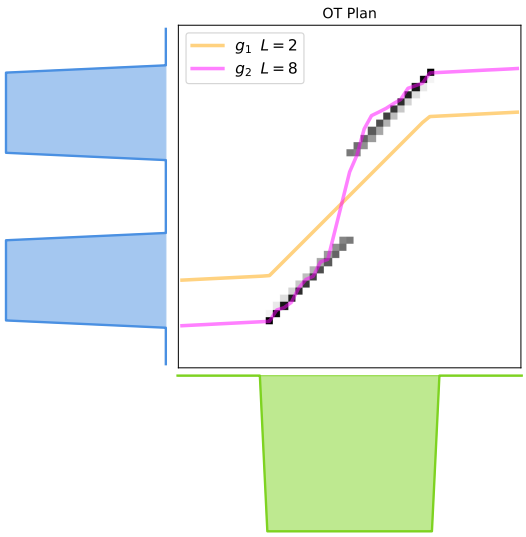
$$\begin{aligned} d_{\mathcal{Y}}(y_0, g(x))^p &\leq (d_{\mathcal{Y}}(y_0, g(x_0)) + d_{\mathcal{Y}}(g(x_0), g(x)))^p \\ &\leq 2^{p-1}(d_{\mathcal{Y}}(y_0, g(x_0))^p + d_{\mathcal{Y}}(g(x_0), g(x))^p) \\ &\leq 2^{p-1} d_{\mathcal{Y}}(y_0, g(x_0))^p + 2^{p-1} L^p d_{\mathcal{X}}(x_0, x)^p, \end{aligned}$$

where we used the inequality $(a+b)^p = 2^p(\frac{a}{2} + \frac{b}{2})^p \leq 2^{p-1}(a^p + b^p)$ for $a, b \geq 0$, by convexity of $t \longmapsto t^p$. Now the constant $2^{p-1} d_{\mathcal{Y}}(y_0, g(x_0))^p$ is μ-integrable since μ is a probability measure, and the function $d_{\mathcal{X}}(x_0, \cdot)^p$ is integrable since μ has a finite moment of order p. □

In Fig. B.I.1, we illustrate a solution of the map problem using numerical methods introduced in Section B.I.4.2, for two different values of L (the Lipschitz constant of the maps g).



(a) Illustration of the source, target and image measures using their samples and (approximate) densities.



(b) Illustration of map solutions and comparison with the (discontinuous) Optimal Transport coupling.

Figure B.I.1: Illustration of solutions of maps problems (Eq. (B.I.4)) on a toy dataset with a source measure $\mu = \mathcal{U}([-1, 1])$ and a target measure $\nu = \frac{1}{2}\mathcal{U}([2, 4]) + \frac{1}{2}\mathcal{U}([6, 8])$. The two solutions are respectively $L = 2$ and $L = 8$ Lipschitz.

B.I.2.2 Existence of a Solution

To formulate an existence result, we shall apply the direct method of calculus of variations, which requires a technical condition on the stability of the class of functions G with respect to certain limits. To formulate this condition, we will introduce the notion of closedness of a class of continuous functions with respect to the compact-open topology. By [Kel17, Chapter 7, Theorem 11], in our setting, this topology is equivalent to the topology of uniform convergence on compact sets, which allows us to formulate Definition B.I.1 in terms of local uniform convergence.

Definition B.I.1. We say that a set of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$ is **closed for the compact-open topology** if there exists a sequence $(\mathcal{K}_m)$ of compact sets of $\mathcal{X}$ verifying $\cup_m \mathcal{K}_m = \mathcal{X}$ such that:
for any sequence $(g_n)_{n\in\mathbb{N}} \in G^{\mathbb{N}}$ such that for all m, $(g_n|_{\mathcal{K}_m})_{n\in\mathbb{N}}$ converges uniformly towards a function $g_{\mathcal{K}_m} : \mathcal{K}_m \longrightarrow \mathbb{R}^d$, there exists $g \in G$ such that $g|_{\mathcal{K}_m} = g_{\mathcal{K}_m}$ for all m.

One can understand this condition as a form of "local uniform closedness" of the class G.

Theorem B.I.1. Let $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ be a lower semi-continuous cost function, $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure on a locally compact Polish space $(\mathcal{X}, d_{\mathcal{X}})$, and $\nu \in \mathcal{P}(\mathbb{R}^d)$. Assume that:

i) **(Coercive cost)** There exists $\eta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ non-decreasing and such that $\eta(t) \xrightarrow[t \longrightarrow +\infty]{} +\infty$, and $\alpha \in \mathbb{R}$ such that $\forall y, y' \in \mathbb{R}^d,\ c(y, y') \geq \alpha + \eta(\|y - y'\|_2)$ and $\int \eta(\|\cdot\|_2) \mathrm{d}\nu < +\infty$;

ii) **(Lipschitzness and Closedness of G)** G is a subset of the space of L-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}^d$, that is closed for the compact-open topology (see Definition B.I.1);

iii) **(Problem finiteness)** There exists $g \in G$ such that $\mathcal{T}_c(g\#\mu, \nu) < +\infty$.

Then the problem $\underset{g \in G}{\operatorname{argmin}}\ \mathcal{T}_c(g\#\mu, \nu)$ has a solution.

Proof. — *Step 1*: Defining a minimising sequence.

We introduce the notation $J(g) := \mathcal{T}_c(g\#\mu, \nu)$ for convenience, and J^* the problem value, which is finite by Assumption iii). Consider a minimising sequence $(g_n)_{n \in \mathbb{N}} \in G^{\mathbb{N}}$ such that

$$\forall n \in \mathbb{N},\ J(g_n) \leq J^* + 2^{-n}.$$

— *Step 2*: Bounding g_n.

First, we fix $n \in \mathbb{N}$, and take $a \in \mathcal{X}$ in the support of μ and $r > 0$, then set $A := B_{d_{\mathcal{X}}}(a, r)$ the ball of centre a and radius r for the distance $d_{\mathcal{X}}$, so that $\mu(A) > 0$. The transport problem has a solution since c is lower semi-continuous ([San15, Theorem 1.7]). We introduce $\pi_n^* \in \Pi(\mu, \nu)$ optimal for the OT cost $\mathcal{T}_c(g_n\#\mu, \nu)$. We lower-bound:

$$J(g_n) \geq \int_{A \times \mathbb{R}^d} c(g_n(x), y) \mathrm{d}\pi_n^*(x, y) \geq \int_{A \times \mathbb{R}^d} \eta(\|g_n(x) - y\|_2) \mathrm{d}\pi_n^*(x, y) + \alpha\mu(A).$$

To separate variables, we will use an elementary inequality: let $z, w \in \mathbb{R}^d$, the triangle inequality yields $\|z\|_2 \leq 2\max(\|w - z\|_2, \|w\|_2)$, applying the non-decreasing and non-negative function η provides $\eta(\|z\|_2/2) \leq \max(\eta(\|w - z\|_2), \eta(\|w\|_2)) \leq \eta(\|w - z\|_2) + \eta(\|w\|_2)$. Finally, we have

$$\forall w, z \in \mathbb{R}^d,\ \eta(\|w - z\|_2) \geq \eta(\|z\|_2/2) - \eta(\|w\|_2). \tag{B.I.8}$$

By assumption, we remind that $\int \eta(\|\cdot\|_2) \mathrm{d}\nu < +\infty$, and resume lower-bounding using Eq. (B.I.8) with $w := y$ and $z := g_n(x)$:

$$J(g_n) \geq \int_A \eta\left(\frac{\|g_n(x)\|_2}{2}\right) \mathrm{d}\mu(x) - \mu(A) \int \eta(\|\cdot\|_2) \mathrm{d}\nu + \alpha\mu(A).$$

Let $x \in A$, we apply again Eq. (B.I.8) with $w := (g_n(a) - g_n(x))/2$ and $z := g_n(a)/2$:

$$\eta(\|g_n(x)\|_2/2) \geq \eta(\|g_n(a)\|_2/4) - \eta(\|g_n(a) - g_n(x)\|_2/2) \geq \eta(\|g_n(a)\|_2/4) - \eta(Lr/2),$$

where the second inequality comes from the fact that g_n is L-Lipschitz, $d_{\mathcal{X}}(x, a) \leq r$ and that η is non-decreasing. Gathering our inequalities leads to the following lower-bound:

$$J^* + 1 \geq J(g_n) \geq \mu(A) \left(\eta(\|g_n(a)\|_2/4) - \eta(Lr/2) - \int \eta(\|\cdot\|_2) \mathrm{d}\nu + \alpha\right).$$

This implies that there exists $M > 0$ independent of n such that $\|g_n(a)\|_2 \leq M$. (Since by coercivity of η, the right-hand side of the equation above would tend to $+\infty$ if $\|g_n(a)\|_2 \xrightarrow[n \longrightarrow +\infty]{} +\infty$).

— *Step 3*: Applying Arzelà-Ascoli's Theorem.

For $n \in \mathbb{N}$, we use the upper-bound from Step 2 and the fact that each g_n is L-Lipschitz:

$$\forall x \in \mathcal{X},\ \|g_n(x)\|_2 \leq M + L d_{\mathcal{X}}(x, a),$$

which shows that $\forall x \in \mathcal{X},\ \{g_n(x),\ n \in \mathbb{N}\}$ has compact closure in $\mathbb{R}^d$. The sequence (g_n) is equi-Lipschitz and thus equi-continuous, and is closed for the compact-open topology (Definition B.I.1) by assumption. By Arzelà-Ascoli's theorem (as stated in [Kel17, Chapter 7, Theorem 17]), we can choose $\beta : \mathbb{N} \longrightarrow \mathbb{N}$ an extraction such that $g_{\beta(n)} \xrightarrow[n \longrightarrow +\infty]{} g$ locally uniformly on $\mathcal{X}$, for a certain function $g \in G$.

— *Step 4*: Showing that the limit g is optimal.

First, the sequence $(g_{\beta(n)}\#\mu)$ converges weakly towards the probability measure $g\#\mu$: take a continuous and compactly supported test function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}$, the dominated convergence theorem shows that

$$\int_{\mathcal{X}} \phi \circ g_{\beta(n)} \mathrm{d}\mu \xrightarrow[n \longrightarrow +\infty]{} \int_{\mathcal{X}} \phi \circ g \mathrm{d}\mu,$$

where convergence of the integrands is ensured by the point-wise convergence of $(g_{\beta(n)})$, and domination by $\|\phi\|_\infty$ suffices. Since c is lower semi-continuous, the OT cost is itself lower semi-continuous for the weak convergence of measures (see [ABS+21, Theorem 2.6]), we obtain the following inequality:

$$\liminf_{n \longrightarrow +\infty} J(g_{\beta(n)}) \geq J(g),$$

where J was introduced in Step 1, where we also chose g_n such as $J(g_n) \leq J^* + 2^{-n}$, thus we conclude $J^* \geq J(g)$, hence g is optimal. □

Theorem B.I.1 can be extended to the case where the regularity of the functions of G is only assumed *on a partition* of $\mathcal{X}$. Note that to avoid pathological ambiguity and unnecessary complications, we will consider partitions whose borders have no mass for μ, such that the problem objective can be split according to the partition.

Theorem B.I.2. Let $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ be a continuous cost function, a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ on a locally compact Polish space $(\mathcal{X}, d_{\mathcal{X}})$, and $\nu \in \mathcal{P}(\mathbb{R}^d)$. Consider $(E_i)_{[\![1,K]\!]}$ a partition of $\mathcal{X}$ such that for every $i \in [\![1, K]\!]$, $\mu(\partial E_i) = 0$. Under the same conditions as B.I.1, and replacing assumption ii) by

ii') The class of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$ is of the form

$$G = \left\{ g : \mathcal{X} \longrightarrow \mathbb{R}^d \mid \forall i \in [\![1, K]\!],\ g|_{\mathring{E}_i} = g_i,\ g_i \in G_i \right\},$$

where for every $i \in [\![1, K]\!]$, the set of functions $G_i \subset (\mathbb{R}^d)^{\mathring{E}_i}$ is a subset of the space of L-Lipschitz functions from $\mathring{E}_i$ to $\mathbb{R}^d$, that is closed for the compact-open topology (see Definition B.I.1),

then the problem $\underset{g \in G}{\operatorname{argmin}}\ \mathcal{T}_c(g\#\mu, \nu)$ has a solution.

Proof. We shall follow closely the proof of Theorem B.I.1, and point out the technical differences. We introduce a minimising sequence exactly identically to Step 1. The computations from Step 2 can be done verbatim, choosing instead $A_i \subset \mathring{E}_i$, and concluding $\|g_n(a_i)\|_2 \leq M_i$ for a fixed $a_i \in A_i$.

Step 3 is then done separately on each $\mathring{E}_i$, yielding extractions (β_i) such that each $g_{\beta_i(n)}$ converges locally uniformly on $\mathring{E}_i$ towards a function $g_i \in G_i$. Considering the extraction $\beta := \beta_1 \circ \cdots \circ \beta_K$, we have for all $i \in [\![1, K]\!]$ the uniform convergence of $(g_{\beta(n)})$ towards $g \in G$ on all compact sets of $\mathring{E}_i$.

Finally, Step 4 is done likewise to Theorem B.I.1, with the technicality that since $\mu(\partial E_i) = 0$, the pointwise convergence of $(g_{\beta(n)})$ towards g at each point of $\mathring{E}_i$ suffices to show that

$g_{\beta(n)}(x) \xrightarrow[n\longrightarrow+\infty]{} g(x)$ for μ-almost-every $x \in \mathcal{X}$, which yields the convergence in law

$$g_{\beta(n)}\#\mu \xrightarrow[n\longrightarrow+\infty]{w} g\#\mu.$$

The rest follows verbatim. □

In Remarks B.I.1 and B.I.2, we present some natural extensions of Theorems B.I.1 and B.I.2, which we kept separate for legibility.

Remark B.I.1. The existence results of Theorems B.I.1 and B.I.2 also hold if the objective functional is changed into a regularised version

$$J(g) = \mathcal{T}_c(g\#\mu, \nu) + R(g),$$

where $R : G \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ is lower semi-continuous with respect to uniform local convergence. One also has to assume that there still exists $g \in G$ such that the new cost J is finite. The proofs can be written almost identically: in Step 1, it suffices to lower-bound $R(g_n) \geq 0$, and in Step 4, one obtains $\liminf\ J(g_{\beta(n)}) \geq J(g)$ thanks to the lower semi-continuity of R.

Remark B.I.2. Condition i) on c can be generalised to the case where the target space $\mathbb{R}^d$ is instead a Polish space $\mathcal{Y}$ verifying the Heine-Borel property (i.e. all closed and bounded sets are compact), in which case Condition i) can be replaced with the condition that $c(\cdot, y_0)$ be **proper**, which is to say that its preimage by any compact set $S \subset \mathbb{R}_+$ is a compact set of $\mathcal{Y}$. This property would be used in Step 2 to show that $g_n(a) \in C$ for some compact set $C \subset \mathcal{Y}$ independent of n, then in Step 3, we would use the Lipschitz property of g_n and the triangle inequality on $d_{\mathcal{Y}}$ to show that $\forall x \in \mathcal{K},\ g_n(x) \in \overline{B}_{\mathcal{Y}}(y_0, Lr + d_{\mathcal{Y}}(y_0, C))$, for a compact set $\mathcal{K} \subset \mathcal{X}$ of diameter r and $y_0 \in \mathcal{Y}$. This would show that for each $x \in \mathcal{K}$, the set $\{g_n(x)\}_{n\in\mathbb{N}}$ is pre-compact in $\mathcal{Y}$, and allow one to apply Arzelà-Ascoli likewise.

A natural context for Optimal Transport is the case where the ground cost is of the from $c(x, y) = \|x - y\|^p$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$ and $p \geq 1$. In Proposition B.I.1, we show that such costs verify the assumptions to our existence results.

Proposition B.I.1. Cost functions of the form $c(x, y) := \|x - y\|^p$, where $p > 0$ and $\|\cdot\|$ is a norm on $\mathbb{R}^d$ satisfy Assumption i) of Theorems B.I.1 and B.I.2, as long as $\nu \in \mathcal{P}_p(\mathbb{R}^d)$.

Proof. Take $\eta := t \longmapsto (Kt)^p$, where $K > 0$ is provided by the norm equivalence inequality $\|\cdot\| \geq K\|\cdot\|_2$. □

B.I.2.3 Function Class Example: Gradients of Convex Functions

An interesting class of functions G to optimise over is the set of L-Lipschitz functions that are gradients of (ℓ-strongly) convex functions. Indeed, this can be seen as a regularising assumption, and was studied in [PdC20] for the cost $c(x, y) = \|x-y\|_2^2$. We shall see in Proposition B.I.2 that classes of such functions on *arc-connected* partitions verify the conditions of our existence result Theorem B.I.2. In particular, [PdC20, Definition 1] (which states existence, with a simplified proof due to lack of space) is a consequence of Theorem B.I.2. Before this result, we will present a technical lemma on arc-connectedness. In this chapter, we will say that a set $A \subset \mathbb{R}^d$ is *arc-connected* if any pair of points of A can be joined by a Lipschitz curve contained in A.

Lemma B.I.3. Let $\mathcal{O}$ be an arc-connected open set of $\mathbb{R}^d$. There exists $(C_k)_{k\in\mathbb{N}}$ a sequence of arc-connected compact sets such that $\forall k \in \mathbb{N},\ C_k \subset C_{k+1}$ and $\bigcup_{k\in\mathbb{N}} C_k = \mathcal{O}$.

Proof. Consider the collection $(\overline{B}(q,r_q))_{q\in\mathcal{O}\cap\mathbb{Q}^d}$ where for each $q\in\mathcal{O}\cap\mathbb{Q}^d$, we take $r_q>0$ such that $\overline{B}(q,r_q)\subset\mathcal{O}$. Using a bijection between $\mathbb{N}$ and $\mathcal{O}\cap\mathbb{Q}^d$, we can introduce sequences $(q_k)\in(\mathcal{O}\cap\mathbb{Q}^d)^{\mathbb{N}}$ and $(r_k)\in(0,+\infty)^{\mathbb{N}}$ such that the sequence of the $A_k:=\overline{B}(q_k,r_k)$ enumerates the previous collection. The sequence (A_k) is made of compact arc-connected sets and verifies $\mathcal{O}=\cup_k A_k$. We can now defined recursively the sequence (C_k) by $C_0:=A_0$ and $C_{k+1}:=C_k\cup A_{k+1}\cup w_k([0,1])$, where for $k\in\mathbb{N}$, $w_k:[0,1]\longrightarrow\mathcal{O}$ is a Lipschitz curve between q_k and q_{k+1} contained in $\mathcal{O}$ (which exists by assumption on $\mathcal{O}$). By induction, the sequence (C_k) verifies the desired properties. □

We now have the technical tools to prove that L-Lipschitz functions that are gradients of (ℓ-strongly) convex functions verifies the local convergence stability assumption of Theorem B.I.2 on partitions of $\mathbb{R}^d$.

Proposition B.I.2. Consider $\mathcal{X}:=\mathbb{R}^d$, and a partition $\mathcal{E}:=(E_i)_{[\![1,K]\!]}$, where each $\mathring{E}_i$ is arc-connected. Let $0\leq\ell\leq L$, the set of functions

$$\mathcal{F}_{\mathcal{E},L,\ell}:=\Big\{g:\mathbb{R}^d\longrightarrow\mathbb{R}^d\mid\forall i\in[\![1,K]\!],\ g|_{\mathring{E}_i}\ L-\text{Lipschitz};\\ g|_{\mathring{E}_i}=\nabla\varphi_i,\ \varphi_i\in\mathcal{C}^1(\mathring{E}_i,\mathbb{R}),\ \varphi_i\ \ell-\text{strongly convex}\Big\}$$

verifies Assumption ii') of Theorem B.I.2.

Proof. Let $i\in[\![1,K]\!]$, and define for notational convenience $\mathbb{U}:=\mathring{E}_i$. We want to show that the set of functions

$$G:=\Big\{g:\mathbb{U}\longrightarrow\mathbb{R}^d\ L-\text{Lipschitz}\mid g=\nabla\varphi,\ \varphi\in\mathcal{C}^1(\mathbb{U},\mathbb{R}),\ \varphi\ \ell-\text{strongly convex}\Big\}$$

is closed for the compact-open topology (Definition B.I.1). By Lemma B.I.3, since $\mathbb{U}$ is open and arc-connected, we can choose an increasing sequence of arc-connected compact sets $\mathcal{K}_m\subset\mathbb{U}$ such that $\cup_m\mathcal{K}_m=\mathbb{U}$. We fix $a\in\mathcal{K}_0$.

Take a sequence $(g_n)_{n\in\mathbb{N}}\in G^{\mathbb{N}}$ such that for each $m\in\mathbb{N}$, $g_n|_{\mathcal{K}_m}$ converges uniformly to some function $h_m\in\mathcal{C}^0(\mathcal{K}_m,\mathbb{R}^d)$. We will show that there exists $g\in G$ that coincides with h_m on each $\mathcal{K}_m$. Regarding the Lipschitz constraint, by point-wise convergence, each function h_m is L-Lipschitz.

For any $n\in\mathbb{N}$, since $g_n\in G$, we can introduce an ℓ-strongly convex function $\varphi_n\in\mathcal{C}^1(\mathbb{U},\mathbb{R})$ such that $g_n=\nabla\varphi_n$. Since φ_n can be chosen up to an additive constant, we can assume $\varphi_n(a)=0$. We study the point-wise convergence of (φ_n) on $\mathcal{K}_m$ for $m\in\mathbb{N}$ fixed, so we fix $x\in\mathcal{K}_m$. Since $\mathcal{K}_m$ is arc-connected, we can choose $w:[0,1]\longrightarrow\mathcal{K}_m$ a Lipschitz curve such that $w(0)=a$ and $w(1)=x$. Noticing that for almost-every $t\in[0,1]$, $\frac{\mathrm{d}}{\mathrm{d}t}\varphi_n(w(t))=\langle\nabla\varphi_n(w(t)),\dot{w}(t)\rangle$ and using $\varphi_n(a)=0$, we write (by absolute continuity of $\varphi_n\circ w$):

$$\varphi_n(x)=\int_0^1\langle\nabla\varphi_n(w(t)),\dot{w}(t)\rangle\mathrm{d}t\xrightarrow[n\longrightarrow+\infty]{}\int_0^1\langle h_m(w(t)),\dot{w}(t)\rangle\mathrm{d}t=:\psi_m(x),$$

where the convergence is obtained by the dominated convergence theorem.

Our objective is now to prove that ψ_m is $\mathcal{C}^1$-smooth on $\mathring{\mathcal{K}}_m$, and that $\nabla\psi_m=h_m$. Let $x\in\mathring{\mathcal{K}}_m$, $v\in\mathbb{R}^d$ and $\delta>0$ such that $\forall t\in[-\delta,\delta]$, $x+tv\in\mathring{\mathcal{K}}_m$. For $n\in\mathbb{N}$ and $t\in[-\delta,\delta]$, let $f_n(t):=\varphi_n(x+tv)$. We have shown that the sequence (f_n) converges pointwise to $f:=t\longmapsto\psi_m(x+tv)$. Furthermore, by convergence of (g_n), the derivative sequence $f_n'=t\longmapsto\langle\nabla\varphi_n(x+tv),v\rangle$ converges uniformly on $[-\delta,\delta]$ to $t\longmapsto\langle h_m(x+tv),v\rangle$. A standard calculus theorem then shows that f is differentiable on $(-\delta,\delta)$, with $f'(t)=\frac{\mathrm{d}}{\mathrm{d}t}\langle h_m(x+tv),v\rangle$. In particular, by setting $t=0$ we have shown that the directional derivative $D_v\psi_m(x)$ exists and has the value $\langle h_m(x),v\rangle$. Since h_m is continuous (we saw that it is Lipschitz), this shows that ψ_m is of class $\mathcal{C}^1$, with $\nabla\psi_m=g_m$ on $\mathring{\mathcal{K}}_m$.

For $x\in\mathbb{U}$, letting $m:=\min\{m\in\mathbb{N}:x\in\mathcal{K}_m\}$, we define $\psi(x):=\psi_m(x)$, which is well-defined since $x\in\mathcal{K}_m$. For $m<m'$, since $\mathcal{K}_m\subset\mathcal{K}_{m'}$, we have $\psi_{m'}|_{\mathcal{K}_m}=\psi_m$, as a consequence, for any

$m \in \mathbb{N}$, $\psi|_{\mathcal{K}_m} = \psi_m$ without ambiguity. The previous result implies in particular that ψ is of class $\mathcal{C}^1$ on each $\mathring{\mathcal{K}}_m$, and thus everywhere on $\mathbb{U}$. We define $g : \mathbb{U} \longrightarrow \mathbb{R}^d$ similarly, with the same property $g|_{\mathcal{K}_m} = h_m$. With this construction, on each $\mathring{\mathcal{K}}_m$, one has $g = g_m = \nabla\psi_m = \nabla\psi$. As a result, we have $g = \nabla\psi$ on all of $\mathbb{U}$. Since each g_m is L-Lipschitz, it follows that g is L-Lipschitz on $\mathbb{U}$.

To see that $g \in G$, it only remains to show that ψ is ℓ-strongly convex, which is a consequence of the fact that it is everywhere a point-wise limit of a ψ_m, which is itself ℓ-strongly convex. □

B.I.2.4 Function Class Example: Neural Networks

Another natural idea is to consider classes G of parametrised functions, in particular Neural Networks (NNs) with Lipschitz activation functions. We will consider a relatively general expression for NNs borrowed from Section A.III.7.4. We consider a class G_{NN} of functions $g_\theta = h_N(\theta, \cdot) : \mathbb{R}^k \longrightarrow \mathbb{R}^d$ for a parameter vector $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is a compact set, and where h_N is the N-th layer of a recursive NN structure defined by

$$h_0(\theta, x) = x, \quad \forall n \in [\![1, N]\!],\ h_n = \left\{ \begin{array}{ccc} \mathbb{R}^p \times \mathbb{R}^k & \longrightarrow & \mathbb{R}^{d_n} \\ (\theta, x) & \longmapsto & a_n\left(\sum_{i=0}^{n-1} A_{n,i}(\theta) h_i(\theta, x) + b_n\theta\right) \end{array} \right., \tag{B.I.9}$$

$$N \in \mathbb{N},\ d_0 = k,\ d_N = d,\ \forall n \in [\![1, N]\!],\ d_n \in \mathbb{N}^*,$$

$$a_n : \mathbb{R}^{d_n} \longrightarrow \mathbb{R}^{d_n} \text{ Lipschitz},\ b_n \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^{d_n}),\ \forall i \in [\![0, n-1]\!],\ A_{n,i} \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^{d_n \times d_i}),$$

where $\mathcal{L}(A, B)$ is the space of linear maps from A to B. The terms $A_{n,i}$ and b_n correspond to the weights matrices and biases respectively, and we allow the use of the entire parameter vector $\theta \in \Theta \subset \mathbb{R}^p$ at each layer for generality. The summation over the previous layers allows the inclusion of "skip-connections" in the architecture. Thanks to the assumption that the parameters lie in a compact set, we will show that the class G_{NN} verifies the conditions of our existence result in Theorem B.I.1.

Proposition B.I.3. Let $\Theta \subset \mathbb{R}^p$ be a compact set and G_{NN} the class of functions $\mathbb{R}^p \longrightarrow \mathbb{R}^d$ of the form $g_\theta = h_N(\theta, \cdot)$, with $\theta \in \Theta$ and h_N as in Eq. (B.I.9). Then G_{NN} verifies Assumption ii) of Theorem B.I.1.

Proof. An immediate induction over the layers shows that for $g_\theta \in G_{\mathrm{NN}}$, there exists a constant $L > 0$ independent of θ such that g_θ is L-Lipschitz on $\mathbb{R}^k$.
Concerning closedness for the compact-open topology (Definition B.I.1), we will show the following stronger property: if $(g_m) \in (G_{\mathrm{NN}})^{\mathbb{N}}$ converges pointwise towards a function $f : \mathbb{R}^k \longrightarrow \mathbb{R}^d$, then there exists $\theta \in \Theta$ such that $f = g_\theta$. For $m \in \mathbb{N}$, we can write $g_m = g_{u_m}$ for $u_m \in \Theta$. Since the sequence (u_m) lies in the compact set Θ, there exists a converging subsequence $(u_{\alpha(m)})$ which converges towards $\theta \in \Theta$. Let $x \in \mathbb{R}^k$, we have the convergence $g_{u_m}(x) \longrightarrow f(x)$. By induction over the layers, the function $v \longmapsto g_v(x)$ is continuous, thus $g_{u_{\alpha(m)}}(x) \longrightarrow g_\theta(x)$. By uniqueness of the limit, $f(x) = g_\theta(x)$, and since $x \in \mathbb{R}^k$ was chosen arbitrarily, we conclude $f \in G$. □

Remark B.I.3. For simplicity, we presented NNs taking $x \in \mathbb{R}^k$ as input, yet the theory holds if $\mathcal{X}$ is a locally compact Polish space, just as in Theorem B.I.1. For instance, one could take a Riemannian manifold.

B.I.2.5 On the Necessity of the Lipschitz Constraint for Existence

Beyond the theoretical usefulness of the constraint that g be L-Lipschitz, this constraint may add substantial difficulty to the numerical implementation (see Section B.I.4). As a result, one could consider the map problem Eq. (B.I.4) without the Lipschitz assumption on G. Unfortunately, this variant has no solution in general. We illustrate this in the light of the class of functions

$\mathcal{F}_{\mathcal{E},L,\ell}$ introduced in Proposition B.I.2, in the 1D case and consider G the cone of continuous non-decreasing functions, yielding the problem:

$$\underset{g\in\mathcal{C}^0(\mathbb{R}),\ \text{non-decreasing}}{\operatorname{argmin}} \quad \mathrm{W}_2^2(g\#\mu,\nu), \tag{B.I.10}$$

where we choose the specific measures $\mu := \mathcal{U}([-1,1])$ and $\nu := \frac{1}{2}\mathcal{U}([-2,-1]) + \frac{1}{2}\mathcal{U}([1,2])$. In this setting, no continuous function $g : \mathbb{R} \longrightarrow \mathbb{R}$ can satisfy $g\#\mu = \nu$. Indeed, suppose that such a continuous function g were to exist. On the one hand, since g is continuous, $\mathrm{supp}(g\#\mu) = g(\mathrm{supp}(\mu)) = g([-1,1])$. On the other hand, by assumption $\mathrm{supp}(g\#\mu) = \mathrm{supp}(\nu) = [-2,-1] \cup [1,2]$. However, since g is continuous and $[-1,1]$ is connected, $g([-1,1])$ is also connected, thus $[-2,-1] \cup [1,2]$ is connected, which is a contradiction.

We now consider a specific function g which satisfies $g\#\mu = \nu$:

$$g := \left\{ \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \left\{ \begin{array}{c} x-1 \text{ if } x<0 \\ 0 \text{ if } x=0 \\ x+1 \text{ if } x>0 \end{array} \right. \end{array} \right. , \tag{B.I.11}$$

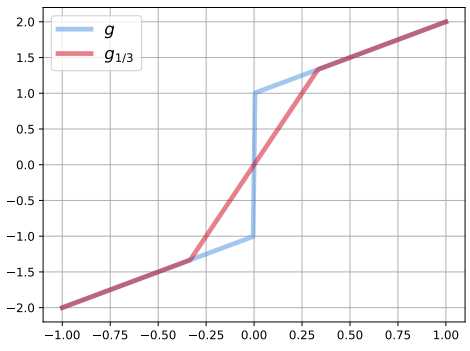
note that the value at 0 can be chosen arbitrarily. This function is not continuous, so we approach it by functions g_ε, with $\varepsilon \in (0,1)$, which are continuous and non-decreasing:

$$g_\varepsilon := \left\{ \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \left\{ \begin{array}{c} x-1 \text{ if } x \leq -\varepsilon \\ \frac{1+\varepsilon}{\varepsilon} x \text{ if } x \in [-\varepsilon,\varepsilon] \\ x+1 \text{ if } x \geq \varepsilon \end{array} \right. \end{array} \right. . \tag{B.I.12}$$

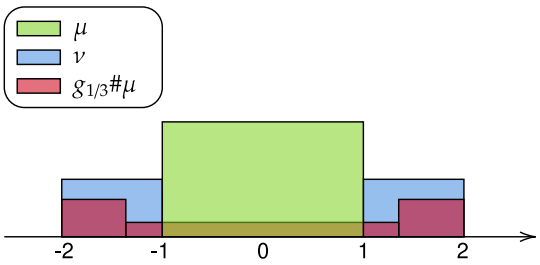
A straightforward computation yields:

$$g_\varepsilon\#\mu = \frac{1-\varepsilon}{2}\mathcal{U}([-2,-1-\varepsilon]) + \varepsilon\mathcal{U}([-1-\varepsilon,1+\varepsilon]) + \frac{1-\varepsilon}{2}\mathcal{U}([1+\varepsilon,2]), \tag{B.I.13}$$

which we illustrate in Fig. B.I.2. It follows that $g_\varepsilon\#\mu$ converges weakly towards ν as $\varepsilon \longrightarrow 0$. As



(a) Illustration of the maps g from Eq. (B.I.11) and g_ε from Eq. (B.I.12) with $\varepsilon = 1/3$.



(b) Illustration of the image measure $g_{1/3}\#\mu$ with $\mu = \mathcal{U}([-1,1])$ and g_ε from Eq. (B.I.12).

Figure B.I.2: Illustration of the counter-example to existence.

a result, since the measures are compactly supported, $\mathrm{W}_2^2(g_\varepsilon\#\mu,\nu) \xrightarrow[\varepsilon\longrightarrow 0]{} 0$, thus the value of Problem Eq. (B.I.10) is 0. To conclude, if Problem Eq. (B.I.10) had a solution g, then it would be continuous and verify $\mathrm{W}_2^2(g\#\mu,\nu) = 0$ (since the problem value is 0), thus $g\#\mu = \nu$, which is impossible by the connectivity argument. Therefore, the problem defined in Eq. (B.I.10) does not have a solution.

B.I.2.6 Discussion on Uniqueness

A natural question is the uniqueness of a solution of the problem

$$\operatorname*{argmin}_{g \in G} \mathcal{T}_c(g\#\mu, \nu),$$

in the case where the measures, the cost and the class G satisfy the conditions of Theorem B.I.1, guaranteeing existence. A first negative answer concerns the simple case where μ, ν are discrete and at least two-dimensional. For instance, consider

$$\mu := \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)}), \quad \nu := \frac{1}{2}(\delta_{(0,-1)} + \delta_{(0,1)}).$$

Then there are two distinct maps g_1, g_2 both verifying $g_i\#\mu = \nu$, which are characterised in $L^2(\mu)$ by their values on the two points $(\pm 1, 0)$.

$$g_1((-1,0)) = (0,-1),\ g_1((1,0)) = (0,1), \quad g_2((-1,0)) = (0,1),\ g_2((1,0)) = (0,-1),$$

as we illustrate in Fig. B.I.3. The previous example illustrates a potential issue for uniqueness,

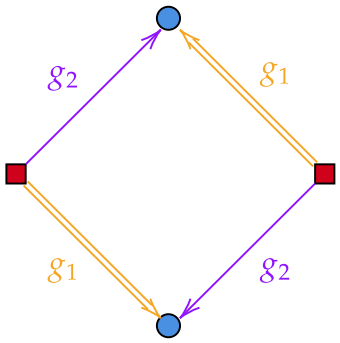


Figure B.I.3: A simple case with two transport maps between 2-point discrete measures in $\mathbb{R}^2$.

which is the multiplicity of the set $\{g \in G \mid g\#\mu = \nu\}$. Another simple counter-example to uniqueness which stems from this property is for $\mu = \nu = \mathcal{N}(0, I)$ the standard d-variate Gaussian distribution. In this case, any rotation R verifies $R\#\mathcal{N}(0, I) = \mathcal{N}(0, I)$. More generally, Brenier's polar factorisation theorem [Bre91] sheds a light on our invariance issue. We present the theorem below for completeness, see also [San15, Section 1.7.2].

Theorem B.I.3 (Brenier's Polar Factorisation [Bre91])**.** Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact set, and $g : \mathcal{K} \longrightarrow \mathbb{R}^d$. Consider $\mathbb{U}_\mathcal{K}$ the probability measure that is the uniform distribution on $\mathcal{K}$, suppose that $g\#\mathbb{U}_\mathcal{K} \ll \mathscr{L}$, then there exists a unique ($\mathscr{L}$-almost-everywhere) decomposition $g = (\nabla\varphi) \circ s$ such that:

- $\varphi : \mathcal{K} \longrightarrow \mathbb{R}^d$ is convex;
- $s : \mathcal{K} \longrightarrow \mathcal{K}$ is measure-preserving, which is to say that $s\#\mathbb{U}_\mathcal{K} = \mathbb{U}_\mathcal{K}$.

To fix the ideas, if we consider $\mu = \mathbb{U}_{[0,1]^d}$, we can fix $g \in G$ and assume $g\#\mathbb{U}_\mathcal{K} \ll \mathscr{L}$ (see sufficient conditions for this in Lemma B.I.6 in the Appendix), then decompose $g = \nabla\varphi \circ s$. Then any map h of the form $\nabla\varphi \circ r$ with r a measure-preserving map will verify $h\#\mathbb{U}_{[0,1]^d} = g\#\mathbb{U}_{[0,1]^d}$. To avoid such potential counter-examples, we will focus on the case where G is a subset of gradients of convex functions.

We provide a uniqueness result for the W_2 case, under the simplifying assumption that $\nu = \mu$. Note that if $L < 1$, the identity map does not belong to G, and there does not exist a $g \in G$ such that $g\#\mu = \mu$.

Proposition B.I.4. Suppose that

$$G = \left\{ g : \mathbb{R}^d \longrightarrow \mathbb{R}^d : g = \nabla\varphi \ \mathscr{L} - \text{a.e.}, \ \varphi \text{ convex}, \ g \ L - \text{Lipschitz} \right\},$$

and that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \mathscr{L}$. Then if g_0 and g_1 are solutions of the problem

$$\operatorname*{argmin}_{g \in G} \mathrm{W}_2^2(g\#\mu, \mu),$$

then $g_0 = g_1$ everywhere on $\mathrm{supp}(\mu)$.

Proof. We will show that if g_0 and g_1 are solutions, then $g_0\#\mu = g_1\#\mu$. First, one may write $g_i = \nabla\varphi_i$ with φ_i convex (for $i = 0, 1$). By [San15, Theorem 1.48], since φ_i is convex, g_i is the optimal transport map between μ and $\nabla\varphi_i\#\mu$. Consider for $t \in [0,1]$ the interpolation $g_t := (1-t)g_0 + tg_1$. Then by definition (see [AGS05, Section 9.2]), the curve $(g_t\#\mu)_{t\in[0,1]}$ is a[4] generalised geodesic between $g_0\#\mu$ and $g_1\#\mu$ with respect to the base measure μ. This allows us to apply [AGS05, Lemma 9.2.1], specifically [AGS05, Equation 9.2.7c], which yields

$$\forall t \in [0,1], \ \mathrm{W}_2^2(g_t\#\mu, \mu) \leq (1-t)\mathrm{W}_2^2(g_0\#\mu, \mu) + t\mathrm{W}_2^2(g_1\#\mu, \mu) - t(1-t)\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu).$$

The curvature of this generalised geodesic will allow us to build a better solution if $g_0\#\mu \neq g_1\#\mu$, as we illustrate in Fig. B.I.4.

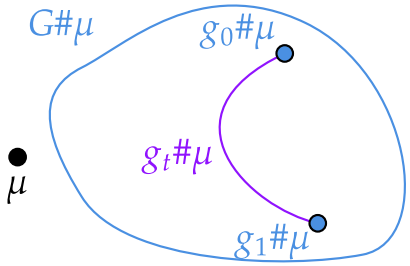


Figure B.I.4: The generalised geodesic based on μ between $g_0\#\mu$ and $g_1\#\mu$.

Taking $t = 1/2$ yields, using the optimality of g_0 and g_1 and writing v for the problem value:

$$\mathrm{W}_2^2(g_{\frac{1}{2}}\#\mu, \mu) \leq v - \frac{1}{4}\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu).$$

Since G is convex, we have $g_{\frac{1}{2}} \in G$, which imposes $\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu) = 0$, since v is the optimal problem value. We conclude $g_0\#\mu = g_1\#\mu$. However, as stated earlier, by [San15, Theorem 1.48], g_i is the optimal transport map between μ and $g_i\#\mu$ for $i = 0, 1$. By uniqueness of the optimal transport map in this setting, we conclude $g_0 = g_1$. (The equality holds μ-a.e., then since g_0 and g_1 are assumed Lipschitz, this shows equality everywhere on $\mathrm{supp}(\mu)$.) □

Remark B.I.4. One could replace the set G in Proposition B.I.4 by a convex subset of G, the proof of the result would follow verbatim.

Remark B.I.5. The problem in Proposition B.I.4 is related to the problem of the Wasserstein metric projection, which was studied in [De +16, Section 5], from which the curvature argument in our proof was closely inspired. This Wasserstein projection problem was also studied for W_p^p in [AM24].

[4] In this case, since $\mu \ll \mathscr{L}$, there is even uniqueness of *the* generalised geodesic between $g_0\#\mu$ and $g_1\#\mu$, but we do not use that fact.

Remark B.I.6. Under some assumptions, it may be possible to find subclasses of gradients of convex functions G such that the set $G\#\mu \subset \mathcal{P}_2(\mathbb{R}^d)$ is geodesically convex (with respect to W_2 geodesics): take $g_0, g_1 \in G$, assume that $g_0\#\mu \ll \mathscr{L}$ (Lemma B.I.6 provides a sufficient condition on g_0 and μ for this to be the case). Then the W_2 geodesic from $g_0\#\mu$ to $g_1\#\mu$ is

$$\nu_t := ((1-t)I + tT)\#g\#\mu_0,$$

where T is the optimal transport map from $g_0\#\mu$ to $g_1\#\mu$, which is uniquely defined thanks to Brenier's Theorem (see [San15], for a possible reference without compactness assumptions). Since $(T \circ g_0)\#\mu = g_1\#\mu$, under some regularity assumptions, it may be possible to show that $T \circ g_0 = g_1$ using the Monge-Ampère equation, then $((1-t)I + tT) \circ g_0 = (1-t)g_0 + tg_1 \in G$. In this case, the generalised geodesic based on μ coincides with the W_2 geodesic between $g_0\#\mu$ and $g_1\#\mu$.
Unfortunately, $\rho \longmapsto \mathrm{W}_2^2(\rho, \nu)$ is not convex along W_2 geodesics, since it satisfies the opposite inequality ([AGS05, Theorem 7.3.2]). As a result, even if we found a convex class G of gradients of convex functions such that $G\#\mu$ were geodesically convex, curvature arguments would not yield uniqueness immediately. Intuition suggests that in some sense, the problem minimises a concave function over a convex set, which bodes poorly with uniqueness.

In Section B.I.3.2, we shall study the case $d = 1$ and show uniqueness and an explicit expression for the minimiser of the map problem for non-decreasing functions g and the squared Euclidean cost. To conclude this discussion, even for the favourable case where $\mu \ll \mathscr{L}^d$, G is a subset of gradients of convex functions and $c(x, y) = \|x - y\|_2^2$, we conjecture that uniqueness is not guaranteed in general for $d \geq 2$.

B.I.2.7 The Plan Approximation Problem

In some cases, one may have access to a transport plan between two measures μ, ν, which poses the natural question of finding a map that approximates this transport plan. For instance, one may compute the optimal entropic plan [Cut13], a Gaussian-Mixture-Model optimal plan [DD20], or an optimal transport plan for a cost that does not verify the twist condition (see [San15, Definition 1.16]), or more generally an optimal plan when the Monge problem is not equivalent to the Kantorovich problem (see [Bre91; GM96; Pra07] for some known equivalence cases).

Given a cost $C : (\mathbb{R}^k \times \mathbb{R}^d) \times (\mathbb{R}^k \times \mathbb{R}^d) \longrightarrow \mathbb{R}_+$, and measures $\mu \in \mathcal{P}(\mathbb{R}^k), \nu \in \mathcal{P}(\mathbb{R}^d)$, we will want to approximate a plan $\gamma \in \Pi(\mu, \nu)$ by the image measure $(I, g)\#\mu$, where I denotes the identity map of $\mathbb{R}^k$. We define the Constrained Approximate Transport Plan problem as:

$$\operatorname*{argmin}_{g \in G} \mathcal{T}_C((I, g)\#\mu, \gamma). \tag{B.I.14}$$

Similarly to Eq. (B.I.4), the transport cost in Eq. (B.I.14) can be re-written using the change-of-variables formula (Lemma B.I.1):

$$\mathcal{T}_C((I, g)\#\mu, \gamma) = \min_{\rho \in \Pi(\mu, \gamma)} \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} C\left((x, g(x)), (y_1, y_2)\right) \mathrm{d}\rho(x, y_1, y_2). \tag{B.I.15}$$

To begin with, one may cast Eq. (B.I.14) as a map problem (Eq. (B.I.4)), providing existence automatically under adequate conditions.

Corollary B.I.1. Consider the class of functions

$$\widetilde{G} := \{\tilde{g} : \mathbb{R}^k \longrightarrow \mathbb{R}^k \times \mathbb{R}^d \ : \ \tilde{g} = (x, y) \longmapsto (x, g(y)),\ g \in G\},$$

the map problem (Eq. (B.I.4)) is a particular map problem (Eq. (B.I.14)):

$$\min_{g \in G} \mathcal{T}_C((I, g)\#\mu, \gamma) = \min_{\tilde{g} \in \widetilde{G}} \mathcal{T}_C(\tilde{g}\#\mu, \gamma),$$

hence existence holds by Theorem B.I.1 if the conditions of the theorem are verified by $C, \widetilde{G}$ and the measures μ, γ.

Remark B.I.7. In the light of Remark B.I.2, one could replace the input space $\mathbb{R}^k$ and the target space $\mathbb{R}^d$ by Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ verifying the Heine-Borel property, in which case condition 1) would ask for $(x_1, x_2) \longmapsto C((x_1, x_2), (y_1, y_2))$ to be proper.

We shall see that in certain cases, the two problems Eq. (B.I.14) and Eq. (B.I.4) are in fact equivalent.

Proposition B.I.5. Consider a cost C of the separable form $C((x_1, x_2), (y_1, y_2)) = h(c_1(x_1, y_1), c_2(x_2, y_2))$, where $h : \mathbb{R}_+ \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, $c_1 : \mathbb{R}^k \times \mathbb{R}^k \longrightarrow \mathbb{R}_+$ and $c_2 : \mathbb{R}^d \times \mathbb{R}^d$ are lower semi-continuous, with $\forall x \in \mathbb{R}^k$, $c_1(x, x) = 0$, and $\forall u, v \in \mathbb{R}_+$, $h(u, v) \geq v$ and $h(0, v) = v$. Let $g : \mathbb{R}^k \longrightarrow \mathbb{R}^d$ be a measurable function, $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $\mu \in \mathcal{P}(\mathbb{R}^k)$. Let $\gamma \in \Pi(\mu, \nu)$ be a plan between μ and ν.
We assume that the value $\mathcal{T}_C((I, g)\#\mu, \gamma)$ is finite. We have the equality

$$\mathcal{T}_{c_2}(g\#\mu, \nu) = \mathcal{T}_C((I, g)\#\mu, \gamma).$$

Proof. For $\rho \in \Pi(\mu, \gamma)$, let $A(\rho) := \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} C((x, g(x)), (y_1, y_2)) \mathrm{d}\rho(x, y_1, y_2) < +\infty$, and denote $A^* := \mathcal{T}_C((I, g)\#\mu, \gamma)$. Likewise, for $\pi \in \Pi(\mu, \nu)$, let $B(\pi) := \int_{\mathbb{R}^k \times \mathbb{R}^d} c_2(g(x), y) \mathrm{d}\pi(x, y)$, and $B^* := \mathcal{T}_{c_2}(g\#\mu, \nu)$.

First, we prove $A^* \leq B^*$. By [San15, Theorem 1.7], there exists $\pi^* \in \Pi(\mu, \nu)$ such that $B^* = B(\pi^*)$. Define $\rho \in \Pi(\mu, \gamma)$ a measure such that for each test function f,

$$\int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} f(x, y_1, y_2) \mathrm{d}\rho(x, y_1, y_2) = \int_{\mathbb{R}^k \times \mathbb{R}^d} f(y_1, y_1, y_2) \mathrm{d}\pi^*(y_1, y_2),$$

or symbolically "$\rho(\mathrm{d}x\mathrm{d}y_1\mathrm{d}y_2) = \delta_{y_1}(\mathrm{d}x)\pi^*(\mathrm{d}y_1\mathrm{d}y_2)$". Then, since $h(c_1(y_1, y_1), c_2(g(y_1), y_2)) = c_2(g(y_1), y_2)$, we have

$$A^* \leq A(\rho) = \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} h(c_1(y_1, y_1), c_2(g(y_1), y_2)) \mathrm{d}\pi^*(y_1, y_2) = \int_{\mathbb{R}^k \times \mathbb{R}^d} c_2(g(y_1), y_2) \mathrm{d}\pi^*(y_1, y_2) = B^*.$$

Now for $A^* \geq B^*$, we let $\rho \in \Pi(\mu, \gamma)$. Using $h(u, v) \geq v$, we have

$$A(\rho) = \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} h(c_1(x, y_1), c_2(g(x), y_2)) \mathrm{d}\rho(x, y_1, y_2) \geq \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} c_2(g(x), y_2) \mathrm{d}\rho(x, y_1, y_2).$$

Again, we can define $\pi \in \Pi(\mu, \nu)$ such that for any test function f,

$$\int_{\mathbb{R}^k \times \mathbb{R}^d} f(x, y_2) \mathrm{d}\pi(x, y_2) = \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} f(x, y_2) \mathrm{d}\rho(x, y_1, y_2),$$

and notice

$$\int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} c_2(g(x), y_2) \mathrm{d}\rho(x, y_1, y_2) = B(\pi) \geq B^*,$$

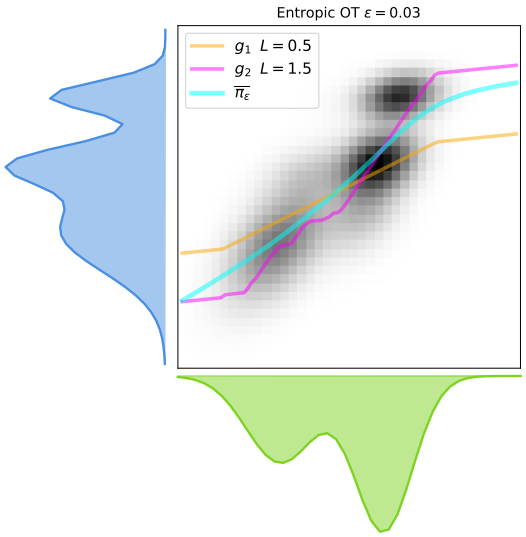
which yields $A^* \geq B^*$. □

For example, the cost $C(x, y) = \|x - y\|_2^2$ satisfies these conditions (with $h(u, v) = u + v$), and thus the problems Eq. (B.I.4) and Eq. (B.I.14) are equivalent. This is still the case for costs of the form $C = \| \cdot - \cdot \|_p^{qp}$ for $p \geq 1$ and $q > 0$, in which case one takes $h(u, v) = (u^{1/q} + v^{1/q})^q$. For $C((x_1, x_2), (y_1, y_2)) = \|(x_1, x_2) - (y_1, y_2)\|_\infty^p$, this is also the case with $c_1(x, y) = c_2(x, y) = \|x - y\|_\infty^p$ and $h(u, v) = \max(u, v)$.

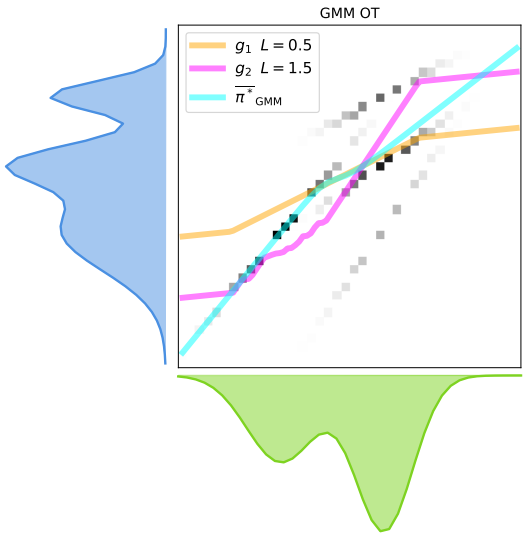
In contrast, a possible choice of norm on the product space is $\|x\|_\Sigma = (x^\top \Sigma^{-1} x)^{1/2}$ for Σ symmetric positive-definite. This choice is of interest since the cost $C((x_1, x_2), (y_1, y_2)) =$

$\|(x_1, x_2) - (y_1, y_2)\|^2_\Sigma$ is quadratic (which is desirable for numerics), but does not satisfy the equivalence condition from Proposition B.I.5 as soon as Σ is not block-diagonal.

In Fig. B.I.5, we illustrate the plan approximation problem for the quadratic cost for two different plans: the Entropic Optimal Transport plan [PC19b] and the Gaussian Mixture Model OT plan [DD20]. The numerics where done using the tools presented in Section B.I.4.2. Note that the plan approximation is equivalent to a map problem in this case, and has a particular structure due to the one-dimensional setting, hence we emphasise that Fig. B.I.5 is merely an illustration of the problem at hand.



(a) plan approximation solutions for the Entropic-OT plan [Cut13].



(b) Illustration of plan approximation solutions for the GMM-OT plan [DD20].

Figure B.I.5: Illustration of solutions of plan approximation problems (Eq. (B.I.14)), for two different plans between Gaussian Mixtures. We compare the plans with $L = 1/2$ and $L = 3/2$-Lipschitz solutions, as well as to the barycentric projection of the given plans (see Section B.I.3.1).

B.I.3 Alternate Minimisation in the Squared Euclidean Case

The map problem Eq. (B.I.4) is a minimisation problem over $\pi \in \Pi(\mu, \nu)$ and $g \in G$:

$$\min_{g \in G} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathbb{R}^d} c(g(x), y) \mathrm{d}\pi(x, y).$$

In this section, we study this alternate minimisation problem in the case where $c(x, y) = \|x - y\|_2^2$ and $\mathcal{X} = \mathbb{R}^d$, thus with maps $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

When $\pi \in \Pi(\mu, \nu)$ is fixed, the sub-problem has the particular structure

$$\min_{g \in G} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y). \tag{B.I.16}$$

To ensure the finiteness of the cost, we assume that $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. We shall see in Section B.I.3.1 that the problem in Eq. (B.I.16) is equivalent to the L^2 projection of the barycentric map $\overline{\pi}$ onto the set G, provided that G is a convex and closed subset of $L^2(\mu)$.

When $g \in G$ is fixed, the problem reads

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y), \tag{B.I.17}$$

and can be seen from two different viewpoints: either as the squared Euclidean optimal transport problem between $g\#\mu$ and ν (i.e. $\mathrm{W}_2^2(g\#\mu, \nu)$), or as the optimal transport problem with cost

$c(x, y) := \|g(x) - y\|_2^2$ between μ and ν. If $g\#\mu$ is absolutely continuous and ν is discrete, then Eq. (B.I.17) is a semi-discrete OT problem. We provide sufficient conditions on g for this to be the case in Section B.I.6.1.

This alternate minimisation viewpoint poses a natural question: if $\pi := \pi^* \in \Pi^*(\mu, \nu)$ is an optimal plan between μ and ν for the quadratic cost $c(y, y') := \|y - y'\|_2^2$, does the following equality holds?

$$\operatorname*{argmin}_{g \in G} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y) \stackrel{?}{=} \operatorname*{argmin}_{g \in G} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi^*(x, y). \quad \text{(B.I.18)}$$

In Section B.I.3.2, we prove that this equality holds in the one-dimensional case $\mathcal{X} = \mathbb{R}^d = \mathbb{R}$ and if G is a subclass of non-decreasing functions, thus generalizing a result of [PdC20]. We also provide a counter-example of this property when $d \geq 2$ in Section B.I.3.3.

B.I.3.1 Projection of the Barycentric Map

In this section, we will show that the sub-problem with $\pi \in \Pi(\mu, \nu)$ fixed can be written as the following L^2 projection problem:

$$\operatorname*{argmin}_{g \in G} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y) = \operatorname*{argmin}_{g \in G} \|g - \overline{\pi}\|^2_{L^2(\mu)},$$

where $\overline{\pi}$ is the barycentric projection of π (defined below), and $L^2(\mu)$ is a shorthand for $L^2(\mu; \mathbb{R}^d)$, the space of measurable functions $T : \mathcal{X} \longrightarrow \mathbb{R}^d$ such that $\int_{\mathcal{X}} \|T(x)\|_2^2 \mathrm{d}\mu(x) < +\infty$. We begin by briefly introducing the notion of barycentric projection.

The *barycentric projection* of π is the map $\overline{\pi} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ defined for μ-almost all $x \in \mathbb{R}^d$ by

$$\overline{\pi}(x) = \mathbb{E}_{(X,Y)\sim\pi}[Y|X = x]. \quad \text{(B.I.19)}$$

As illustrated in Fig. B.I.6, if π admits a disintegration with respect to its first marginal μ of the form $\pi(\mathrm{d}x\mathrm{d}y) = \pi_x(\mathrm{d}y)\mu(\mathrm{d}x)$, then

$$\overline{\pi}(x) = \int_{\mathbb{R}^d} y \, \mathrm{d}\pi_x(y).$$

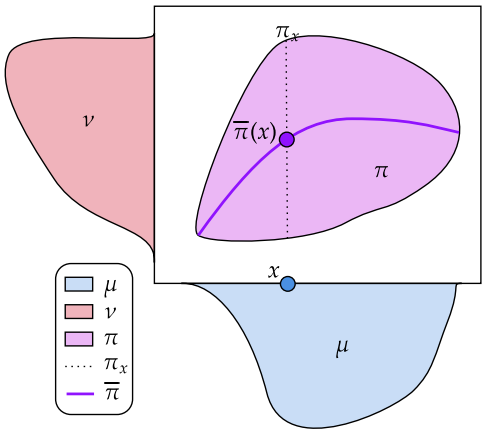


Figure B.I.6: Illustration of a barycentric projection. The disintegration of the coupling π with respect to its first marginal μ at x is the measure π_x concentrated on the dotted line. The barycentric projection of π evaluated at x is the mean of the measure π_x.

Since the conditional expectation minimises the L^2 distance, we also have

$$\overline{\pi} = \operatorname*{argmin}_{T \in L^2(\mu)} \int_{\mathbb{R}^{2d}} \|y - T(x)\|_2^2 \mathrm{d}\pi(x, y), \quad \text{(B.I.20)}$$

where the equality is to be understood in $L^2(\mu)$. Another interesting property is that if $\pi = \pi^* \in \Pi^*(\mu, \nu)$ is an optimal transport plan between μ and ν with respect to the squared Euclidean

distance cost, then by [AGS05, Section 6.2.3], there exists $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ convex such that for π^*-almost-every $(x, y) \in \mathbb{R}^{2d}$, we have $y \in \partial\varphi(x)$, where $\partial\varphi(x)$ denotes the *Fréchet sub-differential* of φ:

$$y \in \partial\varphi(x) \iff \liminf_{z \longrightarrow x} \frac{\varphi(z) - \varphi(x) - \langle y, z - x\rangle}{\|x - z\|_2} \geq 0.$$

Since the Fréchet sub-differential of a convex function is convex, it follows that for μ-almost every $x \in \mathbb{R}^d$, $\overline{\pi^*}(x) \in \partial\varphi(x)$.

If we require constraints on T in Eq. (B.I.20), we obtain exactly the sub-problem of the map problem with a fixed plan π (Eq. (B.I.16)), which we reproduce below:

$$\operatorname*{argmin}_{g \in G} \int_{\mathcal{X}\times\mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y).$$

For this reason, we call this problem the Constrained Barycentric Map problem. A consequence of the proof of Theorem B.I.1 is that this problem has a solution. If G is a convex set and closed in $L^2(\mu)$, then existence and uniqueness are guaranteed by the Hilbert projection Theorem. Since $\overline{\pi}$ minimises the L^2 distance, it is a solution of Eq. (B.I.16) if it is in G.

Using the fact that the barycentric projection is an L^2 projection (Eq. (B.I.20)), one may re-write the Projected Barycentric Map Problem Eq. (B.I.16) as an L^2 minimisation with respect to the barycentric projection. In Proposition B.I.6, we need not assume that $\mathcal{X} = \mathbb{R}^d$, but we shall apply it later in Section B.I.3.2 to the case $\mathcal{X} = \mathbb{R}$.

Proposition B.I.6. Let $\pi \in \Pi(\mu, \nu)$ and $f : \mathcal{X} \longrightarrow \mathbb{R}^d$ be a measurable function. Then one has

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|f(x)-y\|_2^2 \mathrm{d}\pi(x, y) = \int_{\mathcal{X}} \|f(x)-\overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x) + \int_{\mathcal{X}\times\mathbb{R}^d} \|y-\overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x, y), \quad \text{(B.I.21)}$$

and as a result, the Projected Barycentric Map problem Eq. (B.I.16) is equivalent to the problem

$$\operatorname*{argmin}_{g \in G} \int_{\mathcal{X}} \|g(x) - \overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x). \quad \text{(B.I.22)}$$

Moreover, the second term on the right-hand side of Eq. (B.I.21) only depends on $\overline{\pi}$ and the measures μ, ν (it doesn't depend on π). More precisely, we have

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x, y) = m_2(\nu) - m_2(\overline{\pi}\#\mu),$$

where $m_2(\rho) := \int \|x\|_2^2 \mathrm{d}\rho(x)$ for a positive measure ρ.

Proof. Denote J the left-hand-side of Eq. (B.I.21), and compute (taking the expectation under $(X, Y) \sim \pi$)

$$\begin{aligned} J &= \mathbb{E}\left[\|Y - f(X)\|_2^2\right] = \mathbb{E}\left[\|Y - \overline{\pi}(X) + \overline{\pi}(X) + f(X)\|_2^2\right] \\ &= \mathbb{E}\left[\|Y - \overline{\pi}(X)\|_2^2\right] + \mathbb{E}\left[\|\overline{\pi}(X) + f(X)\|_2^2\right] + 2\mathbb{E}\left[(Y - \overline{\pi}(X))^\top(\overline{\pi}(X) + f(X))\right], \end{aligned}$$

then since $\overline{\pi}(X)$ is the orthogonal projection of Y onto the set of random variables that are functions of X, the inner product $\mathbb{E}\left[(Y - \overline{\pi}(X))^\top(\overline{\pi}(X) + f(X))\right]$ is zero, yielding Eq. (B.I.21).

We can expand the norm in the second term of the right-hand side of Eq. (B.I.21) using $m_2(\nu)$ the second moment of ν and get

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x, y) = m_2(\nu) - 2\int_{\mathcal{X}\times\mathbb{R}^d} y \cdot \overline{\pi}(x) \mathrm{d}\pi(x, y) + \int_{\mathcal{X}} \|\overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x).$$

Writing the disintegration of π w.r.t. μ as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\mu(\mathrm{d}x)$, we re-write the second term as

$$\int_{\mathcal{X}\times\mathbb{R}^d} y \cdot \overline{\pi}(x) \mathrm{d}\pi(x, y) = \int_{\mathcal{X}} \overline{\pi}(x) \cdot \left(\int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y)\right) \mathrm{d}\mu(x) = \int_{\mathcal{X}} \overline{\pi}(x) \cdot \overline{\pi}(x) \mathrm{d}\mu(x) = m_2(\overline{\pi}\#\mu).$$

Putting our computations together yields

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x,y) = m_2(\nu) - m_2(\overline{\pi}\#\mu).$$

□

Remark B.I.8 (Ties to the Convex Least Squares Estimator [Man+24].)**.** In [Man+24], Manole et al. study the statistical properties of various estimators of Optimal Transport maps, assuming some regularity on the input distributions. Specifically, they introduce the so-called Convex Least Squares Estimator: given $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with the (x_i) being i.i.d. samples of μ and $\hat{\nu}_m := \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ with the (y_j) i.i.d. samples of ν, the estimator is defined as

$$\hat{T}_{n,m} = \nabla\hat{\varphi}, \quad \hat{\varphi} \in \operatorname*{argmin}_{\varphi\in\Phi_L} \sum_{i=1}^n \sum_{j=1}^m \hat{\pi}^*_{i,j} \|\nabla\varphi(x_i) - y_j\|_2^2, \tag{B.I.23}$$

where $\hat{\pi}^*$ is an optimal transport plan between $\hat{\mu}_n$ and $\hat{\nu}_m$, and where Φ_L is the set of $\mathcal{C}^1$ convex functions from $\Omega \subset \mathbb{R}^d$ to $\mathbb{R}$ with a L-Lipschitz gradient. Notice that Eq. (B.I.23) is a Constrained Barycentric Projection problem Eq. (B.I.16) with a specific (discrete) transport plan $\hat{\pi}^*$, chosen to be the optimal transport plan between $\hat{\mu}_n$ and $\hat{\nu}_m$, and with the particular class $G := \mathcal{F}_{\mathcal{E},L,\ell}$ (introduced in Section B.I.2.3).

B.I.3.2 Equivalence to a Constrained Barycentric Projection in Dimension 1

In this section, we shall prove that the Constrained Approximate Transport Map problem (Eq. (B.I.4)) is equivalent to the Constrained Barycentric Projection Problem (Eq. (B.I.16)) for the quadratic cost in dimension 1. This provides a positive answer to the question raised in Eq. (B.I.18) in this particular case. The idea behind this equivalence stems from the fact that in dimension one, optimal transport maps are non-decreasing, and the composition of two optimal transport maps remains an optimal transport map.

Proposition B.I.7. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, and G a subclass of the non-decreasing functions $g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $g\#\mu \in \mathcal{P}_2(\mathbb{R})$, we have the equality

$$\operatorname*{argmin}_{g\in G} \mathrm{W}_2^2(g\#\mu, \nu) = \operatorname*{argmin}_{g\in G} \|g - \overline{\pi^*}\|^2_{L^2(\mu)}, \tag{B.I.24}$$

where π^* is an optimal transport plan between μ and ν for the squared Euclidean cost.

Proposition B.I.7 generalises [PdC20, Proposition 1], which proves the same equivalence for a specific class of functions G, and assuming μ to be either discrete or absolutely continuous with respect to the Lebesgue measure.

The proof of Proposition B.I.7 hinges on Lemma B.I.4, which is intuitive in the absolutely continuous or discrete case, but a bit more technical in full generality. We write below the cumulative distribution function of a probability measure ρ as $F_\rho := x \longmapsto \rho((-\infty, x])$. Since it is non-decreasing, we can define its **right-inverse** as (using the notation $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$):

$$F_\rho^{\leftarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ F_\rho^{\leftarrow}(p) := \inf\left\{x \in \mathbb{R} \mid\ F_\rho(x) \geq p\right\}.$$

Lemma B.I.4. Let $\mu \in \mathcal{P}(\mathbb{R})$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a non-decreasing function, we have the following almost-everywhere change of variables formula for the quantile functions of $g\#\mu$ and μ:

$$F^{\leftarrow}_{g\#\mu} = g \circ F^{\leftarrow}_\mu, \quad \mathscr{L}_{[0,1]}\text{-almost-everywhere.}$$

Proof. The proof is provided in Section B.I.6.2. □

Proof of Proposition B.I.7 . Let $g \in G$. By [San15, Proposition 2.17] and by Lemma B.I.4

successively, we have

$$\mathrm{W}_2^2(g\#\mu,\nu)=\int_0^1|F^{\leftarrow}_{g\#\mu}(p)-F^{\leftarrow}_{\nu}(p)|^2\mathrm{d}p=\int_0^1|g\circ F^{\leftarrow}_{\mu}(p)-F^{\leftarrow}_{\nu}(p)|^2\mathrm{d}p=\int_{\mathbb{R}^2}|g(x)-y|^2\mathrm{d}\pi(x,y),$$

where $\pi:=(F^{\leftarrow}_{\mu},F^{\leftarrow}_{\nu})\#\mathscr{L}_{[0,1]}$, which by [San15, Theorem 2.9] is the unique optimal plan between μ and ν for the squared Euclidean cost. We apply Proposition B.I.6, which yields

$$\mathrm{W}_2^2(g\#\mu,\nu)=\int_{\mathbb{R}^2}|g(x)-y|^2\mathrm{d}\pi(x,y)=\int_{\mathbb{R}}|g(x)-\overline{\pi}(x)|^2\mathrm{d}\mu(x)+m_2(\nu)-m_2(\overline{\pi}\#\mu).$$

Given the expression of the right-hand-side above, we conclude that

$$\operatorname*{argmin}_{g\in G}\mathrm{W}_2^2(g\#\mu,\nu)=\operatorname*{argmin}_{g\in G}\|g-\overline{\pi^*}\|^2_{L^2(\mu)},$$

(for any optimal transport plan π^* between μ and ν for the squared Euclidean cost, and we have even remarked that such a plan is in fact unique) since the costs are equal up to a constant independent of g. □

B.I.3.3 Non-Equivalence to Constrained Barycentric Projection in Dimension 2

In this section, we provide a negative example to the question formulated in Eq. (B.I.18), namely that

$$\operatorname*{argmin}_{g\in G}\mathrm{W}_2^2(g\#\mu,\nu)\neq\operatorname*{argmin}_{g\in G}\|g-\overline{\pi^*}\|_{L^2(\mu)},$$

where π^* is an optimal transport plan (for the squared Euclidean cost) between μ and ν, in dimension $d\geq 2$. We take G to be the class of *monotone* continuous functions $g:\mathbb{R}^2\longrightarrow\mathbb{R}^2$, which is to say that

$$\forall x,y\in\mathbb{R}^2,\ \langle g(x)-g(y),x-y\rangle\geq 0.$$

Note that gradients of convex functions are monotone, but the converse does not hold. For $(a,b,x)\in(0,+\infty)^3$, we consider the following measures:

$$\mu:=\frac{2}{3}\delta_{(0,0)}+\frac{1}{3}\delta_{(x,0)}\text{ and }\nu:=\frac{2}{3}\delta_{(0,0)}+\frac{1}{3}\delta_{(-a,b)}.$$

There is a unique optimal transport plan π^* between μ and ν, given by

$$\pi^*=\frac{1}{3}\delta_{(0,0)\otimes(0,0)}+\frac{1}{3}\delta_{(0,0)\otimes(-a,b)}+\frac{1}{3}\delta_{(x,0)\otimes(0,0)}.$$

Its barycentric projection is characterised by the following equation

$$\overline{\pi^*}(0,0)=(-a/2,b/2)\text{ and }\overline{\pi^*}(x,0)=(0,0).$$

We now consider the problem $\min_{g\in G}\|g-\overline{\pi^*}\|_{L^2(\mu)}$. A solution of this problem is characterised by its values on the support of μ, and one may reduce the problem to an optimisation over $g(0,0)$ and $g(x,0)$, with the monotonicity constraint $\langle g(0,0)-g(x,0),(0,0)-(x,0)\rangle\geq 0$. Since $\overline{\pi^*}$ itself verifies this condition, it is the only solution (in the sense of $L^2(\mu)$). We conclude

$$\operatorname*{argmin}_{g\in G}\|g-\overline{\pi^*}\|^2_{L^2(\mu)}=\left\{\overline{\pi^*}\right\}.$$

We now show that the problem $\operatorname*{argmin}_{g\in G}\mathrm{W}_2^2(g\#\mu,\nu)$ has a different solution set. First, we compute

$$\mathrm{W}_2^2(\overline{\pi^*}\#\mu,\nu)=\frac{a^2+b^2}{6}.$$

However, if we introduce $g\in G$ such that $g(0,0)=(0,0)$ and $g(x,0)=(0,b)$, we have

$$\mathrm{W}_2^2(g\#\mu,\nu)=\frac{a^2}{3}.$$

For instance, $(a,b,x):=(1,10,1)$ yields

$$\mathrm{W}_2^2(\overline{\pi^*}\#\mu,\nu)=\frac{a^2+b^2}{6}=\frac{101}{6}>\mathrm{W}_2^2(g\#\mu,\nu)=\frac{a^2}{3}=\frac{1}{3}.$$

We illustrate the point configurations for $(a,b,x):=(1,3,1)$ in Fig. B.I.7.

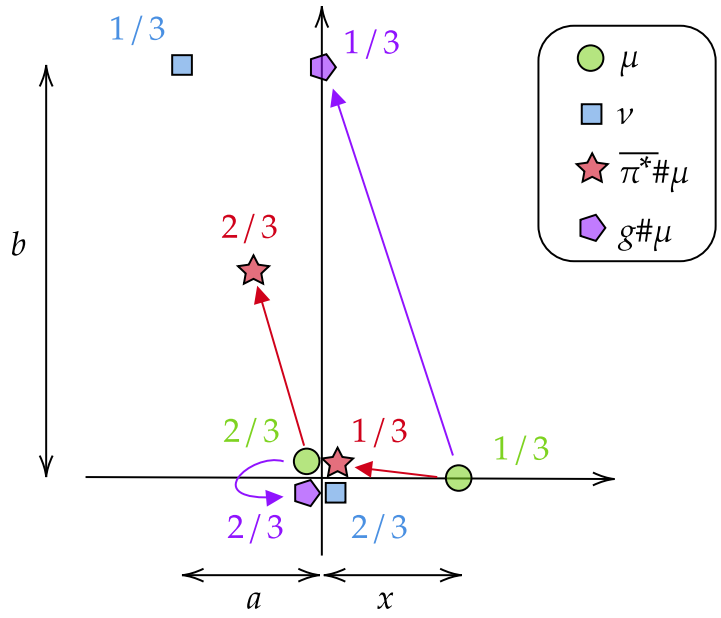


Figure B.I.7: Illustration of the two-dimensional counter-example to the equivalence of the map problem to the L^2 projection of the barycentric projection. The four points close to $(0,0)$ are represented with an offset for legibility, and represent four points equal to $(0,0)$ exactly.

B.I.4 Discrete Measures and Numerical Methods

In this section, we consider some numerical methods to solve the approximate map problem for some specific function classes. To prepare for convergence results, we dedicate Section B.I.4.1 to regularity properties of the transport cost in the discrete case. In Section B.I.4.2, we present methods in the case where G is the class of L-Lipschitz gradients of ℓ-strongly convex potentials (presented in Section B.I.2.3). For the squared Euclidean cost, these methods were introduced in [PdC20], using convex interpolation results from [Tay17]. In Section B.I.4.3, we consider a simple kernel method which solves a regularised version of Eq. (B.I.4). This type of method hinges on the fact that kernel methods yield a finite-dimensional parametrisation of the function g, and allows for provably convergent gradient descent methods. In Section B.I.4.4, we consider a Stochastic Gradient Descent method for the case where the map g is a Neural Network. Finally, in Section B.I.4.5, we illustrate the use of the methods presented in this section on the problem of colour transfer.

B.I.4.1 Regularity of Discrete Optimal Transport Costs

To study the convergence of sub-gradient descent methods theoretically, we will introduce standard notions from non-smooth non-convex analysis, in particular a specific generalisation of sub-gradients, which are in practice computed by automatic differentiation. A central notion in this analysis will be the notion of semi-algebraicity, which we remind in Definition B.I.2 (and refer to [Wak08] and [BP21] for more details).

Definition B.I.2. A set $S \subset \mathbb{R}^d$ is said to be semi-algebraic if it can be written under the form $S = \cup_{n=1}^{N} \cap_{m=1}^{M} S_{n,m}$, where each $S_{n,m}$ is either of the form $\{x \in \mathbb{R}^d : p_{n,m}(x) = 0\}$ or $\{x \in \mathbb{R}^d : p_{n,m}(x) \geq 0\}$, where $p_{n,m}$ is a d-variate polynomial with real coefficients.
A function $f : \mathbb{R}^{d_1} \longrightarrow \mathbb{R}^{d_2}$ is semi-algebraic if its graph $\{(x, f(x)) : x \in \mathbb{R}^{d_1}\}$ is a semi-algebraic set.
A multifunction $f : \mathbb{R}^{d_1} \rightrightarrows \mathbb{R}^{d_2}$ is semi-algebraic if its graph $\bigcup_{x \in \mathbb{R}^{d_1}} \{x\} \times f(x)$ is a semi-algebraic set.

Another central notion will be a generalisation of the notion of gradient for locally Lipschitz

functions called the Clarke differential.

Definition B.I.3. Given a locally Lipschitz function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, its Clarke sub-gradient at $x \in \mathbb{R}^d$ is the set

$$\partial_C f(x) = \mathrm{conv}\left\{ \lim_{t \longrightarrow +\infty} \nabla f(x_t) : x_t \xrightarrow[t \longrightarrow +\infty]{} x,\ x_t \in D_f \right\},$$

where D_f is the set of differentiability of f and conv denotes the convex hull.

In Proposition B.I.8, we show that the discrete OT cost is semi-algebraic and locally Lipschitz as a function of the cost matrix, and that its Clarke sub-gradient is itself semi-algebraic.

Proposition B.I.8. Consider weights $a \in \triangle_n$ (the n-simplex) and $b \in \triangle_m$ (the m-simplex), and the discrete Kantorovich cost function

$$W(a,b,\cdot) := \left\{ \begin{array}{ccc} \mathbb{R}^{n\times m} & \longrightarrow & \mathbb{R} \\ M & \longmapsto & \min\limits_{\pi \in \Pi(a,b)} \pi \cdot M \end{array} \right. .$$

Then the map $W(a,b,\cdot)$ is semi-algebraic, Lipschitz, and its Clarke sub-gradient is semi-algebraic and writes for $M \in \mathbb{R}^{n\times m}$:

$$\partial_C W(a,b,\cdot)(M) = \operatorname*{argmin}_{\pi \in \Pi(a,b)} \pi \cdot M.$$

Proof. Writing the extremal points of the polytope $\Pi(a,b)$ as $(\pi_i)_{i=1}^N$, the function $W(a,b,\cdot)$ can be written as a finite minimisation over $\pi \in (\pi_i)_{i=1}^N$ of linear functions of M, hence $W(a,b,\cdot)$ is semi-algebraic. The fact that $W(a,b,\cdot)$ is Lipschitz is also its consequence of its expression as a minimum of a finite amount of linear maps. The expression of sub-gradients of $W(a,b,\cdot)$ is a consequence of Danskin's Theorem [Dan66]. Finally, the semi-algebraic property of $\partial_C W(a,b,\cdot)$ is a consequence of the fact that $W(a,b,\cdot)$ is semi-algebraic and locally Lipschitz, or can alternatively be seen using the extremal point method as done for $W(a,b,\cdot)$. □

In Lemma B.I.5 we remind some useful properties of semi-algebraic maps that we will use later on. In particular, semi-algebraic maps are *generalised differentiable* (see [EN97, Definition 3.1]), which can be understood as a generalised first-order Taylor expansion.

Lemma B.I.5. Any locally Lipschitz semi-algebraic map $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is generalised differentiable (see [EN97, Definition 3.1]) and its set of critical values $f\{x \in \mathbb{R}^d : 0 \in \partial_C f(x)\}$ is finite.

Proof. Since f is semi-algebraic and locally Lipschitz, by [BDL09, Theorem 3.6] it is semi-smooth, which in turn implies generalised differentiability (by [MGN24, Theorem 1.4]). Next, by definable Morse-Sard (from [BP21, Theorem 5]), the set of critical values $f\{x \in \mathbb{R}^d : 0 \in \partial_C f(x)\}$ is finite. □

B.I.4.2 Numerical Method for Gradients of Convex Functions

In this section, we present numerical methods to solve the approximate problem in the case of the function class $\mathcal{F}_{\mathcal{E},L,\ell}$ of functions $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ that is L-Lipschitz and gradient of an ℓ-strongly convex function $\varphi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, on each part E_k of the fixed partition $\mathcal{E}$. We already introduced this class in Section B.I.2.3, and it was first considered in the context of map problems by [PdC20]. The numerical methods will aim to solve the problem

$$\operatorname*{argmin}_{\varphi \in \mathcal{F}_{\mathcal{E},L,\ell}} \mathcal{T}_c(g\#\mu, \nu), \tag{B.I.25}$$

with a particular emphasis on the case where c is quadratic, i.e. $c(x,y) = (x-y)^\top Q(x-y) + b^\top(x-y)$, where $Q \in S_d^+(\mathbb{R})$ is a positive-semi-definite matrix, and $b \in \mathbb{R}^d$. For our numerical questions, we consider the discrete case

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}.$$

Obviously, we need to assume that the measure μ is compatible with the partition, which is to say the the x_i are never at the boundary of a part E_k: $\forall i \in [\![1,n]\!],\ x_i \in (\cup_k \partial E_k)^c$. The objective in Eq. (B.I.25) only depends on the values $\varphi_i := \varphi(x_i)$ and $g_i := g(x_i)$, the immediate question is that given a candidate $(\varphi_i, g_i) \in (\mathbb{R}\times\mathbb{R}^d)^n$, does there exist a function $g \in \mathcal{F}_{\mathcal{E},L,\ell}$ of the form $\nabla\varphi$ which interpolates these values, i.e. $g(x_i) = g_i$ and $\varphi(x_i) = \varphi_i$? This question, which is called $\mathcal{F}_{\mathcal{E},L,\ell}$-interpolation, was studied by Taylor [Tay17][5]. We write $\mathcal{F}_{L,\ell} := \mathcal{F}_{\mathcal{E},L,\ell}$ in the case $\mathcal{E} = \{\mathbb{R}^d\}$, and present the results in the restricted case where the space is $\mathbb{R}^d$, as opposed to any vector space.

Proposition B.I.9. (Multiple results from Taylor [Tay17; THG17]). Let $S = (x_i, g_i, \varphi_i)_{i\in[\![1,n]\!]} \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^n$. The set S is said to be $\mathcal{F}_{L,\ell}$-interpolable [Tay17, Definition 3.1] if there exists $\varphi \in \mathcal{F}_{L,\ell}$ such that $\forall i \in [\![1,n]\!],\ \nabla\varphi(x_i) = g_i$ and $\varphi(x_i) = \varphi_i$. Consider the quadratic function

$$Q(x,x',\varphi,\varphi',g,g') := \varphi - \varphi' - \langle g', x-x'\rangle - c_1\|g-g'\|_2^2 - c_2\|x-x'\|_2^2 + c_3\langle g'-g, x'-x\rangle, \quad \text{(B.I.26)}$$

for $x, x' \in \mathbb{R}^d,\ \varphi, \varphi' \in \mathbb{R},\ g, g' \in \mathbb{R}^d$, with

$$c_1 := \frac{1}{2L(1-\ell/L)}, \quad c_2 := \frac{\ell}{2(1-\ell/L)}, \quad c_3 := \frac{\ell}{L(1-\ell/L)}.$$

- [Tay17, Theorem 3.8] The set S is $\mathcal{F}_{L,\ell}$-interpolable if and only if for all $i, j \in [\![1,n]\!]$,

$$Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0. \quad \text{(B.I.27)}$$

- [Tay17, Theorem 3.14] For $x \in \mathbb{R}^d$, let:

$$\varphi_l(x) = \min_{t\in\mathbb{R},\ g\in\mathbb{R}^d} t, \quad \text{(B.I.28)}$$
$$\text{s.t. } \forall j \in [\![1,n]\!],\ Q(x, x_j, t, \varphi_j, g, g_j) \geq 0;$$

$$\varphi_u(x) = \max_{t\in\mathbb{R},\ g\in\mathbb{R}^d} t, \quad \text{(B.I.29)}$$
$$\text{s.t. } \forall i \in [\![1,n]\!],\ Q(x_i, x, \varphi_i, t, g_i, g) \geq 0.$$

If S is $\mathcal{F}_{L,\ell}$-interpolable, then any interpolating function φ satisfies $\varphi_l \leq \varphi \leq \varphi_u$, and the potentials φ_l, φ_u are valid interpolations.

Proposition B.I.9 shows that the constraint on $(\varphi_i, g_i)_i$ can be written as a set of quadratic constraints. It follows immediately that any problem that only depends on the values $g(x_i)$ for a variable $G \in \mathcal{F}_{\mathcal{E},L,\ell}$ can be written as a problem over $(\varphi_i, g_i)_i$ under quadratic constraints, as stated in Corollary B.I.2.

Corollary B.I.2. Consider an objective $J : \mathcal{F}_{\mathcal{E},L,\ell} \longrightarrow \mathbb{R}_+$ such that for $g \in \mathcal{F}_{\mathcal{E},L,\ell}$, the

[5]Note that [Tay17, Theorem 3.14] writes an erroneous argmin for φ_u: in the light of [Tay17, Remark 3.13], it should instead read argmax, especially given the fact that the minimisation problem is unbounded.

value $J(g)$ can be written $J\left(g(x_1), \cdots, g(x_n)\right)$. Then the problem

$$\min_{g \in \mathcal{F}_{\mathcal{E},L,\ell}} J(g) \tag{B.I.30}$$

is equivalent to the problem

$$\begin{aligned} &\min_{\substack{\varphi_1,\cdots,\varphi_n \in \mathbb{R} \\ g_1,\cdots,g_n \in \mathbb{R}^d}} J(g(x_1), \cdots, g(x_n)) \\ &\text{s.t. } \forall k \in [\![1, K]\!],\ \forall i, j \in I_k : Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0, \end{aligned} \tag{B.I.31}$$

where $I_k := \{i \in [\![1, n]\!] \mid x_i \in E_k\}$, and Q is defined in Eq. (B.I.26). Given a solution $(\varphi_i^*, g_i^*)_i$ of Eq. (B.I.31), any solution $\nabla\varphi^*$ of Eq. (B.I.30) satisfies $\varphi_l \leq \varphi^* \leq \varphi_u$ on $\cup_k \mathring{E}_k$, where for $x \in \mathring{E}_k$, the bounding potentials and their gradients are solutions of:

$$\begin{aligned} (\varphi_l(x), \nabla\varphi_l(x)) = &\operatorname*{argmin}_{t \in \mathbb{R},\ g \in \mathbb{R}^d} t, \\ &\text{s.t. } \forall j \in I_k,\ Q(x, x_j, t, \varphi_j^*, g, g_j^*) \geq 0; \end{aligned} \tag{B.I.32}$$

$$\begin{aligned} (\varphi_u(x), \nabla\varphi_u(x)) = &\operatorname*{argmax}_{t \in \mathbb{R},\ g \in \mathbb{R}^d} t, \\ &\text{s.t. } \forall i \in I_k,\ Q(x_i, x, \varphi_i^*, t, g_i^*, g) \geq 0. \end{aligned} \tag{B.I.33}$$

The potentials (φ_l, φ_u) themselves are both solutions of Eq. (B.I.30).

Note that the values of the potentials can be chosen arbitrarily on the boundaries ∂E_k.

We can now provide an algorithm for $\operatorname{argmin}_{g \in \mathcal{F}_{\mathcal{E},L,\ell}} \mathcal{T}_c(g\#\mu, \nu)$ (Eq. (B.I.25)) using Corollary B.I.2: the objective is

$$J(g(x_1), \cdots g(x_n)) = \min_{\pi \in \Pi(a,b)} \sum_{i,j} \pi_{i,j} c(g(x_i), y_j), \tag{B.I.34}$$

and the resulting problem defined in Eq. (B.I.31) can be solved by alternating over π (solving a discrete Kantorovich problem, using `ot.emd` from the PythonOT library, for instance [Fla+21]), and over (φ_i, g_i), for which the constraints are quadratic and the objective depends on the cost c. For smooth cost, one may use projected gradient descent, and for (convex) quadratic costs, the problem becomes a (convex) Quadratically Constrained Quadratic Program (QCQP). In the case $c(x, y) = \|x - y\|_2^2$, this method is already known, and is the core contribution of [PdC20] summarised in Algorithm B.I.1, with a generalisation to convex quadratic costs $c_P(x, y) := (x - y)^\top P(x - y)$ with $P \in S_d^{++}(\mathbb{R})$ (a positive-definite symmetric matrix). We remind the notation $\Pi_{c_P}^*(g\#\mu, \nu)$ as the set of optimal couplings between $g\#\mu$ and ν for the cost c_P.

Time complexity.. From a time complexity standpoint, the QCQP problem at lines 3-4-5 bears a substantial cost. As a coarse analysis, standard methods such as [YT89] have $\mathcal{O}(L^2N^4)$ complexity, where N is the dimension of variables, here $N = (d+1)n$, and where $L = N^2 + NM + R$, where M is the number of constraints, here $M = \sum_k \#I_k^2$, and $R = [\log |T|]$, with T the sum of the non-zero integers in the float representation of P and the constraint matrix. For simplicity, we will continue with $K = 1$ and thus $M = n^2$. This yields the final (prohibitive) complexity: $\mathcal{O}\left((n(d+1) + n^3 + R)^2(d+1)^4 n^4\right)$. For the transport cost, using the network simplex (see the explanation in [PC19b, Section 3.5]), omitting multiplicative logarithmic terms, the time complexity of solving the linear Kantorovich problem between measures with n and m points and cost matrix M is $\mathcal{O}((n+m)nm \log(n+m) \log((n+m)\|M\|_\infty))$ [Tar97].

In Fig. B.I.8, we present a numerical example of the method for a map fitting a two-dimensional standard Gaussian to a two-dimensional Gaussian Mixture. Since no public implementation of the QCQP problem from [PdC20] is available, we contributed a solver for

Algorithm B.I.1: Alternate Minimisation for the Gradient of Strongly Convex Functions.

Data: Strongly convex constant $\ell \geq 0$, Lipschitz constant $L \geq \ell$, disjoint point classes $I_k \subset [\![1, n]\!]$ and discrete probability distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$.

Initialisation: Compute $\pi \in \Pi^*_{c_P}(\mu, \nu)$.;

for $t \in [\![0, T_{\max} - 1]\!]$ **do**

  Update $(\varphi_i, g_i)_{i \in [\![1,n]\!]}$ by solving the QCQP:

  $\min_{\substack{\varphi_1, \cdots, \varphi_n \in \mathbb{R} \\ g_1, \cdots, g_n \in \mathbb{R}^d}} \sum_{i,j} (g_i - y_j)^\top P (g_i - y_j) \pi_{i,j}$

  s.t. $\forall k \in [\![1, K]\!],\ \forall i, j \in I_k : Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0.$

  Update π by solving the discrete Kantorovich problem: $\pi \in \Pi^*_{c_P}(g\#\mu, \nu)$.

end

Return: $(\varphi_i, g_i)_{i \in [\![1,n]\!]}$.

Algorithm B.I.1 for the squared-Euclidean cost in the Python OT library [Fla+21], with an example.

B.I.4.3 Numerical Method for Maps in a RKHS

We introduce a relatively straightforward kernel method to solve the map problem (Eq. (B.I.4)). We fix a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $\mathcal{X} \longrightarrow \mathbb{R}^d$ of kernel $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^{d \times d}$. We denote by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the inner product on $\mathcal{H}$, and $\| \cdot \|_{\mathcal{H}}$ the associated RKHS norm on $\mathcal{H}$.

Given discrete measures $\mu = \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}(\mathbb{R}^d)$, we will solve a regularised variant of the map problem (Eq. (B.I.4)):

$$\operatorname*{argmin}_{h \in \mathcal{H}} \mathcal{T}_c(h\#\mu, \nu) + \lambda \|h\|^2_{\mathcal{H}}, \tag{B.I.35}$$

for some constant $\lambda > 0$ that penalises the norm of h, which equates to imposing regularity on the function h. Given the support of μ, the cost $\mathcal{T}_c(h\#\mu, \nu)$ only depends on $(h(x_1), \cdots, h(x_n))$. A well known reduction method in RKHS theory (detailed in Section B.I.6.3 for completeness) then allows to look for solutions in an n-dimensional linear subspace V of $\mathcal{H}$:

$$V := \left\{ \sum_{k=1}^n K(\cdot, x_k) u_k \ :\ \forall k \in [\![1, n]\!],\ u_k \in \mathbb{R}^d \right\}, \quad \text{of} \quad \operatorname*{argmin}_{h \in V} \mathcal{T}_c(h\#\mu, \nu) + \lambda \|h\|^2_{\mathcal{H}}. \tag{B.I.36}$$

Since any element $h \in V$ is characterised by its coefficients $(u_1, \cdots, u_n) \in (\mathbb{R}^d)^n$, we can formulate Eq. (B.I.36) as a problem over the (u_i). First, using the kernel reproducing property, we compute

$$\left\| \sum_{k=1}^n K(\cdot, x_k) u_k \right\|^2_{\mathcal{H}} = \sum_{k=1}^n \sum_{l=1}^n u_k^\top K(x_k, x_l) u_l. \tag{B.I.37}$$

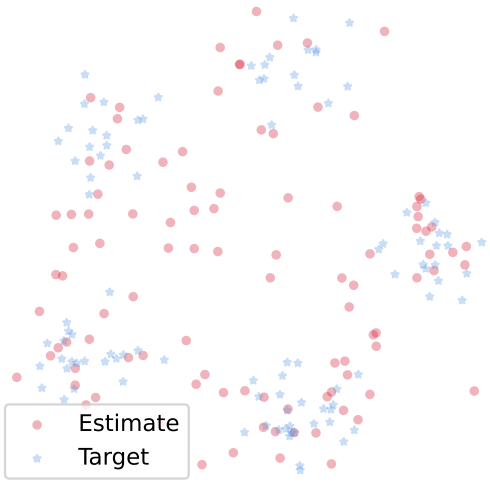
Concerning the transport cost term, we remind the notation for the value of the Kantorovich discrete problem

$$\mathrm{W}(a, b, M) := \min_{\pi \in \Pi(\mu, \nu)} M \cdot \pi,$$

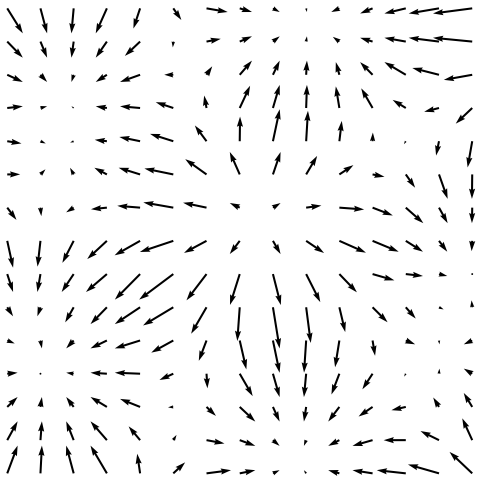
and in this case, the cost matrix M can be computed using the expression

$$\forall (i, j) \in [\![1, n]\!] \times [\![1, m]\!],\ M_{i,j} = c\left(\sum_{k=1}^n K(x_i, x_k) u_k, y_j \right). \tag{B.I.38}$$

The dependency in the (u_i) lies in the cost M. Numerically, provided that c is sufficiently regular, this allows a minimisation through classical algorithms such as gradient descent, using



(a) Positions of the optimal value positions (g_i^*) for the map estimated with Algorithm B.I.1.



(b) Evaluation of the learned map $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ from Algorithm B.I.1 on a grid.

Figure B.I.8: Illustration of the method described in Algorithm B.I.1 for the map problem from samples of a standard Gaussian distribution to samples of a Gaussian mixture. The map g is constrained to be 2-Lipschitz and the gradient of a 1/2-strongly convex function. The cost is taken as $c(x, y) = \|x - y\|_2^2$. Note that due to the constraints, we obtain an inexact matching, with in particular leakage between the modes of the target distribution.

differentiable implementations of the discrete Kantorovich cost, such as `ot.emd2` [Fla+21]. By introducing the $nd \times nd$ matrix $\mathbf{K}$ defined by $n \times n$ blocks $K(x_i, x_j)$ of size $d \times d$:

$$\mathbf{K} = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}, \tag{B.I.39}$$

and the stacked vector $\mathbf{u} \in \mathbb{R}^{nd}$, Eqs. (B.I.37) and (B.I.38) can be re-written as matrix products. This yields our final expression for Eq. (B.I.35):

$$\min_{\mathbf{u} \in \mathbb{R}^{nd}} \mathrm{W}(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^\top \mathbf{K} \mathbf{u}, \quad M(\mathbf{u})_{i,j} := c\left(\mathbf{K}_{[i,:]}\mathbf{u}, y_j\right), \tag{B.I.40}$$

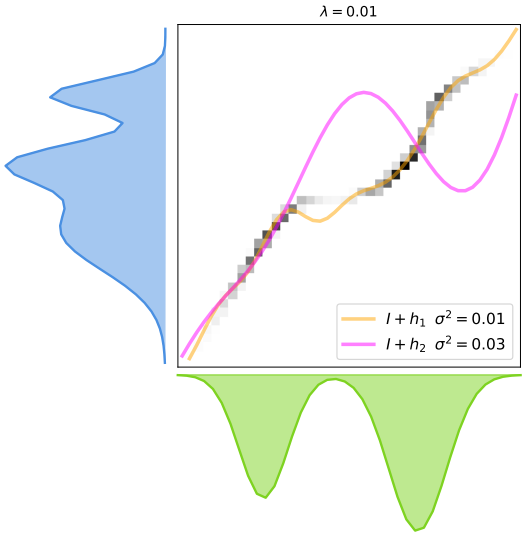
where $\mathbf{K}_{[i,:]}$ denotes the sub-matrix of $\mathbf{K}$ with the n lines $((i-1)d+1, \cdots, id)$, which corresponds to the i-th $d \times d$ block line of $\mathbf{K}$. Given optimal coefficients $\mathbf{u} = (u_1, \cdots, u_n) \in (\mathbb{R}^d)^n$, a solution h is defined everywhere in $\mathcal{X}$ using the kernel:

$$\forall x \in \mathcal{X},\ h(x) = \sum_{i=1}^{n} K(x, x_i) u_i.$$

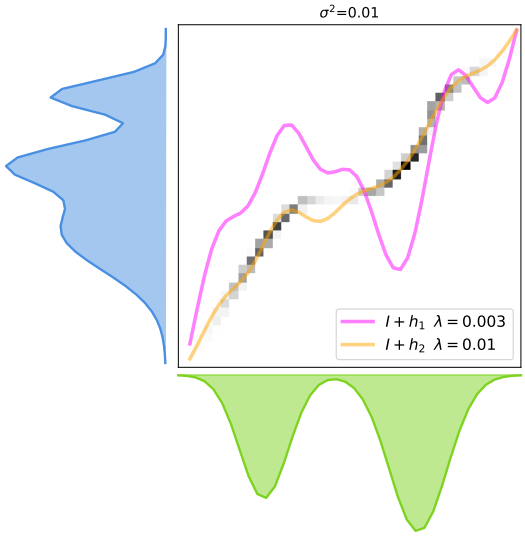
Remark B.I.9. The only constraints that are imposed upon a solution of Eq. (B.I.35) come from the choice of the kernel K (or equivalently of the space $\mathcal{H}$) and of the regularisation coefficient $\lambda > 0$. A natural idea would be to add a regularisation term $R(h)$, for instance to enforce h to be a gradient of a convex function. For Lemma B.I.8 to apply, one would need to have a regularisation which only depends on the values $(h(x_i))$, which is very restrictive. A possible heuristic would be to look for $h \in V$ regardless of this property on R, however the resulting problem would have no theoretical link to the problem over $h \in \mathcal{H}$, unlike in our case. Finally, a regularisation which depends on an infinite amount of values $h(x)$ are numerically challenging, in the specific case of *dense inequality constraints*, we refer to [RMB24] as a useful tool.

Remark B.I.10. A natural idea is to consider class of functions that are perturbations of a simple map, for instance $g = sI + h$, where h is in a RKHS $\mathcal{H}$, and $s > 0$ is a scale factor. Given Lemma B.I.8, this tweak comes without numerical or theoretical cost.

We illustrate this kernel method in Fig. B.I.9 with the Gaussian kernel (a.k.a. RBF kernel) $K(x, y) = \exp\left(-\|x - y\|_2^2/(2\sigma^2)\right) I_d$ and maps of the form $g = I + h$, where h is in the RKHS generated by the Gaussian kernel.



(a) Kernel map solution for a regularisation $\lambda = 0.01$ and multiple scales σ^2.



(b) Kernel map solution for a scale $\sigma^2 = 0.01$ and multiple regularisations λ.

Figure B.I.9: Illustration of the kernel method for the map problem between two Gaussian mixtures, using the Gaussian kernel. In greyscale, the OT plan is represented for reference.

From an algorithmic viewpoint, we propose in Algorithm B.I.2 a simple (sub-)gradient descent method (GD) for the discrete kernel map problem Eq. (B.I.40), and provide a convergence result in Proposition B.I.10 using results from Section B.I.4.1 and Ermoliev-Norkin [EN97].

Algorithm B.I.2: GD on the Kernel Map Parameters.

Data: Gradient steps $\alpha_t > 0$, kernel regularisation $\lambda > 0$, discrete probability distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$, kernel function K.

Pre-Processing: Compute the matrix $\mathbf{K}$ from Eq. (B.I.39);

Initialisation: Draw $\mathbf{u}_0 \in \mathbb{R}^{nd}$;

for $t \in [\![0, T_{\max} - 1]\!]$ **do**

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \alpha_t \left[\frac{\partial}{\partial \mathbf{u}} \left(W(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^\top \mathbf{K} \mathbf{u}\right)\right]_{\mathbf{u}=\mathbf{u}_t} \quad \text{(see Eq. (B.I.40))}.$$

end

Concerning time complexity, there are two main bottlenecks: the matrix-vector computations $\mathbf{Ku}$ which incur a $\mathcal{O}(n^2 d^2)$ cost, and solving the discrete Kantorovich problem, which is in $\mathcal{O}((n+m)nm \log(n+m) \log((n+m)\|M\|_\infty))$, as discussed in Section B.I.4.2. Note that for memory efficiency, one may use map-reduce methods such as proposed in [Cha+21] to avoid storing the matrix $\mathbf{K}$, at the cost of a higher time complexity. For scalar kernels $K(x, x') = k(x, x')I_d$, it suffices to store $\mathbf{K} := (k(x_i, x_j))_{i,j} \in \mathbb{R}^{n \times n}$, reducing the memory complexity to $\mathcal{O}(n^2 + nd)$, and matrix-vector products to $\mathcal{O}(n^2 d)$.

Proposition B.I.10. Convergence of GD for the Kernel Method, application of [EN97, Theorem 4.1] Take a locally Lipschitz and semi-algebraic (see Definition B.I.2) cost function c, and gradients steps $\alpha_t > 0$ such that $\alpha_t \to 0$ and $\sum_t \alpha_t = +\infty$. The iterates $(\mathbf{u}_t)$ of Algorithm B.I.2 are such that any accumulation point $\mathbf{v}$ is Clarke critical: $0 \in \partial_C J(\mathbf{v})$, with $J(\mathbf{u}) := W(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^\top \mathbf{K}\mathbf{u}$.

Proof. First, by Proposition B.I.8 and by semi-algebraicity and local Lipschitzness of c, J is locally Lipschitz and semi-algebraic. Using now Lemma B.I.5, we have the sufficient regularity conditions to use the convergence result of [EN97, Theorem 4.1]. □

B.I.4.4 Stochastic Gradient Descent for Neural Networks

We now consider the case where the function class G is the class of neural networks introduced in Eq. (B.I.9), with parameters in a compact set $\Theta \subset \mathbb{R}^p$, which we also assume to be convex. We will introduce a technical modification of the neural network from Eq. (B.I.9) and consider the parametrised function:

$$h := (\theta, x) \longmapsto g_{P_\Theta(\theta)}(x), \tag{B.I.41}$$

with g the map defined Eq. (B.I.9), and where the map $P_\Theta : \mathbb{R}^p \longrightarrow \Theta$ denotes the orthogonal projection onto Θ. This re-writing allows us to define the NN h on all of $\mathbb{R}^p \supset \Theta$. In practice, SGD with this network is essentially equivalent to projecting the parameters after each gradient step (with the technicality that in our formalism, the gradient of P_Θ is included in the backpropagation).

To solve the map problem of minimising $\mathcal{T}_c(h(\theta, \cdot)\#\mu, \nu)$ in practice, we consider a commonly used minibatch surrogate loss $F(\theta)$, which we define in Eq. (B.I.42). Given a dataset $X^{(n)} \in \mathbb{R}^{n\times k} = (x_1, \cdots, x_n)$, we will denote abusively $h(\theta, X^{(n)}) \in \mathbb{R}^{n\times d} := (h(\theta, x_1), \cdots, h(\theta, x_n))$. Given a dataset $X^{(n)} \in \mathbb{R}^{n\times k}$, the measure $\delta_{X^n} \in \mathcal{P}(\mathbb{R}^k)$ denotes $\frac{1}{n}\sum_i \delta_{x_i}$. Similarly, we will denote a target dataset $Y^{(m)}$. The loss F we consider is

$$F(\theta) := \int \mathcal{T}_c(\delta_{h(\theta, X^{(n)})}, \delta_{Y^{(m)}}) \mathrm{d}\mu^{\otimes n}(X^{(n)}) \mathrm{d}\nu^{\otimes m}(Y^{(m)}). \tag{B.I.42}$$

The loss F will be minimised by Stochastic Gradient Descent over θ, where the stochasticity is on the data batches $X^{(n)}$ and $Y^{(m)}$, as described in Algorithm B.I.3.

Algorithm B.I.3: Training a NN map for the cost $\mathcal{T}_c$.

Data: Gradient steps $\alpha_t > 0$, probability distributions $\mu \in \mathcal{P}(\mathbb{R}^k)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, NN $h(\theta, .)$.

1 **Initialisation:** Draw $\theta_0 \in \Theta$;

2 **for** $t \in [\![0, T_{\max} - 1]\!]$ **do**

3 Draw $X^{(n)} \sim \mu^{\otimes n}$, $Y^{(m)} \sim \nu^{\otimes m}$.

4 $\theta_{t+1} = \theta_t - \alpha_t \left[\frac{\partial}{\partial\theta} \mathcal{T}_c(\delta_{h(\theta, X^{(n)})}, \delta_{Y^{(m)}})\right]_{\theta=\theta_t}$.

5 **end**

An important remark is that this formalism bears strong similarities to the alternate minimisation framework studied in Section B.I.3 for the squared-Euclidean cost. Indeed, it can be seen as an alternation of the map parameters θ and the (minibatch) OT plan π in $\mathcal{T}_c$ (line 4): the optimisation over π is done by solving the linear program when computing the cost $\mathcal{T}_c$, and then one gradient step of optimisation over θ is performed. To study Algorithm B.I.3 theoretically, we will give precise meaning to the partial derivative at line 4 using the notions introduced in Section B.I.4.1. Numerically, the sub-gradients in question are computed by automatic differentiation. Note that P_Θ is Lipschitz and semi-algebraic.

Thanks to the regularity result on the OT cost proved in Proposition B.I.8, we can use recent SGD convergence results by [BLP23] to show that the iterates of Algorithm B.I.3 converge in

a certain sense. First, to give sense to the gradient in Algorithm B.I.3, we remark that for locally Lipschitz semi-algebraic activation functions, the map $h(\cdot,\cdot)$ is semi-algebraic and locally Lipschitz. By composition using Proposition B.I.8, the sample loss function:

$$f(\cdot, X^{(n)}, Y^{(m)}) := \theta \longmapsto \mathcal{T}_c(\delta_{h(\theta,X^{(n)})}, \delta_{(Y^{(m)})}),$$

is locally Lipschitz and semi-algebraic. We can select a semi-algebraic sub-gradient $\varphi : \mathbb{R}^p \times \mathbb{R}^{n\times k} \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^p$ such that

$$\forall \theta \in \mathbb{R}^p,\ \forall X^{(n)} \in \mathbb{R}^{n\times k},\ \forall Y^{(m)} \in \mathbb{R}^{m\times d},\ \varphi(\theta, X^{(n)}, Y^{(m)}) \in \partial_C f(\theta, X^{(n)}, Y^{(m)}),$$

where the selection can be done by lexicographic order on coordinates, for example. Note that $f(\cdot, X^{(n)}, Y^{(m)})$ is differentiable almost-everywhere, and that at differentiable points, φ equates its usual gradient. The choice of sub-gradient performed by automatic differentiation satisfies this condition (see a discussion on this procedure in [BP21; Dav+20].) We remind that the population loss function is $F = \theta \longmapsto \int f(\theta, X^{(n)}, Y^{(m)}) \mathrm{d}\mu^{\otimes n}(X^{(n)}) \mathrm{d}\nu^{\otimes m}(Y^{(m)})$ in this setting.

Proposition B.I.11. Convergence of SGD for NN maps, application of [BLP23, Theorem 3] Assume that μ, ν are discrete measures or compactly supported measures with semi-algebraic densities with respect to the Lebesgue measure. Assume that the NN h is defined as in Eq. (B.I.41), with locally Lipschitz semi-algebraic activation functions. Assume that Θ is compact, convex and semi-algebraic. Suppose that the cost function $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ is locally Lipschitz and semi-algebraic. Take gradient steps $(\alpha_t)_{t\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ such that $\sum_t \alpha_t = +\infty$ with $\alpha_t = o(1/\log(t))$.
Then there exists a set of possible steps $A \subset (0, +\infty)$ whose complement is finite, and a set of possible initialisations $\Theta_0 \subset \Theta$ of full measure, such that for each step sequence $(\alpha_t) \in A^{\mathbb{N}}$ verifying the conditions, the stochastic gradient descent iterates:

$$\theta_0 \in \Theta_0,\ \forall t \in \mathbb{N},\ \theta_{t+1} = \theta_t - \alpha_t \varphi(\theta_t, X_t^{(n)}, Y_t^{(m)}),\ X_t^{(n)} \sim \mu^{\otimes n},\ Y_t^{(m)} \sim \nu^{\otimes m},$$

verify that almost-surely, $(F(\theta_t))$ converges, and almost-surely any accumulation point $\overline{\theta}$ of (θ_t) is such that $0 \in \partial_C F(\overline{\theta})$, under the (mild) additional assumption that that the trajectories (θ_t) are almost-surely bounded.

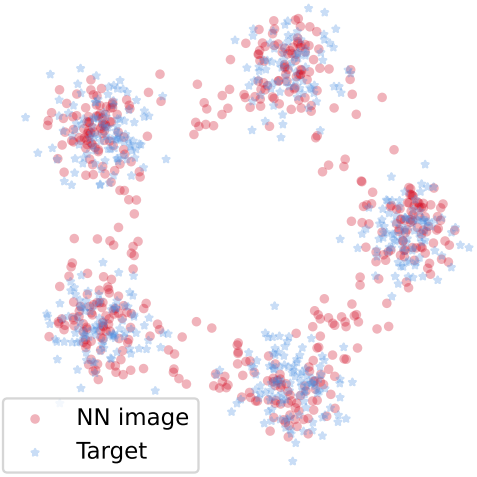
Proof. We apply [BLP23, Theorem 3], to the NN h with the discrete OT loss from Proposition B.I.8. Thanks to the assumptions formulated in the result statement, to Proposition B.I.8 and to the construction of a semi-algebraic sub-gradient selection φ, we have collected all the conditions for [BLP23, Theorem 3], yielding the result. For the case where one or both of $\{\mu, \nu\}$ is/are discrete, we applied their Remark 3. □

In Fig. B.I.10, we present a numerical example of the method for a map fitting a two-dimensional standard Gaussian to a two-dimensional Gaussian Mixture.

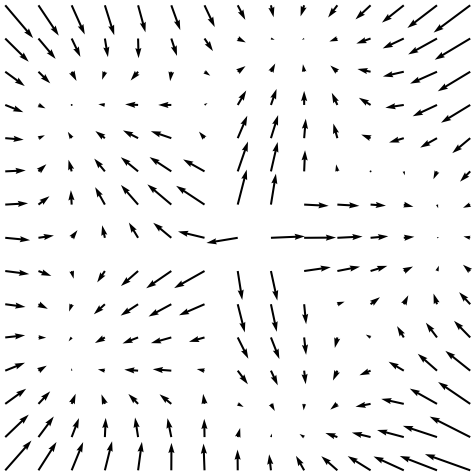
B.I.4.5 Illustrative Application to Colour Transfer

In this section, we consider the task of colour transfer, which consists in transforming the colour distribution of a source image onto the colour distribution of a target image. An $(n \times m)$ RGB image is seen as a 3-tensor $\mathbf{I} \in [0,1]^{n\times m\times 3}$, and its colour distribution is then a discrete measure $\mu = \frac{1}{nm}\sum_{i,j} \delta_{\mathbf{I}_{i,j}}$ on $\mathbb{R}^3$.

We illustrate that a learned map $g : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ which is optimised to transfer the colours of a source image $\mathbf{I}_s$ onto a target image $\mathbf{I}_t$ can be used on a new image $\mathbf{I}$ to transfer its colours. This is made possible since the map g is defined everywhere, and not only at the points of μ. We consider the cost $c(x,y) = \|x-y\|_2^2$, and a simple NN map g using Algorithm B.I.3 on the source and target images, and then apply the map to new images. We present the results in Fig. B.I.11 for three different training tasks. Notice how the constraint on the map g allows us to have a colour transfer that is robust to outliers. In Fig. B.I.12, we present the results in the RGB space, seeing the images as pixel point clouds.



(a) Comparison of images of samples of the source Gaussian by the learned NN map to the target Gaussian Mixture.



(b) Evaluation of the learned NN map $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ on a grid.

Figure B.I.10: Illustration of the method described in Algorithm B.I.3 for the map problem from samples of a standard Gaussian distribution to samples of a Gaussian mixture. The map g is of the form $g = I + h$, where h is a small 4-layer NN with ReLU activation functions, and weights constrained to $[-1/2, 1/2]$. The cost is taken as $c(x, y) = \|x - y\|_2^2$. Note that due to the (indirect) constraint on the Lipschitz constant, we obtain an inexact matching, with in particular leakage between the modes of the target distribution.

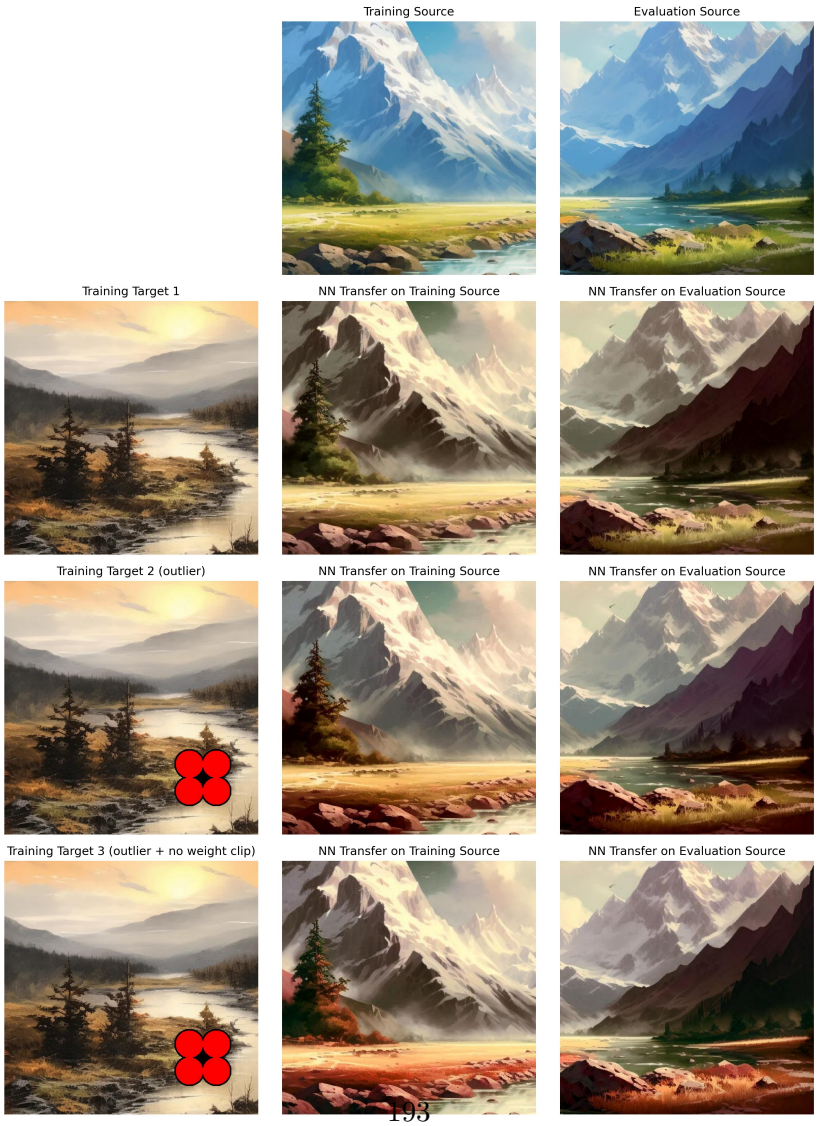


Figure B.I.12: RGB space visualisation of the colour transfer from Fig. B.I.11.

B.I.5 Conclusion and Outlook

In this chapter, we have considered the problem of finding an optimal transport map g between two probability measures μ and ν under the constraint that $g \in G$, where G is a given set of functions (L-Lipschitz, gradient of a convex function, for instance). We have given general assumptions to ensure the existence of an optimal map g, we have studied the relationship between our problem and many other concepts in Optimal Transport, and also the link with kernel methods. We have also explained how to solve the problem from a practical point a view with convergence guarantees, and an application to colour transfer.

We believe that there are two important but difficult questions that should be investigated in future work. The first is the question of the uniqueness of an optimal map. We have given a partial answer to this question, but it seems to be a difficult question in its whole generality. Having a result of uniqueness would then open the way to new questions, such as the use of g to compare measures in a way similar to Linearised Optimal Transport, or the study of the statistical properties of g (related to the sample complexity). The second important

question is the addition of constraints in the kernel method, more precisely: how to translate a set of functions G (like the set of gradients of convex functions for instance) into a RKHS representation?

Acknowledgements

We extend our warmest thanks to Nathaël Gozlan for his valuable input regarding technical assumptions for the existence result. We thank Joan Glaunès for the fruitful time we spent working together on kernel problems. We also want to thank Tam Le for providing references and insight for non-smooth and non-convex optimisation questions. We are grateful for the remarks and feedback of two anonymous reviewers, which in particular allowed significant improvement of the map existence results.

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

B.I.6 Appendix

B.I.6.1 Continuous-to-Discrete Case: Semi-discrete OT

In the alternate optimisation scheme proposed in Section B.I.3, the step with g fixed can be seen as semi-discrete Optimal Transport, whenever the target measure ν is discrete, and when the measure $g\#\mu$ is absolutely continuous with respect to the Lebesgue measure. The condition $g\#\mu \ll \mathscr{L}$ arises naturally whenever the source measure μ is itself absolutely continuous, which we will assume for this section. Specifically, the sub-problem of computing

$$\mathcal{T}_c(g\#\mu, \nu)$$

can be seen as a semi-discrete optimal transport problem between $g\#\mu$ and ν (see [MT21] for a course with a detailed section on semi-discrete OT). To apply semi-discrete optimal transport methods to this sub-problem, we need to verify $g\#\mu \ll \mathscr{L}$. First, it follows from the definition that if $g\#\mathscr{L} \ll \mathscr{L}$, then, since we assume μ is absolutely continuous, $g\#\mu \ll \mathscr{L}$ would follow. In Lemma B.I.6, we provide relatively general sufficient conditions on the map $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

Lemma B.I.6. Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ locally Lipschitz such that for $\mathscr{L}$-a.e. $x \in \mathbb{R}^d$, $\det \partial g(x) \neq 0$. Then $g\#\mathscr{L} \ll \mathscr{L}$.

Remark B.I.11. By Rademacher's theorem ([Eva18, Theorem 3.2]), a locally Lipschitz function is differentiable $\mathscr{L}$-a.e..

Proof. First, we remind that $J_g := x \longmapsto |\det \partial g(x)|$ (defined $\mathscr{L}$-almost-everywhere) is locally integrable since g is locally Lipschitz. We now prove that $g\#\mathscr{L} \ll \mathscr{L}$ by considering the intersection of compact sets and $\mathscr{L}$-null sets. Let $\mathcal{K} \subset \mathbb{R}^d$ a compact set and $E \subset \mathbb{R}^d$ a Borel set such that $\mathscr{L}(E) = 0$. By the area formula ([Eva18, Theorem 3.8]), the following equality holds

$$\int_{g^{-1}(E)\cap\mathcal{K}} J_g(x)\mathrm{d}x = \int_{\mathbb{R}^d} \mathcal{H}^0(g^{-1}(\{y\}) \cap \mathcal{K} \cap g^{-1}(E))\mathrm{d}y = \int_E \mathcal{H}^0(g^{-1}(\{y\}) \cap \mathcal{K})\mathrm{d}y, \quad \text{(B.I.43)}$$

where $\mathcal{H}^0$ denotes the 0-dimensional Hausdorff measure (the counting measure). The left-side expression in Eq. (B.I.43) is finite. Since $\mathscr{L}(E) = 0$, it follows that the right-most term in Eq. (B.I.43) is 0, thus

$$\int_{g^{-1}(E)\cap\mathcal{K}} J_g(x)\mathrm{d}x = 0.$$

Since by assumption J_g is positive almost-everywhere, it follows that $\mathscr{L}(g^{-1}(E) \cap \mathcal{K}) = 0$. Since the compact set $\mathcal{K}$ was chosen arbitrarily, we conclude that $\mathscr{L}(g^{-1}(E)) = 0$, which shows $g\#\mathscr{L} \ll \mathscr{L}$. □

B.I.6.2 Lemmas on Pseudo-inverses and Quantile Functions

To begin with, we introduce some notions regarding pseudo-inverses of non-decreasing functions.

Definition B.I.4. For $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing, its **right-inverse** is defined as the function:

$$\psi^{\leftarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ \psi^{\leftarrow}(p) := \inf \{x \in \mathbb{R} \mid \psi(x) \geq p\}.$$

For $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing, its **left-inverse** is defined as the function:

$$\phi^{\rightarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ \phi^{\rightarrow}(p) := \sup \{x \in \mathbb{R} \mid \phi(x) \leq p\}.$$

These notions are particularly useful for the definition of the right-inverse of the cumulative distribution function of a probability measure μ: $F_\mu := x \longmapsto \mu((-\infty, x])$, and for the left-inverse of the function $G_\mu := x \longmapsto \mu((-\infty, x))$. We recall and prove some well-known properties of pseudo-inverses (see [EH13] for a detailed presentation of right-inverses). For a non-decreasing function ψ, we define

$$\psi(-\infty) := \lim_{x \searrow -\infty} \psi(x) \in \mathbb{R} \cup \{-\infty\} \text{ and } \psi(+\infty) := \lim_{x \nearrow +\infty} \psi(x) \in \mathbb{R} \cup \{+\infty\}.$$

Lemma B.I.7.

1. Let $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing and right-continuous. Then:
   (a) For all $(x, p) \in \mathbb{R}^2$, $\psi(x) \geq p \iff x \geq \psi^{\leftarrow}(p)$.
   (b) If $\psi^{\leftarrow}(p) < +\infty$, $\psi(\psi^{\leftarrow}(p)) \geq p$.
2. Let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing and left-continuous. Then:
   (a) For all $(x, p) \in \mathbb{R}^2$, $\phi(x) \leq p \iff x \leq \phi^{\rightarrow}(p)$.
   (b) If $\phi^{\rightarrow}(p) > -\infty$, $\phi(\phi^{\rightarrow}(p)) \leq p$.
3. Under the assumptions above, if additionally $\phi \leq \psi$, then $\phi^{\rightarrow} \geq \psi^{\leftarrow}$.

Proof. We detail the proofs for claims 1.(a) and 1.(b), the arguments for 2.(a) and 2.(b) being essentially the same. First, we let $p \in \mathbb{R}$ such that $\psi^{\leftarrow}(p) < +\infty$, which is equivalent to supposing $A_p \neq \varnothing$, with $A_p := \{x \in \mathbb{R} \mid \psi(x) \geq p\}$. We also suppose $\psi^{\leftarrow}(p) > -\infty$, which is equivalent to assuming that A_p is lower-bounded. Since $A_p \neq \varnothing$, we can choose a decreasing sequence $(x_n) \in A_p^{\mathbb{N}}$ such that $x_n \xrightarrow[n \longrightarrow +\infty]{} \psi^{\leftarrow}(p)$. Since ψ is right-continuous and $\psi^{\leftarrow}(p) \in \mathbb{R}$, we have $\psi(x_n) \xrightarrow[n \longrightarrow +\infty]{} \psi(\psi^{\leftarrow}(p))$. However, since each $x_n \in A_p$, we have $\psi(x_n) \geq p$, and by taking the limit in the inequality we deduce $\psi(\psi^{\leftarrow}(p)) \geq p$. If $\psi^{\leftarrow}(p) = -\infty$, then the same argument with $x_n \xrightarrow[n \longrightarrow +\infty]{} -\infty$ and $\psi(-\infty) := \lim_{x \searrow -\infty} \psi(x) \in \mathbb{R} \cup \{-\infty\}$ also shows $\psi(\psi^{\leftarrow}(p)) \geq p$, which concludes the proof of 1.(b).

For 1.(a), we first assume $\psi^{\leftarrow}(p) < +\infty$. In this case, by 1b) we have $\phi(\psi^{\leftarrow}(p)) \geq p$, thus $[\psi^{\leftarrow}(p), +\infty) \subset A_p$. Yet by definition of $\psi^{\leftarrow}(p)$, $x \in A_p \Longrightarrow x \geq \psi^{\leftarrow}(p)$, thus we conclude $A_p = [\psi^{\leftarrow}(p), +\infty)$, which is exactly the same statement as $\psi(x) \geq p \iff x \geq \psi^{\leftarrow}(p)$. If $\psi^{\leftarrow}(p) = +\infty$, then the equivalence still holds, since $\psi(x) \geq p \iff x \in A_p$, with $A_p = \varnothing$.

Regarding 3., let $p \in \mathbb{R}$ such that $\phi^{\leftarrow}(p) > -\infty$. Then $\{x \in \mathbb{R} \mid \phi(x) \leq p\} = (-\infty, \phi^{\leftarrow}(p)]$ by 2.a), thus $\phi^{\leftarrow}(p) = \inf\{x \in \mathbb{R} \mid \phi(x) > p\}$. The previous equality also holds when $\phi^{\leftarrow}(p) = -\infty$. Now since $\phi \leq \psi$, we have $\{x \in \mathbb{R} \mid \phi(x) > p\} \subset \{x \in \mathbb{R} \mid \psi(x) \geq p\}$, and taking the infimum yields $\phi^{\rightarrow}(p) \geq \psi^{\leftarrow}(p)$. □

Using this result, we can now prove Lemma B.I.4.

Proof of Lemma B.I.4 . First, notice that as a cumulative distribution function, F_μ is non-decreasing and right-continuous. Since g is non-decreasing, we have for $p \in (0, 1)$:

$$F_{g\#\mu}(g \circ F_\mu^{\leftarrow}(p)) = \mathbb{P}_{X \sim \mu}\left(g(X) \leq g \circ F_\mu^{\leftarrow}(p)\right) \geq \mathbb{P}_{X \sim \mu}(X \leq F_\mu^{\leftarrow}(p)) = F_\mu \circ F_\mu^{\leftarrow}(p).$$

Now if $F_\mu^{\leftarrow}(p) < +\infty$, we have $F_\mu \circ F_\mu^{\leftarrow}(p) \geq p$ by Lemma B.I.7 1.b). We now turn to the case $F_\mu^{\leftarrow}(p) = +\infty$, which implies that $\forall x \in \mathbb{R},\ F_\mu(x) < p$. Since F_μ is a cumulative distribution function, this implies $p \geq 1$, which we excluded. We have shown that $F_{g\#\mu}(g \circ F_\mu^{\leftarrow}(p)) \geq p$, thus by definition of $F_{g\#\mu}^{\leftarrow}(p)$, we have $F_{g\#\mu}^{\leftarrow}(p) \leq g \circ F_\mu^{\leftarrow}(p)$.

Regarding the converse inequality, we will show that the set

$$N := \left\{ p \in (0,1) \ : \ F_{g\#\mu}^{\leftarrow}(p) < g \circ F_\mu^{\leftarrow}(p) \right\}$$

is Lebesgue-null. Let $p \in N$ and $\alpha \in \left[F_{g\#\mu}^{\leftarrow}(p), g \circ F_{g\#\mu}^{\leftarrow}(p) \right)$. As done earlier with F_μ, using Lemma B.I.7 and the fact that $F_{g\#\mu}$ is a c.d.f. and that $p < 1$, we have $F_{g\#\mu} \circ F_{g\#\mu}^{\leftarrow}(p) \geq p$. Since $F_{g\#\mu}$ is non-decreasing, we obtain $p \leq F_{g\#\mu}(\alpha)$. We re-write $F_{g\#\mu}(\alpha)$ using its definition, then use the fact that g is non-decreasing:

$$\begin{aligned} p \leq F_{g\#\mu}(\alpha) &= \mathbb{P}_{X\sim\mu}\left(g(X) \leq \alpha\right) \leq \mathbb{P}_{X\sim\mu}\left(g(X) < g \circ F_\mu^{\leftarrow}(p)\right) \\ &\leq \mathbb{P}_{X\sim\mu}\left(X < F_\mu^{\leftarrow}(p)\right) =: G_\mu(F_\mu^{\leftarrow}(p)). \end{aligned} \tag{B.I.44}$$

We now want to show that $G_\mu(F_\mu^{\leftarrow}(p)) \leq p$. Since $G_\mu \leq F_\mu$ and since they are non-decreasing and G_μ is left-continuous, and F_μ is right-continuous (by the axiomatic properties of μ), by Lemma B.I.7 item 3, we have $G_\mu^{\rightarrow} \geq F_\mu^{\leftarrow}$. In particular, since G_μ is non-decreasing, we have

$$G_\mu(F_\mu^{\leftarrow}(p)) \leq G_\mu(G_\mu^{\rightarrow}(p)) \leq p,$$

where the final inequality comes from Lemma B.I.7 item 2b), with $\phi^{\rightarrow}(p) > -\infty$ since we chose $p > 0$.

We have shown that $G_\mu(F_\mu^{\leftarrow}(p)) \leq p$, thus every equality in Eq. (B.I.44) is an equality, and as a result, for any $\alpha \in \left[F_{g\#\mu}^{\leftarrow}(p), g \circ F_{g\#\mu}^{\leftarrow}(p) \right)$, we have $F_{g\#\mu}(\alpha) = p$, thus the right-inverse $F_{g\#\mu}^{\leftarrow}$ has a jump-discontinuity at p:

$$\sup_{q<p} F_{g\#\mu}^{\leftarrow}(q) = F_{g\#\mu}^{\leftarrow}(p) < \inf_{p<q} F_{g\#\mu}^{\leftarrow}(q).$$

We conclude that N is a subset of the set J of jump-discontinuities of $F_{g\#\mu}^{\leftarrow}$, and since $F_{g\#\mu}^{\leftarrow}$ is non-decreasing, J is countable and thus of Lebesgue measure 0. As a result, we have for almost-every $p \in [0,1]$, $F_{g\#\mu}^{\leftarrow}(p) = g \circ F_\mu^{\leftarrow}(p)$. □

B.I.6.3 Reminder on Reduction in RKHS methods

The reduction method in RKHS is known since [Aro50, Section 3], but given the simplicity of the arguments and for the sake of self-completeness, we provide a proof and presentation in Lemma B.I.8.

Lemma B.I.8. Consider a cost function $J : \mathcal{H} \longrightarrow \mathbb{R}_+$ which can be written as $J(h) = J\left(h(x_1), \cdots, h(x_n)\right)$, then if $h^* \in \mathcal{H}$ is a solution of

$$\operatorname*{argmin}_{h\in\mathcal{H}} J(h) + \lambda \|h\|_{\mathcal{H}}^2,$$

then h_V, the orthogonal projection of h^* onto V (defined in Eq. (B.I.36)) verifies :

$$\forall i \in [\![1, n]\!],\ h_V(x_i) = h^*(x_i),$$

and as a result $J(h_W) = J(h^*)$, which leads to the following problem reduction:

$$\operatorname*{argmin}_{h\in\mathcal{H}} J(h) + \lambda \|h\|_{\mathcal{H}}^2 = \operatorname*{argmin}_{h\in V} J(h) + \lambda \|h\|_{\mathcal{H}}^2. \tag{B.I.45}$$

Proof. To show that $\forall i \in [\![1,n]\!],\ h_V(x_i) = h^*(x_i)$, we will show that

$$V^\perp = H_0 := \{h \in \mathcal{H} \mid \forall i \in [\![1,n]\!],\ h(x_i) = 0\}.$$

Indeed,

$$\begin{aligned} h \in H_0 &\iff \forall i \in [\![1,n]\!],\ g(x_i) = 0 \\ &\iff \forall i \in [\![1,n]\!],\ \forall u \in \mathbb{R}^d,\ g(x_i) \cdot u = 0 \\ &\iff \forall i \in [\![1,n]\!],\ \forall u \in \mathbb{R}^d,\ \delta_{x_i}^u g = 0 \\ &\iff \forall i \in [\![1,n]\!],\ \forall u \in \mathbb{R}^d,\ \langle g, K(\cdot,x_i)u\rangle_{\mathcal{H}} = 0 \\ &\iff f \in V^\perp, \end{aligned}$$

where δ_x^u is the linear form $h \longmapsto h(x) \cdot u$, whose Riesz representation in $\mathcal{H}$ is $K(\cdot,x)u$ by the kernel reproducing property. We conclude the proof with the fact that as an orthogonal projection, $\|h_V\|_{\mathcal{H}}^2 \leq \|h^*\|_{\mathcal{H}}^2$, which shows that the cost of h_V is less than the cost of h^*. □

B.I.6.4 Extending Kantorovich Potentials to Maps

In this section, we shall recall a relatively uncommon method[6] which extends a discrete Optimal Transport map to the entire space, and provide novel technical details about its theoretical properties. We shall see that in some sense, this method is unsatisfactory, since the resulting extended map is piecewise constant, and may not be well-defined on the support of the input discrete measure. This technique stems from the extension of Kantorovich dual potentials, which we present beforehand. Let $\Omega \subset \mathbb{R}^d$ be a compact set and $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ a $\mathcal{C}^0$ cost function. By [San15, Theorem 1.5, Theorem 1.42, and Proposition 1.11] we have the following duality:

$$\begin{aligned} (\mathrm{KP}) : \min_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega^2} c(x,y)\mathrm{d}\gamma(x,y) &= (\mathrm{KD}) : \max_{\substack{\phi,\psi \in \mathcal{C}_b(\Omega) \\ \phi \oplus \psi \leq c}} \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \psi \mathrm{d}\nu \\ &= (\mathrm{KD}') : \max_{\phi \in \mathrm{c-conc}(\Omega)} \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \phi^c \mathrm{d}\nu, \end{aligned} \tag{B.I.46}$$

where for $\phi : \Omega \longrightarrow \overline{\mathbb{R}}$, its c-transform ϕ^c and $\bar{c}$-transform $\phi^{\bar{c}}$ are respectively:

$$\phi^c : \left\{ \begin{array}{ccc} \Omega & \longrightarrow & \overline{\mathbb{R}} \\ y & \longmapsto & \inf_{x \in \Omega} c(x,y) - \phi(x) \end{array} \right. , \ \phi^{\bar{c}} : \left\{ \begin{array}{ccc} \Omega & \longrightarrow & \overline{\mathbb{R}} \\ x & \longmapsto & \inf_{y \in \Omega} c(x,y) - \phi(y) \end{array} \right. ;$$

and $\mathrm{c-conc}(\Omega)$ denotes the set of "c-concave" functions, i.e. functions $\phi : \Omega \longrightarrow \overline{\mathbb{R}}$ such that there exists $\psi : \Omega \longrightarrow \overline{\mathbb{R}}$ with $\phi = \psi^{\bar{c}}$. For more details, see [San15, Section 1.2]. In fact, it is possible to allow the dual variables from (KD) to be defined only on the supports of μ, ν (see [GSZ18, Theorem 4.12] for instance):

$$(\mathrm{KP}) = (\mathrm{KD}'') : \max_{\substack{\phi \in L^1(\mu), \psi \in L^1(\nu) \\ \phi \oplus \psi \leq c}} \int_\Omega \phi \mathrm{d}\mu + \int_\Omega \psi \mathrm{d}\nu, \tag{B.I.47}$$

where the inequality constraint is to be understood $\mu \otimes \nu$-a.e.. In the spirit of [San15, Remark 1.13], and of [Fey20, Equation (3.173)], a dual solution $(\phi, \psi) \in L^1(\mu) \times L^1(\nu)$ of (KD″) is only known μ-a.e.. To extend it on Ω, we introduce $\chi_\mu : \Omega \longrightarrow \{-\infty, 1\}$ which takes the value 1 on the support of μ, and $-\infty$ elsewhere. One may verify that the pair $\left(((\phi\chi_\mu)^c)^{\bar{c}}, (\phi\chi_\mu)^c\right)$ is optimal for (KD). Another possibility is the pair $\left((\psi\chi_\nu)^{\bar{c}}, (\phi\chi_\mu)^c\right)$, which will be our object of focus.

In the case where $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ are discrete, Kantorovich duality can be written as the following linear program dual ([PC19b, Proposition 2.4]):

$$\min_{\substack{\pi \in \mathbb{R}_+^{n \times m} \\ \pi \mathbf{1} = a, \pi^\top \mathbf{1} = b}} \langle \pi, C \rangle = \max_{\substack{u \in \mathbb{R}^n,\ v \in \mathbb{R}^m \\ u \oplus v \leq C}} u^\top a + v^\top b, \tag{B.I.48}$$

[6] kindly pointed out to us by Jean Feydy.

with $C = (c(x_i, y_j))_{i,j}$. Given our earlier remark (and, again, in the light of [Fey20, Equation (3.173)]), optimal dual potentials $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ can be extended to a pair of functions $(\phi, \psi) \in \mathcal{C}(\mathbb{R}^d)^2$, using the following expression:

$$\forall x \in \mathbb{R}^d,\ \phi(x) := \min_{j \in [\![1,m]\!]} c(x, y_j) - v_j, \quad \forall y \in \mathbb{R}^d,\ \psi(y) := \min_{i \in [\![1,n]\!]} c(x_i, y) - u_i. \tag{B.I.49}$$

To prove some properties of the extensions ϕ, ψ, we will use a well-known[7] necessary condition on optimal potentials u, v solutions of Eq. (B.I.48):

Lemma B.I.9. Assume that entry-wise, $a > 0$ and $b > 0$. Consider (u, v) a solution of Eq. (B.I.48), then necessarily the pair verifies

$$\forall i \in [\![1, n]\!],\ u_i = \min_{j \in [\![1,m]\!]} C_{i,j} - v_i, \quad \forall j \in [\![1, m]\!],\ v_j = \min_{i \in [\![1,n]\!]} C_{i,j} - u_i. \tag{B.I.50}$$

In terms of discrete C-transforms, this result can be written as $(u, v) = (v^{\overline{C}}, u^C)$. In particular, Lemma B.I.9 implies that the extended potentials ϕ, ψ coincide with the associated discrete potentials u, v: $\phi(x_i) = u_i$ and $\psi(y_j) = v_j$.

Proof. Consider (π, u, v) optimal solutions of the primal and dual formulations of Eq. (B.I.48). By complementary slackness ([PC19b, Proposition 3.2]), we have

$$\forall (i, j) \in [\![1, n]\!] \times [\![1, m]\!],\ \pi_{i,j} \neq 0 \Longrightarrow u_i + v_j = C_{i,j}.$$

Let $j \in [\![1, m]\!]$. Since $(\pi^\top \mathbf{1})_j = b_j > 0$, there exists $i_0 \in [\![1, n]\!]$ such that $\pi_{i_0,j} > 0$. By complementary slackness, it follows that $v_j = C_{i_0,j} - u_{i_0}$, which yields

$$v_j = C_{i_0,j} - u_{i_0} \geq \min_{i \in [\![1,n]\!]} C_{i,j} - u_i \geq v_j,$$

where the final inequality comes from the dual constraint $u_i + v_j \leq C_{i,j}$. We conclude $v_j = \min_{i \in [\![1,n]\!]} C_{i,j} - u_i$, and the same reasoning can be applied for u. □

In the context of optimal transport *maps*, a typical assumption to consider is the "twist condition" for the cost c.

Definition B.I.5. $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ satisfies the **twist condition** if $x \longmapsto c(x, y)$ is differentiable for any $y \in \mathbb{R}^d$, and for any $x_0 \in \mathbb{R}^d$, the map $y \longmapsto \nabla_x c(x_0, y)$ is injective. ($\nabla_x c(x_0, y)$ denotes the gradient of $c(\cdot, y)$ taken at x_0.)

Using [San15, Theorem 1.47], since our constructed potential ϕ from Eq. (B.I.49) is c-concave, then if ϕ is differentiable μ-almost-everywhere, any map T that verifies the implicit equation

$$\nabla_x c(x, T(x)) = \nabla \phi(x) \tag{B.I.51}$$

at μ-almost-every $x \in \mathbb{R}^d$ is the optimal transport map between μ and $T\#\mu$ for the cost c.

In the case $c(x, y) = \|x - y\|_2^2$, the condition $\nabla_x c(x, T(x)) = \nabla \phi(x)$ equates to $T(x) = x - \frac{1}{2}\nabla \phi(x)$, which is the usual expression relating Kantorovich potentials to the Brenier map, in the case where $\mu \ll \mathscr{L}$.

To determine an OT map using our extended potential ϕ from Eq. (B.I.49), we study the solutions of Eq. (B.I.51) using our extended potential ϕ , which is the topic of the following result.

Theorem B.I.4. Let c be a cost function satisfying the twist condition (Definition B.I.5). Take $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ a pair of optimal potentials (solutions of Eq. (B.I.48)), and the asso-

[7] for which we could not find a specific reference, which is why we provide a proof.

ciated (ϕ, ψ) defined in Eq. (B.I.49). Then the extended map T defined using Eq. (B.I.51) is **piecewise constant**, with more specifically:

1) ϕ is continuous everywhere and differentiable on $\mathcal{U} := \bigcup_{j=1}^{m} \mathcal{U}_j$, where for all $j \in [\![1, m]\!]$,

$$\mathcal{U}_j := \left\{ x \in \mathbb{R}^d : \forall j' \in [\![1, m]\!] \setminus \{j\},\ c(x, y_j) - v_j < c(x, y_{j'}) - v_{j'} \right\} \tag{B.I.52}$$

$$= \operatorname{Int} \left\{ x \in \mathbb{R}^d : \phi(x) = c(x, y_j) - v_j \right\}. \tag{B.I.53}$$

2) ϕ is not differentiable at any point of $\mathcal{U}^c$.

3) Let π an optimal transport plan (solution of the primal problem in Eq. (B.I.48)). For any $i \in [\![1, n]\!]$, if $\#\{j \in [\![1, m]\!] : \pi_{i,j} > 0\} \geq 2$, then ϕ is not differentiable at x_i.

4) Eq. (B.I.51) at $x \in \mathcal{U}$ has a unique solution, thus T is wel-defined on $\mathcal{U}$ by:

$$\forall j \in [\![1, m]\!],\ \forall x \in \mathcal{U}_j,\ T(x) = y_j. \tag{B.I.54}$$

Proof. Since $c(\cdot, y)$ is differentiable for any $y \in \mathbb{R}^d$, item 1) follows immediately from the definition of ϕ. For item 2), take $x_0 \in \mathcal{U}^c$ and assume that ϕ is differentiable at x_0. Using Eq. (B.I.52), there exists $j \neq j'$ such that $\phi(x_0) = c(x_0, y_j) - v_j = c(x_0, y_{j'}) - v_{j'}$. By definition of ϕ, we have that that $r_j := c(\cdot, y_j) - v_j - \phi$ verifies $r_j \geq 0$ on $\mathbb{R}^d$ and $r_j(x_0) = 0$. By assumption, r_j is differentiable at x_0, and thus its minimality at x_0 yields $\nabla r_j(x_0) = 0$, which implies that $\nabla_x c(x_0, y_j) = \nabla \phi(x_0)$. Similarly, by considering $r_{j'} := c(\cdot, y_{j'}) - v_{j'} - \phi$, we obtain $\nabla_x c(x_0, y_{j'}) = \nabla \phi(x_0)$. Since c satisfies the twist condition, $\nabla_x c(x_0, \cdot)$ is injective, thus $\nabla_x c(x_0, y_j) \neq \nabla_x c(x_0, y_{j'})$ which is a contradiction, and we conclude that ϕ is not differentiable at x_0.

We now turn to item 3). We fix π a solution of the primal problem Eq. (B.I.48), and $i \in [\![1, n]\!]$. Assume that there exists and $j \neq j'$ such that $\pi_{i,j} > 0$ and $\pi_{i,j'} > 0$. Since (u, v) is optimal for the dual problem, by complementary slackness ([PC19b, Proposition 3.2]), we have

$$u_i + v_j = C_{i,j},\ u_i + v_{j'} = C_{i,j'}.$$

By Lemma B.I.9, it follows that $\phi(x_i) = u_i = C_{i,j} - v_j = C_{i,j'} - v_{j'}$, and thus $x_i \in \mathcal{U}^c$, which allows us to conclude that ϕ is not differentiable at x_i by item 2).

For item 4), let $x \in \mathcal{U}$, by item 1), there exists $j \in [\![1, m]\!]$ such that $x \in \mathcal{U}_j$, which allows the computation $\nabla \phi(x) = \nabla_x c(x, y_j)$, thus y_j is a solution of Eq. (B.I.51), which has a unique solution, by the twist assumption on c. We conclude that $T(x) = y_j$. □

Remark B.I.12. Item 3) in Theorem B.I.4 implies that T cannot be defined using Eq. (B.I.51) at a point x_i at which an optimal transport plan may need to "split mass". The fact that the implication does not depend on the choice of an optimal plan π is a consequence of complementary slackness, which holds whatever the choice of primal and dual optimal variables.

For example, consider the cost $c(x, y) = |x - y|^2$ the measure $\mu := \frac{1}{2}(\delta_1 + \delta_2)$ and $\nu := \frac{1}{2}(\delta_4 + \delta_5 + \delta_6)$, as represented in Fig. B.I.13.

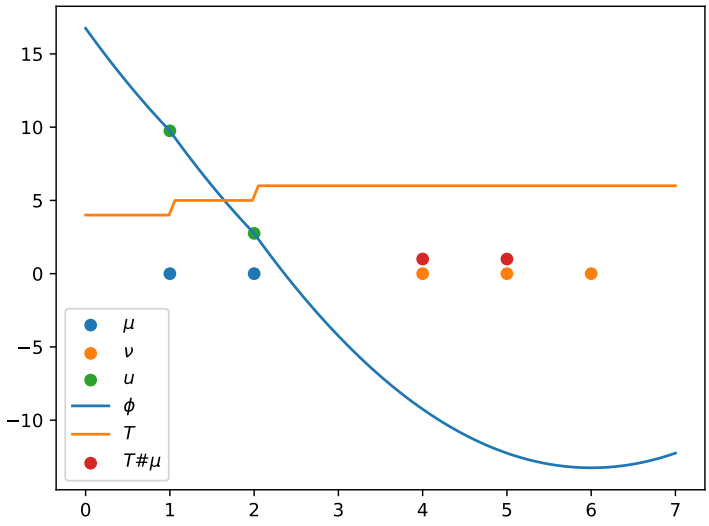


Figure B.I.13: Illustration of the extended potential ϕ from Eq. (B.I.49) and the associated map T defined in Eq. (B.I.51), for the cost $c(x, y) = |x - y|^2$.

In this case, the extended map T of Eq. (B.I.51) is uniquely defined on $\mathbb{R} \setminus \{1, 2\}$ as follows:

$$T_{(-\infty,1)} \equiv 4, \; T_{(1,2)} \equiv 5, \; T_{(2,+\infty)} \equiv 6,$$

and T cannot be defined at the points $\{1, 2\}$, since the extended potential ϕ (defined using Eq. (B.I.49)) is not differentiable at those points. In Fig. B.I.13, we choose to extend T as to be left-continuous: $T(1) := T(1-)$ and $T(2) := T(2-)$, giving an (arbitrary) meaning to the image $T\#\nu$, which obviously does not satisfy $T\#\mu = \nu$.

B.II

Sliced Optimal Transport Plans

B.II.1 Introduction . . . 204
B.II.2 Reminders and New Results on the ν-based Wasserstein Distance . . . 206
B.II.2.1 Wasserstein Geodesics and Generalised Geodesics 206
B.II.2.2 The ν-based Wasserstein Distance 207
B.II.2.3 Reminders on Wasserstein Means 211
B.II.2.4 Another Formulation of W_ν with Measure Disintegration 211
B.II.3 The Pivot Sliced Discrepancy . . . 212
B.II.3.1 Definition with the ν-based Wasserstein Distance 212
B.II.3.2 Semi-Metric Properties of PS_θ 216
B.II.4 Correspondence of Pivot-Sliced and a Constrained Wasserstein Discrepancy . . . 219
B.II.4.1 First Inequality: $\mathrm{PS}_\theta \leq \mathrm{CW}_\theta$ 219
B.II.4.2 Converse Inequality: $\mathrm{PS}_\theta \geq \mathrm{CW}_\theta$ 220
B.II.4.3 Triangle Inequality for PS_θ for Projection-Atomless Measures 221
B.II.5 A Monge Formulation of PS_θ Between Point Clouds . . . 222
B.II.5.1 The Case of Non-Ambiguous Projections 222
B.II.5.2 Problem Formulation and Reduction to Sorted Projections 222
B.II.5.3 A Kantorovich Formulation of CW_θ Between Point Clouds 223
B.II.5.4 Technical Lemmas on Bipartite Graphs Associated to Couplings 227
B.II.5.5 A Constrained Version of the Birkhoff von Neumann Theorem 229
B.II.6 Min-Pivot Sliced . . . 232
B.II.6.1 Min-Pivot Sliced Discrepancy: Definition 232
B.II.6.2 Equality with the Wasserstein Distance for Certain Discrete Measures 233
B.II.7 Expected Sliced Wasserstein . . . 234
B.II.7.1 Lifting Sliced Plans 235
B.II.7.2 Averaging Lifted Plans 238
B.II.8 Numerics . . . 240
B.II.8.1 Evaluation of the Transport Losses and Plans 240
B.II.8.2 Illustration on Colour Transfer 243
B.II.8.3 Experiments on a Shape Registration Task 244
B.II.9 Appendix . . . 245
B.II.9.1 Ambiguity in SWGG from [Mah+23] 245
B.II.9.2 Midpoints are Geodesic Middles 247
B.II.9.3 Reminders on Disintegration of Measures 248
B.II.9.4 Proof of the Disintegration Formula for ν-based Wasserstein 249

Abstract

Since the introduction of the Sliced Wasserstein distance in the literature, its simplicity and efficiency have made it one of the most interesting surrogate for the

Wasserstein distance in image processing and machine learning. However, its inability to produce transport plans limits its practical use to applications where only a distance is necessary. Several heuristics have been proposed in the recent years to address this limitation when the probability measures are discrete. In this chapter, we propose to study these different propositions by redefining and analysing them rigorously for generic probability measures. Leveraging the ν-based Wasserstein distance and generalised geodesics, we introduce and study the Pivot Sliced Discrepancy, inspired by a recent work by Mahey et al.. We demonstrate its semi-metric properties and its relation to a constrained Kantorovich formulation. In the same way, we generalise and study the recent Expected Sliced plans introduced by Liu et al. for completely generic measures. Our theoretical contributions are supported by numerical experiments on synthetic and real datasets, including colour transfer and shape registration, evaluating the practical relevance of these different solutions. This chapter is based on the paper:

[TCD25] Eloi Tanguy, Laetitia Chapel and Julie Delon.
"Sliced Optimal Transport Plans".
arxiv preprint 2508.01243 (Aug. 2025).

B.II.1 Introduction

Known for its ability to capture geometric structure in probability distributions, optimal transport has attracted considerable attention in both theoretical and applied fields. Several studies have developed its mathematical foundations in great detail [San15; Vil09], and its practical impact has been demonstrated on a broad spectrum of applications. Originally developed for applications in logistics, economics [Gal17] and fluid mechanics, computational optimal transport has also emerged in the last fifteen years as a central tool in data science. It is used nowadays for a large variety of applications, ranging from image processing, computer vision and computer graphics [RDG09; HHR22; Fey+17; BD23; Pon+21], to domain adaptation [Cou+16; MM21; Fat+21b], natural language processing [Che+], generative modelling [ACB17; Gul+17; Sal+18; Ton+24; HCD25], quantum chemistry [BDG12] or biology [Bun+24; NS25], to cite just a few.

In these applications, optimal transport is used to define meaningful discrepancies between probability distributions, taking into account the underlying geometry of the data, but also as a way to define optimal plans or maps between such data, in order to transform a given distribution into another in an optimal way. In the continuous setting, we recall that the 2-Wasserstein distance W_2 between two probability measures μ and ν on $\mathbb{R}^d$ is defined as:

$$\mathrm{W}_2^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\pi(x,y),$$

where $\Pi(\mu,\nu)$ is the set of couplings with marginals μ and ν. In the discrete case, with empirical measures supported on finite point clouds, this problem becomes a linear program over a polytope. Computing Wasserstein distances between discrete datasets comes with significant computational expense. Classical linear programming solvers used to evaluate the transport cost between two discrete measures of size n typically have a complexity of $\mathcal{O}(n^3 \log n)$ [PC19b]. This limitation has motivated the development of computationally lighter surrogates or approximations that preserve key characteristics of optimal transport metrics.

One of these popular and efficient surrogate is the Sliced Wasserstein distance (SW) [Rab+12; Bon+15a]. This approach leverages the fact that in one dimension, the Wasserstein distance has a closed-form solution. The Sliced Wasserstein distance is derived by averaging 1D Wasserstein distances over all directions on the unit sphere, offering a simple alternative to W_2:

$$\mathrm{SW}_2^2(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta\#\mu, P_\theta\#\nu)\,\mathrm{d}\theta,$$

where P_θ denotes the projection onto direction θ. Since evaluating the full integral is intractable in practice, it is approximated by Monte Carlo sampling. One draws L random directions,

computes the 1D Wasserstein distance for each, and averages the results. The 1D Wasserstein distance between empirical distributions of n points can be obtained in $\mathcal{O}(n \log n)$, so the approximate SW_2 distance can be computed in $\mathcal{O}(Ln \log n)$. This efficiency makes it especially appealing for large values of n.

The SW distance remains a true distance on the space of probability measures and retains several fundamental features of Wasserstein distances. For probability measures with compact support, it has been shown to be equivalent to the Wasserstein distance [Bon13]. It also has desirable statistical properties, such as sample complexity bounds and robustness [Nad+20b]. Its efficiency has been confirmed in numerous use cases, including domain adaptation [Lee+19], texture generation, colour and style transfer [Hei+21; Bon+15a; EW22], statistical inference [KRH18], generative modelling [DZS18; Wu+19; CTV25], auto-encoder regularisation [Kol+19b], topological data analysis [SDT25] or shape analysis [Le+24; NNH23]. Extensions to Riemannian settings have also been investigated [BDC25]. Nevertheless, a key limitation of SW is that it does not provide a transport plan or a map between distributions, which limits its use in applications that require correspondences between datasets.

To circumvent this issue, several heuristics have been proposed to extract approximate transport plans from SW. A notable example is the use of stochastic gradient descent (SGD) to minimise the objective $X \longmapsto \mathrm{SW}(\delta_X, \delta_Y)$, as a way to gradually move points from a source point cloud X to a target point cloud[1] Y. This strategy has been first explored for colour transfer and image matching tasks in [Rab+12; Bon+15a], and can provide plausible pointwise correspondences in practice, although theoretical guarantees remain partial [CS25; LM25], see also Chapter A.II.

More recently, two alternative strategies have been introduced to build transport plans grounded in Sliced Wasserstein distances. The first one, called Sliced Wasserstein Generalised Geodesics (SWGG) [Mah+23; CTV25], defines a map between two discrete distributions $\delta_X = \frac{1}{n}\sum_i \delta_{x_i}$ and $\delta_Y = \frac{1}{n}\sum_i \delta_{y_i}$ as $\tau_\theta \circ \sigma_\theta^{-1}$, where σ_θ is a permutation which sorts $(\theta^\top x_i)_{i=1}^n$ and τ_θ a permutation sorting $(\theta^\top y_i)_{i=1}^n$. The Sliced Wasserstein Generalised Geodesic distance [Mah+23, Equation 8] is then defined as: (see also Fig. B.II.24)

$$\mathrm{SWGG}_2^2(\mu_1, \mu_2, \theta) := \frac{1}{n}\sum_{i=1}^n \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2. \tag{B.II.1}$$

The second one, called Expected Sliced Transport Plans, was introduced in [Liu+24] (inspired by [Row+19]), also for discrete measures. It aims to construct couplings by averaging the 1D optimal transport plans obtained from projections. Given $\mathfrak{o}$ a probability measure on the hypersphere, with the same notations as above, the Expected Sliced Transport distance is defined as:

$$\mathbb{E}_{\theta\sim\mathfrak{o}}\left[\frac{1}{n}\sum_{i=1}^n \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2\right] \tag{B.II.2}$$

and the average transport plan as $\mathbb{E}_{\theta\sim\mathfrak{o}}[\tau_\theta \circ \sigma_\theta^{-1}]$. This yields a plan between the two d-dimensional measures that reflects the averaged behaviour along slices.

These approaches provide practical and interpretable ways to define approximate transport maps. However, they are currently defined only for discrete measures and lack a rigorous theoretical grounding in more general measure spaces. Moreover, even in the discrete setting, it can easily be shown that the right hand-side quantity in Eq. (B.II.1) depends on the choice of the permutations, rendering the quantity ill-defined, as showcased in Section B.II.9.1.

The goal of this chapter is to rigorously define and analyse these different Sliced Optimal Transport Plans for completely generic probability measures. We introduce the Pivot Sliced Discrepancy PS_θ, a discrepancy measure based on the ν-based generalised geodesics [NP23], and generalising the Sliced Wasserstein Generalised Geodesic distance [Mah+23]. In doing so, we also provide new theoretical insights on the ν-based Wasserstein distance [NP23]. We prove that PS_θ is well-defined, symmetric and separates points. We then establish an equivalence between PS_θ and a constrained version of the Wasserstein distance, showing that PS_θ coincides

[1] For a point cloud $X = (x_i)_{i=1}^N$, we write $\delta_X = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$

with the minimal transport cost among plans that preserve the projected coupling. For empirical measures, we provide Monge and Kantorovich formulations of PS_θ, proving a constrained version of the Birkhoff-von Neumann theorem [Bir46]. Additionally, we study the Min-Pivot Sliced Discrepancy, a variant that matches the true Wasserstein distance for discrete measures when the space dimension is large enough with respect to the number of points. We then study the Expected Sliced Wasserstein Plan [Liu+24], which averages 1D sliced transport plans to obtain high-dimensional (non sparse) couplings. This theoretical study is followed by numerical experiments, illustrating the behaviour of the proposed transport plans on synthetic datasets and shape registration tasks.

The chapter is organised as follows. In Section B.II.2, we recall the necessary background on ν-based Wasserstein geodesics, along with some new theoretical results that will serve as building blocks for the rest of the work. Section B.II.3 presents and analyses the Pivot Sliced Discrepancy. In Section B.II.4, we establish a precise connection between PS_θ and a constrained Wasserstein discrepancy, showing that both quantities coincide. This correspondence is further developed in Section B.II.5, where we explore Monge and Kantorovich formulations of PS_θ for discrete measures. We then study in Section B.II.6 the Min-Pivot Sliced Discrepancy, and show that it recovers the exact Wasserstein distance in certain discrete settings. Section B.II.7 introduces and analyses the concept of Expected Sliced Wasserstein Plans. Finally, Section B.II.8 is dedicated to numerical experiments.

B.II.2 Reminders and New Results on the ν-based Wasserstein Distance

In this section, we lay some pre-requisites for the objects at play in the chapter. We begin by recalling the concept of generalised geodesics in Section B.II.2.1, which allows us to introduce the ν-based Wasserstein distance in Section B.II.2.2. This (semi-)metric was first defined in [AGS05; NP23], and we will sometimes also refer to it as "Pivot Wasserstein", and prove new technical properties that will be useful later. Later in this work, we will consider the Pivot Wasserstein distance using a "Wasserstein Mean" pivot, and to this end we propose some reminders on Wasserstein means in Section B.II.2.3. Finally, in Section B.II.2.4, we revisit a disintegration formulation of the ν-based Wasserstein distance (first proved in [NP23]), which will sometimes be convenient for computations.

B.II.2.1 Wasserstein Geodesics and Generalised Geodesics

Given two measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $\Pi^*(\mu_1, \mu_2)$ the set of Optimal Transport plans between μ_1 and μ_2 for the cost $\|x-y\|_2^2$. Using such plans, we can define a notion of shortest path (i.e. geodesic) between μ_1 and μ_2 in the space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$.

Definition B.II.1. A constant-speed geodesic between $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ is a curve $[0,1] \longrightarrow \mathcal{P}_2(\mathbb{R}^d)$ constructed using an optimal transport plan $\gamma \in \Pi^*(\mu_1, \mu_2)$ as follows:

$$\mu_\gamma^{1\to 2}(t) := ((1-t)P_1 + tP_2)\,\#\gamma, \tag{B.II.3}$$

where $P_1 : (x,y) \longmapsto x$ and $P_2 : (x,y) \longmapsto y$ are the marginal projection operators. Not only is $\mu_\gamma^{1\to 2}$ a geodesic for the W_2 metric, but all (constant-speed) geodesics between μ_1 and μ_2 are of the form $\mu_\gamma^{1\to 2}$ for a suitable $\gamma \in \Pi^*(\mu_1, \mu_2)$ (this is [AGS05, Theorem 7.2.2]).

If the chosen optimal transport plan γ is induced by a transport map T (which is to say that $\gamma = (I, T)\#\mu_1$), then the geodesic takes the intuitive "displacement" formulation:

$$\mu_\gamma^{1\to 2}(t) := ((1-t)I + tT)\,\#\mu_1, \tag{B.II.4}$$

with I denoting the identity map of $\mathbb{R}^d$.

A remarkable property of the 2-Wasserstein space is that it is a Positively Curved (according to Alexandrov's metric definition of curvature) space, as proved in [AGS05, Theorem 7.3.2,

Equation 7.3.12]: for $\mu_1, \mu_2, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in \Pi^*(\mu_1, \mu_2)$ and $t \in [0,1]$, we have

$$\mathrm{W}_2^2(\mu_\gamma^{1\to 2}(t), \nu) \geq (1-t)\mathrm{W}_2^2(\mu_1, \nu) + t\mathrm{W}_2^2(\mu_2, \nu) - (1-t)t\mathrm{W}_2^2(\mu_1, \mu_2). \tag{B.II.5}$$

For $t := \frac{1}{2}$, this can be re-written as

$$\mathrm{W}_2^2(\mu_1, \mu_2) \geq 2\mathrm{W}_2^2(\mu_1, \nu) + 2\mathrm{W}_2^2(\mu_2, \nu) - 4\mathrm{W}_2^2(\mu_\gamma^{1\to 2}(t), \nu). \tag{B.II.6}$$

Unfortunately, the squared distance W_2^2 is not λ-convex along these Wasserstein geodesics ([AGS05, Example 9.1.5]), which motivated [AGS05] to introduce other curves, coined "generalised geodesics", that satisfy this desirable property. First, we consider two optimal plans $\gamma_1 \in \Pi^*(\nu, \mu_1)$ and $\gamma_2 \in \Pi^*(\nu, \mu_2)$. To introduce the notion of generalised geodesics, we will require a 3-plan $\rho \in \Pi(\nu, \mu_1, \mu_2) \in \mathcal{P}_2(\mathbb{R}^{3d})$ (i.e. with marginals $\rho_0 = \mu, \rho_1 = \mu_1, \rho_2 = \mu_2$), such that its bi-marginals coincide with the plans γ_1 and γ_2: we require $\rho_{0,1} := P_{0,1}\#\rho = \gamma_1$ and $\rho_{0,2} := P_{0,2}\#\rho = \gamma_2$, where $P_{0,i} := (y, x_1, x_2) \longmapsto (y, x_i)$. We introduce the following notation for such 3-plans:

$$\Gamma(\nu, \mu_1, \mu_2) := \left\{\rho \in \mathcal{P}_2(\mathbb{R}^{3d}) : \rho_{0,1} \in \Pi^*(\nu, \mu_1) \text{ and } \rho_{0,2} \in \Pi^*(\nu, \mu_2)\right\}. \tag{B.II.7}$$

Definition B.II.2. A generalised geodesic based on ν between μ_1 and μ_2 is then defined as ([AGS05, Definition 9.2.2]), given a $\rho \in \Gamma(\nu, \mu_1, \mu_2)$:

$$\mu_\rho^{1\to 2}(t) := ((1-t)P_1 + tP_2)\,\#\rho. \tag{B.II.8}$$

Note that this curve depends on the choice of the 3-plan ρ, which itself depends on the optimal plans γ_1 and γ_2. The existence of such a ρ can be shown using the gluing lemma (as presented in [San15, Lemma 5.5], for example). As desired, the curvature induced by these curves makes W_2^2 convex along these geodesics (in a certain sense, see [AGS05, Definition 9.2.4]), namely we have the following inequality ([AGS05, Equation 9.2.7c]), which is reversed compared to Eq. (B.II.5):

$$\mathrm{W}_2^2(\mu_\rho^{1\to 2}(t), \nu) \leq (1-t)\mathrm{W}_2^2(\mu_1, \nu) + t\mathrm{W}_2^2(\mu_2, \nu) - (1-t)t\mathrm{W}_2^2(\mu_1, \mu_2). \tag{B.II.9}$$

Like before, setting $t := \frac{1}{2}$ yields the following inequality:

$$\mathrm{W}_2^2(\mu_1, \mu_2) \leq 2\mathrm{W}_2^2(\mu_1, \nu) + 2\mathrm{W}_2^2(\mu_2, \nu) - 4\mathrm{W}_2^2(\mu_\rho^{1\to 2}(t), \nu). \tag{B.II.10}$$

If the optimal transport plans γ_1 and γ_2 are induced respectively by transport maps T_1 and T_2, then the choice of ρ is unique, with $\rho = (I, T_1, T_2)\#\nu$ ([AGS05, Remark 9.2.3], see also [AGS05, Lemma 5.3.2] for a formal proof). This yields the following expression of the generalised geodesic, which is substantially more intuitive:

$$\mu_\rho^{1\to 2}(t) = ((1-t)T_1 + tT_2)\#\nu. \tag{B.II.11}$$

B.II.2.2 The ν-based Wasserstein Distance

A closely related concept is the ν-based Wasserstein (semi)-distance, introduced by Nenna and Pass in [NP23]. This time we use a pivot measure ν to introduce a variant of the Wasserstein distance, yielding the following definition by [NP23, Definition 3][2]:

Definition B.II.3. For $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, the ν-based Wasserstein (semi)-metric between

[2] Their definition seems to have a typo, with $\Pi^*(\mu_i, \nu)$ instead of $\Pi^*(\nu, \mu_i)$. Furthermore, they work with measures supported on a bounded and convex domain of $\mathbb{R}^d$, but as they remark (footnote 4), and given [AGS05, Chapter 9], generalisation to measures on $\mathbb{R}^d$ with a finite moment of order 2 is perfectly natural.

$\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$\mathrm{W}_\nu^2(\mu_1, \mu_2) := \min_{\rho \in \Gamma(\nu, \mu_1, \mu_2)} \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2). \tag{B.II.12}$$

We illustrate the ν-based Wasserstein distance on a simple example in Fig. B.II.1.

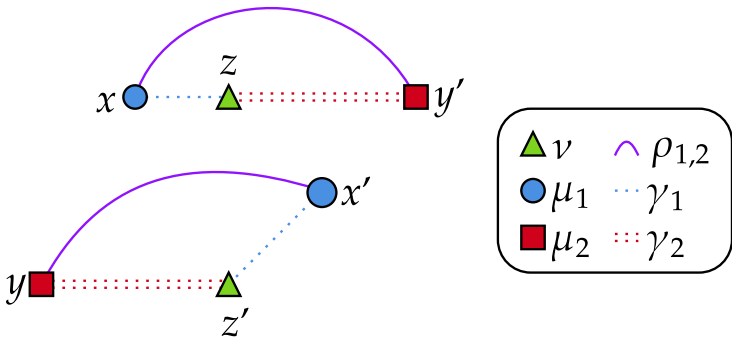


Figure B.II.1: Example of the couplings behind $\mathrm{W}_\nu(\mu_1, \mu_2)$ for discrete measures on $\mathbb{R}^2$. The measure ν is drawn with green triangles, μ_1 with blue circles and μ_2 with red squares. The (unique) OT plan γ_1 between ν and μ_1 is drawn with dotted blue lines, the (also unique) OT plan γ_2 between ν and μ_2 with red double dotted lines. The plans induce a unique valid 3-plan $\rho \in \Gamma(\nu, \mu_1, \mu_2)$, we represent the coupling $\rho_{1,2}$ between μ_1 and μ_2 with curved purple lines. Notice that the coupling $\rho_{1,2}$ differs from the (unique) OT coupling between μ_1 and μ_2.

The question of whether the infimum defining W_ν is attained was not addressed by [NP23], we show that it is indeed the case in Proposition B.II.1, using a technical property of the 3-plan set Γ defined in Eq. (B.II.7). We remind that by Prokhorov's theorem, a subset of $\mathcal{P}_2(\mathbb{R}^d)$ is tight set if and only if it is pre-compact, which means that any sequence of measures in the set has a weakly converging subsequence.

Lemma B.II.1. 1. For tight sets $P, Q_1, Q_2 \subset \mathcal{P}_2(\mathbb{R}^d)$, the set

$$\Gamma(P, Q_1, Q_2) := \{\rho \in \Gamma(\nu, \mu_1, \mu_2) : (\nu, \mu_1, \mu_2) \in P \times Q_1 \times Q_2\}$$

is tight in $\mathcal{P}_2(\mathbb{R}^{3d})$.

2. Consider sequences $\nu^{(n)}, \mu_1^{(n)}, \mu_2^{(n)} \in \mathcal{P}_2(\mathbb{R}^d)^{\mathbb{N}}$ respectively converging to $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ for the weak convergence of measures, and a sequence $(\rho_n) \in \mathcal{P}_2(\mathbb{R}^{3d})^{\mathbb{N}}$ such that $\forall n \in \mathbb{N}$, $\rho_n \in \Gamma(\nu^{(n)}, \mu_1^{(n)}, \mu_2^{(n)})$ with $\rho_n \xrightarrow[n \to +\infty]{w} \rho \in \mathcal{P}_2(\mathbb{R}^{3d})$. Then $\rho \in \Gamma(\nu, \mu_1, \mu_2)$.

3. For $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the set $\Gamma(\nu, \mu_1, \mu_2)$ is compact in $\mathcal{P}_2(\mathbb{R}^{3d})$.

Proof. For 1. we set $\varepsilon > 0$. By tightness of P, Q_1, Q_2 and Prokhorov's theorem, there exists a compact set $\mathcal{K} \subset \mathbb{R}^d$ such that for any $\mu \in P \cup Q_1 \cup Q_2$, $\mu(\mathbb{R}^d \setminus \mathcal{K}) < \varepsilon/3$. It follows that for any $\rho \in \Gamma(P, Q_1, Q_2)$,

$$\begin{aligned}
\rho(\mathbb{R}^{3d} \setminus \mathcal{K}^3) &\leq \rho\left((\mathbb{R}^d \setminus \mathcal{K}) \times \mathbb{R}^d \times \mathbb{R}^d\right) + \rho\left(\mathbb{R}^d \times (\mathbb{R}^d \setminus \mathcal{K}) \times \mathbb{R}^d\right) + \rho\left(\mathbb{R}^d \times \mathbb{R}^d \times (\mathbb{R}^d \setminus \mathcal{K})\right) \\
&= \nu(\mathbb{R}^d \setminus \mathcal{K}) + \mu_1(\mathbb{R}^d \setminus \mathcal{K}) + \mu_2(\mathbb{R}^d \setminus \mathcal{K}) \\
&< \varepsilon,
\end{aligned}$$

and thus $\Gamma(P, Q_1, Q_2)$ is tight.

For 2. we observe that for $i \in \{1,2\}$, $[\rho_n]_{0,i} \in \Pi^*(\nu^{(n)}, \mu_i^{(n)})$. Given that $\nu^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} \nu$ and $\mu_i^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} \mu_i$, and that $[\rho_n]_{0,i} \xrightarrow[n\longrightarrow+\infty]{w} \rho_{0,i}$, [Vil09, Theorem 5.20] shows that $\rho_{0,i} \in \Pi^*(\nu, \mu_i)$ (the result provides the existence of a subsequence converging to an element of $\Pi^*(\nu, \mu_i)$, then uniqueness of the limit shows $\rho_{0,i} \in \Pi^*(\nu, \mu_i)$), and we conclude that $\rho \in \Gamma(\nu, \mu_1, \mu_2)$ by definition.

For 3., take $(\rho_n) \in \Gamma(\nu, \mu_1, \mu_2)^{\mathbb{N}}$. By 1) and tightness of $\{\nu\}, \{\mu_1\}, \{\mu_2\}$, there exists an extraction α such that $\rho_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \rho \in \mathcal{P}_2(\mathbb{R}^{3d})$, then we show that $\rho \in \Gamma(\nu, \mu_1, \mu_2)$ using 2) with $\forall n \in \mathbb{N}$, $\nu^{(n)} := \nu$, $\mu_i^{(n)} := \mu_i$ for $i \in \{1,2\}$. □

Proposition B.II.1. For $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, it holds

$$\inf_{\rho \in \Gamma(\nu,\mu_1,\mu_2)} \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2) = \min_{\rho \in \Gamma(\nu,\mu_1,\mu_2)} \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2)$$

Proof. By Lemma B.II.1 item 3), $\Gamma(\nu, \mu_1, \mu_2)$ is a compact subset of $\mathcal{P}_2(\mathbb{R}^{3d})$. Then the map $J : \rho \in \mathcal{P}_2(\mathbb{R}^{3d}) \longmapsto \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2)$ is lower semi-continuous with respect to the weak convergence of measures ([San15, Lemma 1.6]), hence the infimum is attained. □

Another consequence of Lemma B.II.1 is that the ν-based Wasserstein distance is lower semi-continuous with respect to the weak convergence of measures, which is a property that was not studied in [NP23].

Proposition B.II.2. The map $(\nu, \mu_1, \mu_2) \in \mathcal{P}_2(\mathbb{R}^d)^3 \longmapsto \mathrm{W}_\nu(\mu_1, \mu_2)$ is lower semi-continuous with respect to the weak convergence of measures: for any $\nu^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu_i^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} \mu_i \in \mathcal{P}_2(\mathbb{R}^d)$, $i \in \{1,2\}$, we have:

$$\mathrm{W}_\nu(\mu_1, \mu_2) \leq \liminf_{n\longrightarrow+\infty} \mathrm{W}_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)}). \tag{B.II.13}$$

Proof. Without loss of generality, we can assume that

$$\mathrm{W}_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)}) \xrightarrow[n\longrightarrow+\infty]{} \liminf_{n\longrightarrow+\infty} \mathrm{W}_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)}),$$

up to considering an extraction of all sequences. For $n \in \mathbb{N}$, we can choose $\rho_n \in \Gamma(\nu^{(n)}, \mu_1^{(n)}, \mu_2^{(n)})$ optimal by Proposition B.II.1. By Lemma B.II.1 item 1) and tightness of the sets $\{\nu^{(n)}\}, \{\mu_1^{(n)}\}, \{\mu_2^{(n)}\}$, there exists an extraction α such that $\rho_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \rho \in \mathcal{P}_2(\mathbb{R}^{3d})$. and by Lemma B.II.1 item 2) we have $\rho \in \Gamma(\nu, \mu_1, \mu_2)$. By lower semi-continuity of the map $J : \rho \in \mathcal{P}_2(\mathbb{R}^{3d}) \longmapsto \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2)$ ([San15, Lemma 1.6]), we have:

$$\mathrm{W}_\nu^2(\mu_1, \mu_2) \leq J(\rho) \leq \liminf_{n\longrightarrow+\infty} J(\rho_{\alpha(n)}) = \liminf_{n\longrightarrow+\infty} \mathrm{W}^2_{\nu^{(\alpha(n))}}(\mu_1^{(\alpha(n))}, \mu_2^{(\alpha(n))}) = \liminf_{n\longrightarrow+\infty} \mathrm{W}^2_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)}),$$

where the first inequality follows from the definition of W_ν, since $\rho \in \Gamma(\nu, \mu_1, \mu_2)$ is admissible, and the second inequality follows from the lower semi-continuity of J. The first equality is due to the optimality of $\rho_{\alpha(n)}$, and the second equality follows from our reduction to the case where $\mathrm{W}_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)}) \xrightarrow[n\longrightarrow+\infty]{} \liminf_{n\longrightarrow+\infty} \mathrm{W}_{\nu^{(n)}}(\mu_1^{(n)}, \mu_2^{(n)})$. □

Full continuity with respect to the weak convergence of measures is not guaranteed, as shown in Example B.II.1.

Example B.II.1 ($\mathrm{W}_\nu(\cdot, \mu_2)$ is not continuous)**.** Consider the following empirical measures

in $\mathbb{R}^2$:

$$\begin{aligned}\nu &:= \tfrac{1}{2}(\delta_z + \delta_{z'}),\ z := (0,1),\ z' := (0,-1);\\ \mu_1^{(n)} &:= \tfrac{1}{2}(\delta_{x_n} + \delta_{x'}),\ x_n := (-1, 2^{-n}),\ x' := (1,0);\\ \mu_2 &= \tfrac{1}{2}(\delta_y + \delta_{y'}),\ y := (-2,-1),\ y := (2,1).\end{aligned}$$

For each $n \in \mathbb{N}$, we have $\Pi^*(\nu, \mu_1^{(n)}) = \{\gamma_1^{(n)}\}$ with $\gamma_1^{(n)} := \frac{1}{2}(\delta_{(z,x_n)} + \delta_{(z',x')})$. We also have $\Pi^*(\nu, \mu_2) = \{\gamma_2\}$ with $\gamma_2 := \frac{1}{2}(\delta_{(z,y')} + \delta_{(z',y)})$. This shows that $\Gamma(\nu, \mu_1^{(n)}, \mu_2) = \{\rho_n\}$ where $\rho_n := \frac{1}{2}(\delta_{(z,x_n,y')} + \delta_{(z',x',y)})$, yielding the cost

$$\mathrm{W}_\nu^2(\mu_1^{(n)}, \mu_2) = \tfrac{1}{2}\|x_n - y'\|_2^2 + \tfrac{1}{2}\|x' - y\|_2^2 = \tfrac{1}{2}\left(3^2 + (1-2^{-n})^2\right) + \tfrac{1}{2}\left(3^2 + 1^2\right) \xrightarrow[n\longrightarrow+\infty]{} 10.$$

However, we have $\mu_1^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} \mu_1 = \frac{1}{2}(\delta_x + \delta_{x'})$ with $x := (-1,0)$. We see that $\Pi^*(\nu, \mu_1) = \Pi(\nu, \mu_2)$, and clearly the choice $\gamma_1 := \frac{1}{2}(\delta_{(z,x')} + \delta_{(z',x)})$ will be optimal, such that $\rho := \frac{1}{2}(\delta_{(z,x',y')} + \delta_{(z',x,y)})$ is optimal for $\mathrm{W}_\nu^2(\mu_1, \mu_2) = 2 < \lim_{n\longrightarrow+\infty} \mathrm{W}_\nu^2(\mu_1^{(n)}, \mu_2) = 10$. We illustrate the setting of this example in Fig. B.II.2.

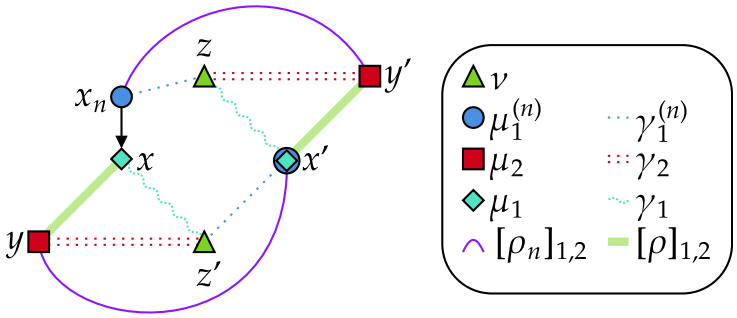


Figure B.II.2: Representation of Example B.II.1. The measure ν is drawn with green triangles, $\mu_1^{(n)}$ with blue circles, μ_2 with red squares, and the limit μ with light blue diamonds. The OT plan $\gamma_1^{(n)}$ between ν and $\mu_1^{(n)}$ is drawn with dotted blue lines, the OT plan γ_2 between ν and μ_2 with red double dotted lines, and the induced plan $[\rho_n]_{1,2}$ between $\mu_1^{(n)}$ and μ_2 with curved purple lines. As for the limit, an OT plan γ_1 between ν and μ_1 is drawn with curved dashed light blue lines, and the induced plan $[\rho]_{1,2}$ between μ_1 and μ_2 using γ_1 and γ_2 is drawn with thick green lines.

As remarked earlier (again, [AGS05, Remark 9.2.3]), if each $\Pi^*(\nu, \mu_i)$ is reduced to a single plan γ_i induced by T_i (for $i \in \{1,2\}$), then the only element $\rho \in \Gamma(\nu, \mu_1, \mu_2)$ is $\rho = (I, T_1, T_2)\#\nu$, yielding the following formulation for the ν-based Wasserstein distance (see also [NP23, Example 9] and the Linear OT framework [Wan+13]):

$$\mathrm{W}_\nu^2(\mu_1, \mu_2) = \int_{\mathbb{R}^d} \|T_1(y) - T_2(y)\|_2^2 \mathrm{d}\nu(y). \tag{B.II.14}$$

A result of interest is [AGS05, Lemma 9.2.1, Equation 9.2.7b], which states that for any $\rho \in \Gamma(\nu, \mu_1, \mu_2)$ (see Eq. (B.II.7))

$$\mathrm{W}_2^2(\mu_1, \mu_\rho^{1\to2}(t)) = (1-t)\mathrm{W}_2^2(\mu_1, \nu) + t\mathrm{W}_2^2(\mu_2, \nu) - (1-t)t \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2). \tag{B.II.15}$$

Taking in particular a 3-plan $\rho^* \in \Gamma(\nu, \mu_1, \mu_2)$ that is optimal for the ν-based Wasserstein

distance (Eq. (B.II.12)), we obtain

$$\mathrm{W}_2^2(\mu_1, \mu_{\rho^*}^{1\to 2}(t)) = (1-t)\mathrm{W}_2^2(\mu_1,\nu) + t\mathrm{W}_2^2(\mu_2,\nu) - (1-t)t\mathrm{W}_\nu^2(\mu_1,\mu_2). \tag{B.II.16}$$

B.II.2.3 Reminders on Wasserstein Means

A natural application of Wasserstein geodesics is the concept of Wasserstein means, which we will require in Section B.II.3. The following result states that Wasserstein means are exactly the middles of Wasserstein geodesics. For the sake of completeness, we provide some reminders on geodesic middles in the Appendix Section B.II.9.2, wherein we recall and prove an analogous result for geodesic spaces.

Proposition B.II.3. For $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the set of Wasserstein Means between μ_1 and μ_2

$$\mathrm{M}(\mu_1,\mu_2) := \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu_1,\mu) + \mathrm{W}_2^2(\mu,\mu_2) \tag{B.II.17}$$

can be expressed using Wasserstein geodesics Eq. (B.II.3):

$$\mathrm{M}(\mu_1,\mu_2) = \left\{\mu_\gamma^{1\to 2}(\tfrac{1}{2}) : \gamma \in \Pi^*(\mu_1,\mu_2)\right\} = \left\{\left(\tfrac{1}{2}P_1 + \tfrac{1}{2}P_2\right)\#\gamma : \gamma \in \Pi^*(\mu_1,\mu_2)\right\}. \tag{B.II.18}$$

Proof. The result is an application of Lemma B.II.12 in the geodesic space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$. □

B.II.2.4 Another Formulation of W_ν with Measure Disintegration

In this work, we will need a convenient formulation of the ν-based Wasserstein distance which uses the notion of disintegration of measures. We recall this notion in Section B.II.9.3, and provide a proof in Section B.II.9.4 of the Theorem by Nenna and Pass ([NP23, Theorem 12, item 1]), adapted to measures in $\mathcal{P}_2(\mathbb{R}^d)$. In Example B.II.2, we illustrate the result on a simple example with discrete measures.

Example B.II.2. We consider measures $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ as in Fig. B.II.3. We consider two optimal plans $\gamma_1 \in \Pi^*(\nu,\mu_1)$ and $\gamma_2 \in \Pi^*(\nu,\mu_2)$, represented in Fig. B.II.3. Writing the disintegrations as $\gamma_i(\mathrm{d}y, \mathrm{d}x_i) = \nu(\mathrm{d}y)\gamma_i^y(\mathrm{d}x_i)$, we can apply Theorem B.II.1 to compute $\mathrm{W}_\nu^2(\mu_1,\mu_2)$:

$$\begin{aligned}\mathrm{W}_\nu^2(\mu_1,\mu_2) &= \tfrac{1}{2}\mathrm{W}_2^2\left(\gamma_1^{z_1},\gamma_2^{z_1}\right) + \tfrac{1}{2}\mathrm{W}_2^2\left(\gamma_1^{z_2},\gamma_2^{z_2}\right)\\ &= \tfrac{1}{2}\mathrm{W}_2^2\left(\tfrac{1}{2}\delta_{x_1} + \tfrac{1}{2}\delta_{x_2}, \tfrac{1}{2}\delta_{y_1} + \tfrac{1}{2}\delta_{y_2}\right) + \tfrac{1}{2}\mathrm{W}_2^2\left(\tfrac{1}{2}\delta_{x_2} + \tfrac{1}{2}\delta_{x_3}, \tfrac{1}{2}\delta_{y_2} + \tfrac{1}{2}\delta_{y_3}\right).\end{aligned}$$

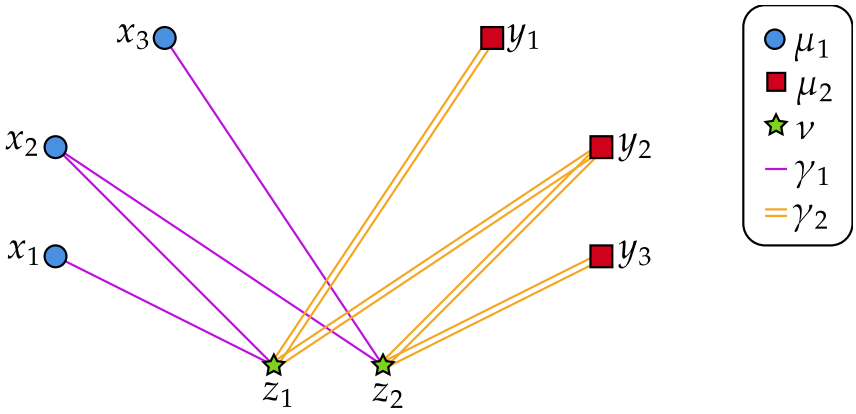


Figure B.II.3: In this example, there is a unique optimal transport plan γ_1 (purple lines) between μ_1 (blue circles) and the pivot ν (green stars), and likewise for γ_2 (orange double lines) between μ_2 (red squares) and ν. The disintegration kernel $\gamma_1^{z_1}$ in the disintegration $\gamma_1(\mathrm{d}z, \mathrm{d}x) = \nu(\mathrm{d}z)\gamma_1^z(\mathrm{d}x)$ is the probability measure $\gamma_1^{z_1} = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$, and likewise for $\gamma_1^{z_2}, \gamma_2^{z_1}, \gamma_2^{z_2}$.

Theorem B.II.1. [NP23, Theorem 12, item 1] Let $\nu, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. The following equality holds:

$$\mathrm{W}_\nu^2(\mu_1, \mu_2) = \min_{\gamma_i \in \Pi^*(\nu, \mu_i),\, i \in \{1,2\}} \int_{\mathbb{R}^d} \mathrm{W}_2^2(\gamma_1^y, \gamma_2^y)\mathrm{d}\nu(y), \tag{B.II.19}$$

where for $i \in \{1, 2\}$, $\gamma_i^y \in \mathcal{P}_2(\mathbb{R}^d)$ is defined using the disintegration $\gamma_i(\mathrm{d}y, \mathrm{d}x) = \nu(\mathrm{d}y)\gamma_i^y(\mathrm{d}x)$.

Proof. We provide a proof in Section B.II.9.4, which generalises that in [NP23] to measures in $\mathcal{P}_2(\mathbb{R}^d)$, following similar ideas. □

B.II.3 The Pivot Sliced Discrepancy

B.II.3.1 Definition with the ν-based Wasserstein Distance

We introduce a generalised version of SWGG introduced in [Mah+23] for general measures in $\mathcal{P}_2(\mathbb{R}^d)$ (and fixing the ambiguity issues that will be discussed in Example B.II.8), using the ν-based Wasserstein distance (Eq. (B.II.12), and see [NP23]), where the base measure ν is taken as a middle of projected versions of the measures:

Definition B.II.4. Let μ_1, μ_2 of $\mathcal{P}_2(\mathbb{R}^d)$, take $\mu_\theta \in \mathrm{M}(Q_\theta \# \mu_1, Q_\theta \# \mu_2)$, where $Q_\theta : x \longmapsto (\theta^\top x)\theta$. Then, we define

$$\mathrm{PS}_\theta(\mu_1, \mu_2) := \mathrm{W}_{\mu_\theta}(\mu_1, \mu_2). \tag{B.II.20}$$

We remind that we consider to related projection operations: $P_\theta : x \longmapsto \theta^\top x$ and $Q_\theta : x \longmapsto (\theta^\top x)\theta$. The first one is valued in $\mathbb{R}$, while the second is valued in $\mathbb{R}\theta \subset \mathbb{R}^d$. To fix ideas, we illustrate the definition of PS_θ in the case of discrete measures without projection ambiguity in Fig. B.II.4.

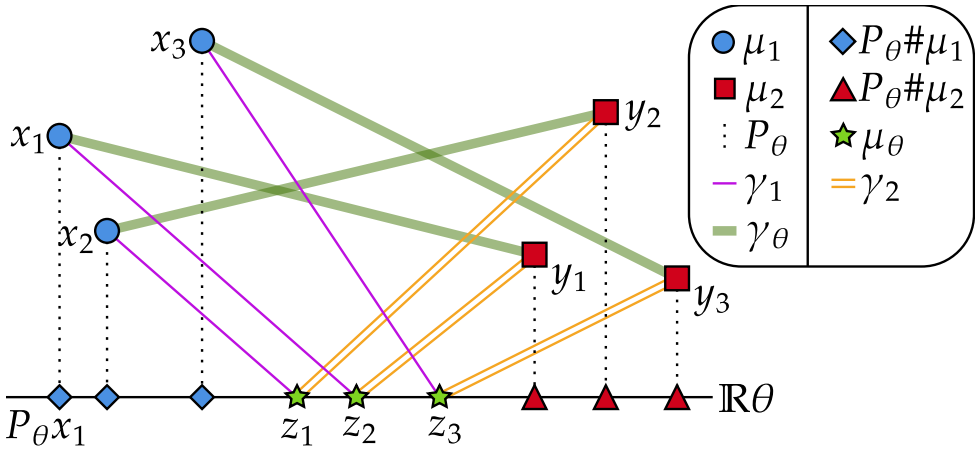


Figure B.II.4: Illustration of the definition of PS_θ in the case of discrete measures without projection ambiguity. The measure μ_1 is represented by blue circles, and μ_2 by red squares. The projected measures $Q_\theta\#\mu_1$ and $Q_\theta\#\mu_2$ are represented by blue diamonds and red triangles respectively. The middle μ_θ of the projections is represented by green stars. Once this middle is determined, we compute optimal transport plans γ_1, γ_2 between μ_θ and μ_1, μ_2 respectively (in this case, they are unique). We represent γ_1 by purple lines and γ_2 by orange double lines. To obtain the coupling corresponding to the cost $\mathrm{PS}_\theta(\mu_1, \mu_2)$, we look at the targets of each point (z_i) of the projected middle μ_θ: since z_1 is mapped to x_1 in μ_1 and to y_1 in μ_2, the coupling γ_θ maps x_1 to x_2, and so on. The coupling γ_θ is represented with thick green lines.

The idea of using a pivot measure is to find an optimal manner of correcting projection ambiguities. To illustrate this, we consider a simple pathological example in Fig. B.II.5, where the projections of the points of the support of μ_1 and μ_2 are not distinct.

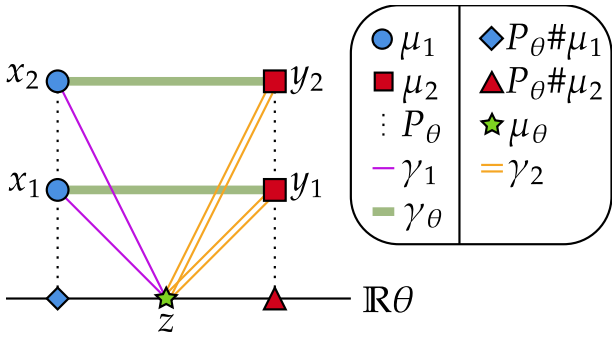


Figure B.II.5: In this example, we notice that $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$ are reduced to Dirac masses, thus their middle μ_θ is the middle Dirac mass. The optimal couplings γ_1 and γ_2 between μ_θ, μ_1 and μ_θ, μ_2 are then unique. It is then easy to see that the optimal $\rho \in \Gamma(\mu_\theta, \mu_1, \mu_2)$ is such that $\rho_{1,2} =: \gamma_\theta$ is the OT coupling between μ_1 and μ_2. In this example, $\mathrm{PS}_\theta(\mu_1, \mu_2) = \mathrm{W}_2(\mu_1, \mu_2)$.

In Fig. B.II.6, we illustrate another simple example where the projections of the points of the support of μ_1 are not distinct, but where they are distinct for μ_2.

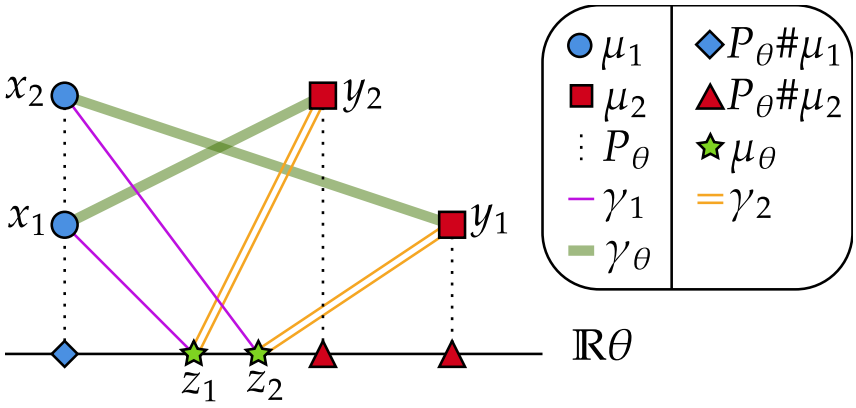


Figure B.II.6: In this illustration, $P_\theta\#\mu_1$ is a Dirac mass but not $P_\theta\#\mu_2$. Since we compare the middle μ_θ with μ_1 and not $P_\theta\#\mu_1$, there is in this case a unique optimal plan γ_1 between μ_θ and μ_1. The optimal plan γ_2 between μ_θ and μ_2 is also unique. The constraint $\rho_{0,1} = \gamma_1$, $\rho_{0,2} = \gamma_2$ imposes that $\rho_{1,2} = \frac{1}{2}\delta_{x_1\otimes y_2} + \frac{1}{2}\delta_{x_2\otimes y_1}$ for any $\rho \in \Gamma(\mu_\theta, \mu_1, \mu_2)$, hence there is no choice in the optimisation over ρ.

Remark B.II.1. As remarked by [NP23, Proposition 16], when ν is absolutely continuous with respect to the one-dimensional Hausdorff on a line, then the ν-based Wasserstein distance equates the *layer-wise Wasserstein metric* introduced by [KPS20]. We will see in Section B.II.4 that PS_θ equals another discrepancy that we call CW_θ, and this equality allows us to show that PS_θ satisfies the triangle inequality (and thus is a distance) on the set of measures with atomless projections, which is a stronger result than assuming absolute continuity of the pivot.

Note that this is a generalisation of SWGG introduced [Mah+23] in the sense that they show that their definition of SWGG coincides with the expression of Eq. (B.II.20) in Proposition 4.2. To prove that the quantity PS^2_θ is well-defined, which is to say that it does not depend on the choice of $\mu_\theta \in \mathrm{M}(Q_\theta\#\mu_1, Q_\theta\#\mu_2)$, we will show that in fact $\mathrm{M}(Q_\theta\#\mu_1, Q_\theta\#\mu_2)$ has only one element.

Lemma B.II.2. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\theta \in \mathbb{S}^{d-1}$. Then

$$\mathrm{M}(Q_\theta\#\mu_1, Q_\theta\#\mu_2) = \{\mu_\theta[\mu_1,\mu_2]\}\,, \quad \mu_\theta[\mu_1,\mu_2] := \left[\left(\tfrac{1}{2}F^{[-1]}_{\nu_1} + \tfrac{1}{2}F^{[-1]}_{\nu_2}\right)\theta\right]\#\mathscr{L}_{[0,1]}, \tag{B.II.21}$$

where for $i = 1, 2$, the measure ν_i is defined as $\nu_i = P_\theta\#\mu_i$, with $P_\theta : x \longmapsto \theta^\top x$, $\mathscr{L}_{[0,1]}$ the Lebesgue measure on $[0,1]$, and where $F^{[-1]}_\nu$ for $\nu \in \mathcal{P}(\mathbb{R})$ denotes the pseudo-inverse of its cumulative distribution:

$$\forall t \in [0,1], \quad F^{[-1]}_\nu(t) := \inf\{s \in \mathbb{R} : \nu((-\infty, s]) \geq t\}. \tag{B.II.22}$$

Proof. First, since the $Q_\theta\#\mu_i, i \in \{1,2\}$ are supported on $\mathbb{R}\theta$, we have

$$\mathrm{M}(Q_\theta\#\mu_1, Q_\theta\#\mu_2) = \{\theta\#\mu : \mu \in \mathrm{M}(P_\theta\#\mu_1, P_\theta\#\mu_2)\}\,, \tag{B.II.23}$$

which amounts to reducing a problem on a line of direction θ to a problem on $\mathbb{R}$, then embedding the result onto the line $\mathbb{R}\theta$. We introduce $\nu_i := P_\theta\#\mu_i$ for $i \in \{1,2\}$ and leverage Proposition B.II.3:

$$\mathrm{M}(\nu_1, \nu_2) = \left\{\left(\tfrac{1}{2}P_1 + \tfrac{1}{2}P_2\right)\#\gamma : \gamma \in \Pi^*(\nu_1, \nu_2)\right\}. \tag{B.II.24}$$

Since the ν_i are measures on $\mathbb{R}$, by [San15, Theorem 2.9], the set of optimal plans $\Pi^*(\nu_1, \nu_2)$ is reduced to the plan $(F^{[-1]}_{\nu_1}, F^{[-1]}_{\nu_2})\#\mathscr{L}_{[0,1]}$. Using Eq. (B.II.24) above and the projection embedding from Eq. (B.II.23), we obtain the result stated in Eq. (B.II.21). □

Remark B.II.2. Consider $\mu_1 = \frac{1}{n}\sum_i \delta_{x_i}$, $\mu_2 = \frac{1}{n}\sum_i \delta_{y_i}$, $\theta \in \mathbb{S}^{d-1}$ and $\sigma_\theta, \tau_\theta$ two permutations sorting respectively $(\theta^\top x_i)_i$ and $(\theta^\top y_i)_i$ (they may not be unique if the families $(\theta^\top x_i)_i$ and $(\theta^\top y_i)_i$ are not injective). Then the projected middle (computed using Eq. (B.II.21)) is explicit:

$$\mu_\theta[\mu_1, \mu_2] = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{\theta^\top\left(x_{\sigma_\theta(i)} + y_{\tau_\theta(i)}\right)}{2}\theta\right). \tag{B.II.25}$$

Note that measure above does not depend on the choice of the sorting permutations $(\sigma_\theta, \tau_\theta)$, since the families $(\theta^\top x_{\sigma_\theta(i)})_i$ and $(\theta^\top y_{\tau_\theta(i)})_i$ do not. This expression is specific to the case where μ_1 and μ_2 are uniform discrete measures with the same amount of atoms.

An interesting property of optimal transport between a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and another measure ν supported on a line $\mathbb{R}\theta$ is that the set of optimal plans and the cost can be related to the one-dimensional projections of μ and ν onto the line $\mathbb{R}\theta$. We remind $Q_\theta : x \longmapsto (\theta^\top x)\theta$, and introduce $Q_{\theta^\perp} := I - Q_\theta$. The following result is a generalisation of [Mah+23, Lemma 4.6], which was written in the case of uniform discrete measures. Note that the exponent 2 in the cost is paramount and allows the separation of orthogonal terms.

Proposition B.II.4. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ such that ν is supported on $\mathbb{R}\theta$, where $\theta \in \mathbb{S}^{d-1}$. Then for any plan $\gamma \in \Pi(\nu, \mu)$, we have

$$\int_{\mathbb{R}^{2d}} \|x - y\|_2^2 \mathrm{d}\gamma(y, x) = \int_{\mathbb{R}^{2d}} (\theta^\top(x - y))^2 \mathrm{d}\gamma(y, x) + \int_{\mathbb{R}^d} \|Q_{\theta^\perp} x\|_2^2 \mathrm{d}\mu(x), \tag{B.II.26}$$

with the alternate expression

$$\int_{\mathbb{R}^{2d}} (\theta^\top(x - y))^2 \mathrm{d}\gamma(y, x) = \int_{\mathbb{R}^2} (s - t)^2 \mathrm{d}(P_\theta, P_\theta)\#\gamma(s, t). \tag{B.II.27}$$

This yields the following expression for the OT cost:

$$\mathrm{W}_2^2(\nu, \mu) = \mathrm{W}_2^2(P_\theta\#\nu, P_\theta\#\mu) + \int_{\mathbb{R}^d} \|Q_{\theta^\perp} x\|_2^2 \mathrm{d}\mu(x), \tag{B.II.28}$$

and the following characterisation of the optimal plans:

$$\Pi^*(\nu, \mu) = \left\{\gamma \in \Pi(\nu, \mu) : (P_\theta, P_\theta)\#\gamma = \left(F^{[-1]}_{P_\theta\#\nu}, F^{[-1]}_{P_\theta\#\mu}\right)\#\mathscr{L}_{[0,1]}\right\}. \tag{B.II.29}$$

Proof. Let $\gamma \in \Pi(\nu, \mu)$. We have

$$\begin{aligned}\int_{\mathbb{R}^{2d}} \|x - y\|_2^2 \mathrm{d}\gamma(y, x) &= \int_{\mathbb{R}^{2d}} \left(\|Q_\theta(x - y)\|_2^2 + \|Q_{\theta^\perp}(x - y)\|_2^2\right) \mathrm{d}\gamma(y, x) \\ &= \int_{\mathbb{R}^{2d}} (\theta^\top(x - y))^2 \mathrm{d}\gamma(y, x) + \int_{\mathbb{R}^d} \|Q_{\theta^\perp} x\|_2^2 \mathrm{d}\mu(x),\end{aligned} \tag{B.II.30}$$

where the last equality comes from the fact that ν is supported on $\mathbb{R}\theta$ and that the second marginal of γ is μ. Since the second term does not depend on γ, by taking the infimum in Eq. (B.II.26), we obtain Eq. (B.II.28), where the equality

$$\inf_{\gamma \in \Pi(\nu,\mu)} \int_{\mathbb{R}^{2d}} (\theta^\top(x - y))^2 \mathrm{d}\gamma(y, x) = \mathrm{W}_2^2(P_\theta\#\nu, P_\theta\#\mu)$$

is justified by [DLV24, Lemma 2]. Furthermore, Eq. (B.II.26) shows that $(P_\theta, P_\theta)\#\Pi^*(\nu, \mu) \subset \Pi^*(P_\theta\#\nu, P_\theta\#\mu)$. Indeed, take $\gamma \in \Pi^*(\nu, \mu)$, then since $\pi_\theta := (P_\theta, P_\theta)\#\gamma \in \Pi(P_\theta\#\nu, P_\theta\#\mu)$, Eq. (B.II.26) yields the optimality of π_θ for the problem $\mathrm{W}_2^2(P_\theta\#\nu, P_\theta\#\mu)$. By [San15, Theorem 2.9],

$$\Pi^*(P_\theta\#\nu, P_\theta\#\mu) = \left\{\left(F^{[-1]}_{P_\theta\#\nu}, F^{[-1]}_{P_\theta\#\mu}\right)\#\mathscr{L}_{[0,1]}\right\} =: \{\pi_\theta^*\},$$

hence we have shown that $(P_\theta, P_\theta)\#\Pi^*(\nu,\mu) = \{\pi^*_\theta\}$. Conversely, take $\gamma \in \Pi(\nu,\mu)$ such that $(P_\theta, P_\theta)\#\gamma = \pi^*_\theta$, where π^*_θ is the unique element of $\Pi^*(P_\theta\#\nu, P_\theta\#\mu)$. Then by plugging γ into Eq. (B.II.26), we obtain $\gamma \in \Pi^*(\nu,\mu)$ using Eq. (B.II.28). We conclude that

$$\Pi^*(\nu,\mu) = \{\gamma \in \Pi(\nu,\mu) : (P_\theta, P_\theta)\#\gamma = \pi^*_\theta\}.$$

□

B.II.3.2 Semi-Metric Properties of PS_θ

We begin by stating straightforward properties of the discrepancy PS_θ:

Proposition B.II.5. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\theta \in \mathbb{S}^{d-1}$. Then the following properties hold:

- (Separation) $\mathrm{PS}_\theta(\mu_1,\mu_2) = 0$ if and only if $\mu_1 = \mu_2$.
- (Symmetry) $\mathrm{PS}_\theta(\mu_1,\mu_2) = \mathrm{PS}_\theta(\mu_2,\mu_1)$.
- (Upper-bound of W_2) $\mathrm{PS}_\theta(\mu_1,\mu_2) \geq \mathrm{W}_2(\mu_1,\mu_2)$.

Proof. If $\mathrm{PS}_\theta(\mu_1,\mu_2) = 0$, then by Proposition B.II.1, there exists $\rho \in \Gamma(\mu_\theta[\mu_1,\mu_2],\mu_1,\mu_2)$ such that $\int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y,x_1,x_2) = 0$, then in particular, for $\gamma := \rho_{1,2} \in \Pi(\mu_1,\mu_2)$, we have $\int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\gamma(x_1,x_2) = 0$ and thus $x_1 = x_2$ for γ-almost-every $(x_1,x_2) \in \mathbb{R}^{2d}$. For a test function $\phi \in \mathcal{C}^0_b$, we compute:

$$\int_{\mathbb{R}^d} \phi(x_1)\mathrm{d}\mu_1(x_1) = \int_{\mathbb{R}^d} \phi(x_1)\mathrm{d}\gamma(x_1,x_2) = \int_{\mathbb{R}^d} \phi(x_2)\mathrm{d}\gamma(x_1,x_2) = \int_{\mathbb{R}^d} \phi(x_2)\mathrm{d}\nu(x_2),$$

and thus $\mu_1 = \mu_2$. The converse is clear, but we detail for completeness. We show that $\mathrm{PS}_\theta(\mu,\mu) = 0$ for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$: notice that $\mu_\theta[\mu,\mu] = Q_\theta\#\mu =: \mu_\theta$, take any $\gamma \in \Pi^*(\mu_\theta,\mu)$ and introduce by disintegration $\rho(\mathrm{d}y,\mathrm{d}x_1,\mathrm{d}x_2) := \mu_\theta(\mathrm{d}y)\gamma^y(\mathrm{d}x_1)\delta_{x_1=x_2}(\mathrm{d}x_1,\mathrm{d}x_2)$. We have $\rho \in \Gamma(\mu_\theta,\mu,\mu)$, and for ρ-almost-every $(y,x_1,x_2) \in \mathbb{R}^{3d}$, we have $\|x_1 - x_2\|_2^2 = 0$, hence $\mathrm{PS}_\theta(\mu,\mu) = 0$.

Symmetry is immediate from the definition (Eq. (B.II.20)). As for the upper-bound, by Proposition B.II.1 we can take $\rho \in \Gamma(\mu_\theta[\mu_1,\mu_2],\mu_1,\mu_2)$ optimal for $\mathrm{PS}_\theta(\mu_1,\mu_2)$, and we have $\rho_{1,2} \in \Pi(\mu_1,\mu_2)$ and thus:

$$\mathrm{PS}^2_\theta(\mu_1,\mu_2) = \int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho_{1,2}(\mathrm{d}x_1,\mathrm{d}x_2) \geq \mathrm{W}^2_2(\mu_1,\mu_2).$$

□

The triangle inequality is not satisfied in general, as shown in Example B.II.3.

Example B.II.3 (PS_θ does not verify the triangle inequality)**.** We represent the counter-example in Fig. B.II.7. Consider $\theta := (1,0)$ and:

$$x_1 := (-1,0),\ x_2 := (1,5),\ \mu_1 := \tfrac{1}{2}\delta_{x_1} + \tfrac{1}{2}\delta_{x_2},$$
$$y_1 := (-1,5),\ y_2 := (1,0),\ \mu_2 := \tfrac{1}{2}\delta_{y_1} + \tfrac{1}{2}\delta_{y_2},$$
$$z_1 := (0,0),\ z_2 := (0,5),\ \mu_3 := \tfrac{1}{2}\delta_{z_1} + \tfrac{1}{2}\delta_{z_2}.$$

First, we compute $\mathrm{PS}_\theta(\mu_1,\mu_2)$: we have

$$u_1 := (-1,0),\ u_2 := (1,0),\ \mu_\theta[\mu_1,\mu_2] = \tfrac{1}{2}\delta_{u_1} + \tfrac{1}{2}\delta_{u_2},$$

and then we see that there are unique optimal plans between $\mu_\theta[\mu_1,\mu_2]$ and each μ_1,μ_2:

$$\Pi^*(\mu_\theta[\mu_1,\mu_2],\mu_1) = \left\{\gamma_{121} := \tfrac{1}{2}\delta_{u_1\otimes x_1} + \tfrac{1}{2}\delta_{u_2\otimes x_2}\right\},$$
$$\Pi^*(\mu_\theta[\mu_1,\mu_2],\mu_2) = \left\{\gamma_{122} := \tfrac{1}{2}\delta_{u_1\otimes y_1} + \tfrac{1}{2}\delta_{u_2\otimes y_2}\right\}.$$

Using Theorem B.II.1, we compute:

$$\mathrm{PS}_\theta(\mu_1,\mu_2) = \sqrt{\tfrac{1}{2}\|x_1 - y_1\|_2^2 + \tfrac{1}{2}\|x_2 - y_2\|_2^2} = 5.$$

We now turn to $\mathrm{PS}_\theta(\mu_1, \mu_3)$. This time, we have

$$v_1 := (-\tfrac{1}{2}, 0),\ v_2 := (\tfrac{1}{2}, 0),\ \mu_\theta[\mu_1, \mu_3] = \tfrac{1}{2}\delta_{v_1} + \tfrac{1}{2}\delta_{v_2}.$$

There is a unique optimal transport plan between $\mu_\theta[\mu_1, \mu_3]$ and μ_1:

$$\Pi^*(\mu_\theta[\mu_1, \mu_3], \mu_1) = \left\{\gamma_{131} := \tfrac{1}{2}\delta_{v_1 \otimes x_1} + \tfrac{1}{2}\delta_{v_2 \otimes x_2}\right\}.$$

On the other hand, there are an infinite number of OT between $\mu_\theta[\mu_1, \mu_3]$ and μ_3, which are convex combinations of two extremal plans (which correspond to the two permutations of $\{1, 2\}$):

$$\Pi^*(\mu_\theta[\mu_1, \mu_3], \mu_3) = \left\{\gamma_{133}(t) := (1-t)\left(\tfrac{1}{2}\delta_{v_1 \otimes z_1} + \tfrac{1}{2}\delta_{v_2 \otimes z_2}\right) + t\left(\tfrac{1}{2}\delta_{v_2 \otimes z_1} + \tfrac{1}{2}\delta_{v_1 \otimes z_2}\right),\ t \in [0, 1]\right\}.$$

Following Theorem B.II.1, we have

$$\mathrm{PS}_\theta(\mu_1, \mu_3) = \min_{t \in [0,1]} \tfrac{1}{2}\mathrm{W}_2^2\left(\delta_{x_1}, \tfrac{1-t}{2}\delta_{z_1} + \tfrac{t}{2}\delta_{z_2}\right) + \tfrac{1}{2}\mathrm{W}_2^2\left(\delta_{x_2}, \tfrac{1-t}{2}\delta_{z_2} + \tfrac{t}{2}\delta_{z_1}\right),$$

which is clearly minimal at $t = 0$, yielding

$$\mathrm{PS}_\theta(\mu_1, \mu_3) = \sqrt{\tfrac{1}{2}\|x_1 - z_1\|_2^2 + \tfrac{1}{2}\|x_2 - z_2\|_2^2} = 1,$$

and by symmetry, $\mathrm{PS}_\theta(\mu_2, \mu_3) = \mathrm{PS}_\theta(\mu_1, \mu_3) = 1$. We conclude that the triangle inequality does not hold:

$$\mathrm{PS}_\theta(\mu_1, \mu_2) = 5 > \mathrm{PS}_\theta(\mu_2, \mu_3) + \mathrm{PS}_\theta(\mu_1, \mu_3) = 2.$$

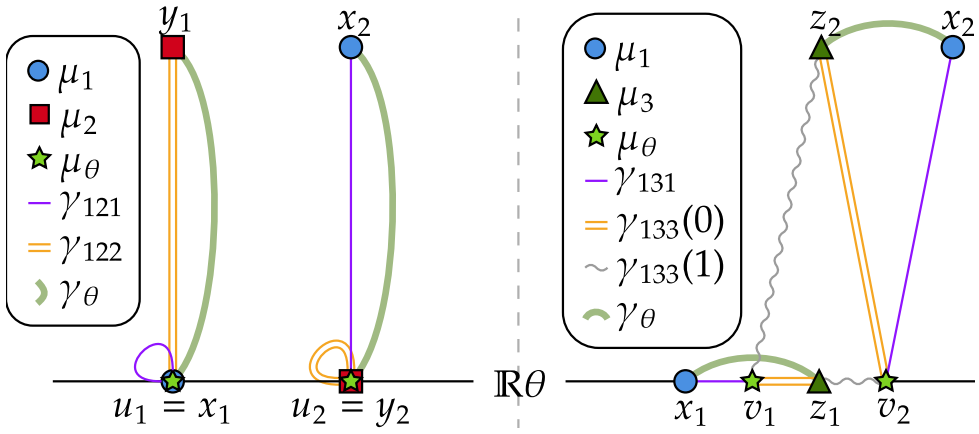


Figure B.II.7: Counter-example from Example B.II.3 to the triangle inequality for PS_θ. Left: illustration of the couplings for $\mathrm{PS}_\theta(\mu_1, \mu_2)$, with the optimal coupling γ_θ between μ_1 and μ_2 for $\mathrm{PS}_\theta(\mu_1, \mu_2)$ represented with thick green lines. Right: illustration of the couplings for $\mathrm{PS}_\theta(\mu_1, \mu_3)$. The optimal coupling γ_θ for $\mathrm{PS}_\theta(\mu_1, \mu_3)$ corresponds to gluing γ_{131} and $\gamma_{133}(0)$.

In the following, we show that PS_θ is lower semi-continuous with respect to the weak convergence of measures, along with a result on continuity of the middle $\mu_\theta[\mu_1, \mu_2]$. We speak of continuity with respect to the Euclidean topology on $\mathbb{S}^{d-1}$, and the weak topology on $\mathcal{P}_2(\mathbb{R}^d)$.

Proposition B.II.6. The map $(\theta, \mu_1, \mu_2) \longmapsto \mu_\theta[\mu_1, \mu_2]$ is continuous, and $(\theta, \mu_1, \mu_2) \longmapsto \mathrm{PS}_\theta(\mu_1, \mu_2)$ is lower semi-continuous.

Proof. Take measure sequences $\mu_1^{(n)} \xrightarrow[n \longrightarrow +\infty]{w} \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mu_2^{(n)} \xrightarrow[n \longrightarrow]{w} \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and a

sequence of projections $\theta_n \xrightarrow[n\longrightarrow+\infty]{} \theta \in \mathbb{S}^{d-1}$. By Lemma B.II.12, we have

$$\mu_{\theta_n}[\mu_1^{(n)}, \mu_2^{(n)}] = \left[\left(\tfrac{1}{2}F^{[-1]}_{P_{\theta_n}\#\mu_1^{(n)}} + \tfrac{1}{2}F^{[-1]}_{P_{\theta_n}\#\mu_2^{(n)}}\right)\theta_n\right]\#\mathscr{L}_{[0,1]} =: f_n\#\mathscr{L}_{[0,1]},$$

and:

$$\mu_\theta[\mu_1, \mu_2] = \left[\left(\tfrac{1}{2}F^{[-1]}_{P_\theta\#\mu_1} + \tfrac{1}{2}F^{[-1]}_{P_\theta\#\mu_2}\right)\theta\right]\#\mathscr{L}_{[0,1]} =: f\#\mathscr{L}_{[0,1]}.$$

Let $i \in \{1,2\}$ and $g_n := F^{[-1]}_{P_{\theta_n}\#\mu_i^{(n)}}$, we show that (g_n) converges pointwise $\mathscr{L}_{[0,1]}$-almost-everywhere to $g := F^{[-1]}_{P_\theta\#\mu_i}$. Since $P_{\theta_n}\#\mu_i^{(n)} \xrightarrow[n\longrightarrow+\infty]{w} P_\theta\#\mu_i$, we have by [Van00, Lemma 21.2] that for all $p \in [0,1]$ such that g is continuous at p, $g_n(p) \xrightarrow[n\longrightarrow+\infty]{} g(p)$. Since g is non-decreasing, it is continuous $\mathscr{L}_{[0,1]}$-almost-everywhere, and thus the convergence happens $\mathscr{L}_{[0,1]}$-almost-everywhere. Having shown that f_n converges pointwise to f $\mathscr{L}_{[0,1]}$-almost-everywhere, we deduce that

$$\mu_{\theta_n}[\mu_1^{(n)}, \mu_2^{(n)}] \xrightarrow[n\longrightarrow+\infty]{w} \mu_\theta[\mu_1, \mu_2].$$

By Proposition B.II.2, we deduce that $(\theta, \mu_1, \mu_2) \longmapsto \mathrm{PS}_\theta(\mu_1, \mu_2)$ is lower semi-continuous. □

We show in Example B.II.4 that full continuity does not holds.

Example B.II.4 (PS_θ is not continuous with respect to the weak convergence)**.** Consider

$$x_n := (-1-2^{-n}, 5),\ x' := (-1, 0),\ \mu_n := \tfrac{1}{2}\delta_{x_n} + \tfrac{1}{2}\delta_{x'},$$

$$y := (1,0),\ y' := (2,5),\ \nu := \tfrac{1}{2}\delta_y + \tfrac{1}{2}\delta_{y'}.$$

Obviously, $\mu_n \xrightarrow[n\longrightarrow+\infty]{w} \mu$, with $\mu = \frac{1}{2}\delta_{(-1,5)} + \frac{1}{2}\delta_{x'}$. For $n \in \mathbb{N}$, we compute easily that:

$$\mathrm{PS}^2_\theta(\mu_n, \nu) = \tfrac{1}{2}\|x_n - y'\|_2^2 + \tfrac{1}{2}\|x' - y\|_2^2 = 36 + 3.2^{-n} + \tfrac{4^{-n}}{2} \xrightarrow[n\longrightarrow+\infty]{} 36.$$

The limit does not coincide with $\mathrm{PS}^2_\theta(\mu, \nu) = \frac{13}{2}$. We summarise this counter-example in Fig. B.II.8.

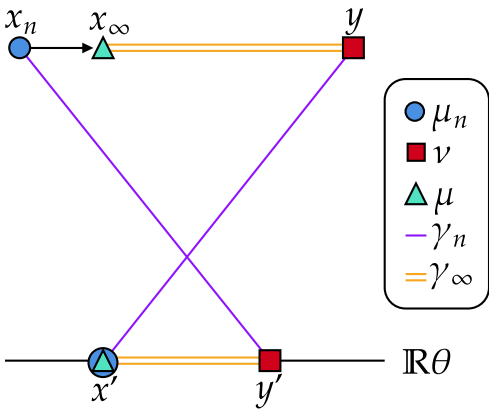


Figure B.II.8: Representation of Example B.II.4, showing a counter-example to the continuity of $\mathrm{PS}^2_\theta(\cdot, \nu)$ with respect to the weak convergence of measures. At each $n \in \mathbb{N}$, the coupling γ_n associated to $\mathrm{PS}^2_\theta(\mu_n, \nu)$ between μ_n and ν (represented by purple lines) is imposed to assign x_n to y' and x' to y. However, the coupling γ_∞ associated to $\mathrm{PS}^2_\theta(\mu, \nu)$ represented by orange double lines has more freedom due to the fact that $P_\theta x_\infty = P_\theta x'$, and therefore can perform the less costly assignment of x_∞ to y and x' to y'.

B.II.4 Correspondence of Pivot-Sliced and a Constrained Wasserstein Discrepancy

In this section, we will compare the quantity $\mathrm{PS}_\theta(\mu_1,\mu_2)$ defined in Eq. (B.II.20) with a particular lifting of the 1D sliced plan between μ_1 and μ_2. Namely, we will compare the two quantities:

$$\begin{aligned}\mathrm{PS}^2_\theta(\mu_1,\mu_2) &:= \min_{\rho\in\Gamma(\mu_\theta[\mu_1,\mu_2],\mu_1,\mu_2)} \int_{\mathbb{R}^{3d}} \|x_1-x_2\|_2^2 \mathrm{d}\rho(y,x_1,x_2)\\ \overset{?}{=}\mathrm{CW}^2_\theta(\mu_1,\mu_2) &:= \min_{\substack{\omega\in\Pi(\mu_1,\mu_2)\\(P_\theta,P_\theta)\#\omega=\pi_\theta[\mu_1,\mu_2]}} \int_{\mathbb{R}^{2d}} \|x_1-x_2\|_2^2 \mathrm{d}\omega(x_1,x_2),\end{aligned} \tag{B.II.31}$$

where $\pi_\theta[\mu_1,\mu_2]$ is the unique optimal plan between $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$. We introduce the following notation for the set of admissible plans ω for $\mathrm{CW}_\theta(\mu_1,\mu_2)$:

$$\Omega_\theta(\mu_1,\mu_2) := \{\omega\in\Pi(\mu_1,\mu_2) : (P_\theta,P_\theta)\#\omega = \pi_\theta[\mu_1,\mu_2]\}\,. \tag{B.II.32}$$

Note that by compactness of $\Pi(\mu_1,\mu_2)$, continuity of $\omega\longmapsto(P_\theta,P_\theta)\#\omega$ and lower semi-continuity of $J:=\omega\longmapsto\int_{\mathbb{R}^{2d}}\|\cdot-\cdot\|_2^2\mathrm{d}\omega$, the infimum in CW^2_θ is attained.

To draw a correspondence between PS_θ and CW_θ, we will compare their optimisation sets, and to this end, we introduce the set $\Gamma_{\theta,1,2}\subset\Pi(\mu_1,\mu_2)$ defined as:

$$\Gamma_{\theta,1,2}(\mu_1,\mu_2) := \{\rho_{1,2} : \rho\in\Gamma(\mu_\theta[\mu_1,\mu_2],\mu_1,\mu_2)\}\,. \tag{B.II.33}$$

We have by definition:

$$\mathrm{CW}^2_\theta(\mu_1,\mu_2) = \min_{\omega\in\Omega_\theta(\mu_1,\mu_2)} J(\omega);\quad \mathrm{PS}^2_\theta(\mu_1,\mu_2) = \min_{\gamma\in\Gamma_{\theta,1,2}(\mu_1,\mu_2)} J(\gamma). \tag{B.II.34}$$

B.II.4.1 First Inequality: $\mathrm{PS}_\theta \le \mathrm{CW}_\theta$

To prove a first inequality between PS_θ and CW_θ, we will show that $\Omega_\theta(\mu_1,\mu_2)\subset\Gamma_{\theta,1,2}(\mu_1,\mu_2)$ (these sets are defined in Eq. (B.II.32) and Eq. (B.II.33)). We start with two Lemmas on Wasserstein means. The first result provides an explicit optimal coupling between a Wasserstein mean and the two measures.

Lemma B.II.3. Let $\mu_1,\mu_2\in\mathcal{P}_2(\mathbb{R}^d)$ and an optimal coupling $\gamma\in\Pi^*(\mu_1,\mu_2)$.
Then $\mu_{\frac12} := ((1-t)P_1+tP_2)\#\gamma = \mathrm{Law}_{(X_1,X_2)\sim\gamma}\left[\frac{X_1+X_2}{2}\right]$ belongs to $\mathrm{M}(\mu_1,\mu_2)$, and furthermore the coupling $\gamma_{\frac12} := \mathrm{Law}_{(X_1,X_2)\sim\gamma}\left[\left(\frac{X_1+X_2}{2},X_1\right)\right]$ belongs to $\Pi^*(\mu_{\frac12},\mu_1)$.

Proof. By Proposition B.II.3 we have $\mu_{\frac12}\in M(\mu_1,\mu_2)$, and $\mathrm{W}_2^2(\mu_{\frac12},\mu_1)=\frac14\mathrm{W}_2^2(\mu_1,\mu_2)$. We compute:

$$\mathrm{W}_2^2(\mu_{\frac12},\mu_1)\le \mathop{\mathbb{E}}_{(X_1,X_2)\sim\gamma}\left[\left\|\tfrac{X_1+X_2}{2}-X_1\right\|_2^2\right] = \tfrac14\mathop{\mathbb{E}}_{(X_1,X_2)\sim\gamma}\left[\|X_1-X_2\|_2^2\right] = \tfrac14\mathrm{W}_2^2(\mu_1,\mu_2),$$

showing optimality of the coupling $\gamma_{\frac12}$, since by Proposition B.II.3, $\mathrm{W}_2^2(\mu_{\frac12},\mu_1)=\frac14\mathrm{W}_2^2(\mu_1,\mu_2)$. □

Note that Lemma B.II.3 is also a consequence of [AGS05, Lemma 7.2.1] (which states a stronger result with more abstract language). The following second lemma relates an admissible plan $\omega\in\Omega_\theta(\mu_1,\mu_2)$ for $\mathrm{CW}_\theta(\mu_1,\mu_2)$ to an explicit optimal coupling between the projected middle $\mu_\theta[\mu_1,\mu_2]$ and the measures μ_1,μ_2, which will be useful to construct an admissible 3-plan for $\mathrm{PS}_\theta(\mu_1,\mu_2)$.

Lemma B.II.4. Let $\omega \in \Pi(\mu_1, \mu_2)$ such that $(P_\theta, P_\theta)\#\omega = \Pi^*(P_\theta\#\mu_1, P_\theta\#\mu_2)$. Let $(X_1, X_2) \sim \omega$ and $Y := \frac{P_\theta X_1 + P_\theta X_2}{2}\theta$. Then $\mathrm{Law}[Y] = \mu_\theta[\mu_1, \mu_2]$, and $\mathrm{Law}[(Y, X_i)] \in \Pi^*(\mu_\theta[\mu_1, \mu_2], \mu_i)$, $i \in \{1, 2\}$.

Proof. First, we apply Lemma B.II.3 to the optimal coupling $(P_\theta X_1, P_\theta X_2)$, which shows that $\mathrm{Law}[P_\theta Y] = P_\theta\#\mu_\theta[\mu_1, \mu_2]$, thus that $\mathrm{Law}[Y] = \mu_\theta[\mu_1, \mu_2]$. Lemma B.II.3 also shows that $(P_\theta Y, P_\theta X_1)$ is the optimal coupling between $P_\theta\#\mu_\theta[\mu_1, \mu_2]$ and $P_\theta\#\mu_1$. Then by Eq. (B.II.29) in Proposition B.II.4, it follows that $\mathrm{Law}[(Y, X_1)] \in \Pi^*(\mu_\theta[\mu_1, \mu_2], \mu_1)$, and the same reasoning applies to $\mathrm{Law}[(Y, X_2)] \in \Pi^*(\mu_\theta[\mu_1, \mu_2], \mu_2)$. □

Using Lemma B.II.3 and Lemma B.II.4, we can now show an inequality between PS_θ and CW_θ:

Proposition B.II.7. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\theta \in \mathbb{S}^{d-1}$. The two sets defined in Eq. (B.II.32) and Eq. (B.II.33) verify $\Omega_\theta(\mu_1, \mu_2) \subset \Gamma_{\theta,1,2}(\mu_1, \mu_2)$, and the two quantities defined in Eq. (B.II.31) verify $\mathrm{PS}_\theta(\mu_1, \mu_2) \leq \mathrm{CW}_\theta(\mu_1, \mu_2)$.

Proof. Let $\omega \in \Pi(\mu_1, \mu_2)$ such that $(P_\theta, P_\theta)\#\omega = \Pi^*(P_\theta\#\mu_1, P_\theta\#\mu_2)$ optimal for $\mathrm{CW}_\theta(\mu_1, \mu_2)$. Consider $(X_1, X_2) \sim \omega$ and $Y := \frac{P_\theta X_1 + P_\theta X_2}{2}\theta$. By Lemma B.II.4, we have $\mathrm{Law}[Y] = \mu_\theta[\mu_1, \mu_2]$, and $\mathrm{Law}[(Y, X_i)] \in \Pi^*(\mu_\theta[\mu_1, \mu_2], \mu_i)$, $i \in \{1, 2\}$. By definition the 3-plan ρ defined by $\rho := \mathrm{Law}[(Y, X_1, X_2)]$ belongs to $\Gamma(\mu_\theta[\mu_1, \mu_2], \mu_1, \mu_2)$, thus $\omega \in \Gamma_{\theta,1,2}(\mu_1, \mu_2)$. We compute:

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) \leq \int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y, x_1, x_2) = \int_{\mathbb{R}^{2d}} \|x_2 - x_2\|_2^2 \mathrm{d}\omega(x_1, x_2) = \mathrm{CW}_\theta^2(\mu_1, \mu_2). \qquad \square$$

B.II.4.2 Converse Inequality: $\mathrm{PS}_\theta \geq \mathrm{CW}_\theta$

To show the converse inequality $\mathrm{CW}_\theta(\mu_1, \mu_2) \leq \mathrm{PS}_\theta(\mu_1, \mu_2)$, we will use more technical arguments from [AGS05, Lemma 7.2.1], which will show that (denoting $\mu_\theta := \mu_\theta[\mu_1, \mu_2]$,) for $i \in \{1, 2\}$, the unique optimal plan π_i between $P_\theta\#\mu_\theta$ and $P_\theta\#\mu_i$ is induced by a transport map T_i, i.e. $\pi_i = (I, T_i)\#P_\theta\#\mu_\theta$. This is a consequence of the fact that μ_θ is chosen as the middle of the geodesic between $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$, and remarkably holds without atomless assumptions on the $P_\theta\#\mu_i$.

Theorem B.II.2. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$. Then the two sets defined in Eq. (B.II.32) and Eq. (B.II.33) verify $\Gamma_{\theta,1,2}(\mu_1, \mu_2) = \Omega_\theta(\mu_1, \mu_2)$, and the two quantities defined in Eq. (B.II.31) verify $\mathrm{PS}_\theta(\mu_1, \mu_2) = \mathrm{CW}_\theta(\mu_1, \mu_2)$.

Proof. We have already shown that $\mathrm{PS}_\theta(\mu_1, \mu_2) \leq \mathrm{CW}_\theta(\mu_1, \mu_2)$ in Proposition B.II.7. We now show that $\Gamma_{\theta,1,2}(\mu_1, \mu_2) \subset \Omega_\theta(\mu_1, \mu_2)$ (we write $\mu_\theta := \mu_\theta[\mu_1, \mu_2]$, and π_θ the unique element of $\Pi^*(P_\theta\#\mu_1, P_\theta\#\mu_2)$). Let $\rho \in \Gamma(\mu_\theta, \mu_1, \mu_2)$. We introduce $\eta := (P_\theta, P_\theta, P_\theta)\#\rho \in \Pi(\nu_{\frac{1}{2}}, \nu_1, \nu_2)$, where for convenience we write $\nu_i := P_\theta\#\mu_i$ for $i \in \{1, 2\}$ and $\nu_{\frac{1}{2}} := P_\theta\#\mu_\theta$. By Eq. (B.II.29) in Proposition B.II.4, we have $\eta_{0,i} \in \Pi^*(\nu_{\frac{1}{2}}, \nu_i)$.

We now write η using OT maps. By [AGS05, Lemma 7.2.1], since $\nu_{\frac{1}{2}} = \mathrm{M}(\nu_1, \nu_2)$ (i.e. it is the middle of the constant-speed geodesic between ν_1 and ν_2, which is unique since the measures are one-dimensional), for $i \in \{1, 2\}$ the transport plan $\eta_{0,i} \in \Pi^*(\nu_{\frac{1}{2}}, \nu_i)$ is induced by a non-decreasing transport map T_i, which is to say that $\eta_{0,i} = (I, T_i)\#\nu_{\frac{1}{2}}$. It follows that for η-almost-every $(t, s_1, s_2) \in \mathbb{R}^3$, we have $s_1 = T_1(t)$ and $s_2 = T_2(t)$.

We now verify that $\eta_{1,2} \in \Pi^*(\nu_1, \nu_2)$ using the cyclical monotonicity criterion: Let $(s_1, s_2), (s_1', s_2') \in \mathrm{supp}\,\eta_{1,2}$ such that $s_1 < s_1'$. Our earlier considerations on η show that there exists $t, t' \in \mathbb{R}$ verifying $s_1 = T_1(t)$, $s_2 = T_2(t)$ and $s_1' = T_1(t')$, $s_2' = T_2(t')$. Since $s_1 = T_1(t) < T_1(t') = s_1'$ and T_1 is non-decreasing, we deduce $t < t'$. Now since T_2 is non-decreasing, $t < t'$ implies that $s_2 = T_2(t) \leq T_2(t') = s_2'$. We have shown the following property of $\eta_{1,2}$:

$$\forall (s_1, s_2), (s_1', s_2') \in \mathrm{supp}\,\eta_{1,2},\ s_1 < s_1' \implies s_2 \leq s_2'. \tag{B.II.35}$$

By [San15, Lemma 2.8], Eq. (B.II.35) implies that $\eta_{1,2}$ is the co-monotone plan between ν_1 and ν_2, and by [San15, Theorem 2.9], we conclude that $\eta_{1,2} \in \Pi^*(\nu_1, \nu_2)$.

Having shown that $\eta_{1,2} \in \Pi^*(\nu_1, \nu_2)$, we conclude that $(P_\theta, P_\theta)\#\rho_{1,2} \in \Pi^*(P_\theta\#\mu_1, P_\theta\#\mu_2)$, and by definition we conclude $\rho_{1,2} \in \Omega_\theta(\mu_1, \mu_2)$, which shows the inclusion $\Gamma_{\theta,1,2}(\mu_1, \mu_2) \subset \Omega_\theta(\mu_1, \mu_2)$, and equality is obtained by combining with Proposition B.II.7. By Eq. (B.II.34) we conclude that $\mathrm{CW}_\theta(\mu_1, \mu_2) = \mathrm{PS}_\theta(\mu_1, \mu_2)$. □

B.II.4.3 Triangle Inequality for PS_θ for Projection-Atomless Measures

Using Theorem B.II.2 and the following technical lemma on one-dimensional 3-plans, we will show that PS_θ is a metric on the set of probability measures whose projections on $\mathbb{R}\theta$ are atomless. We show that in the one-dimensional atomless case, 3-plans with two optimal bi-marginals automatically verify that *all* their bi-marginals are optimal. In terms of random variables, Lemma B.II.5 states that if (X_1, X_2) and (X_1, X_3) are optimal couplings, then so is (X_2, X_3).

Lemma B.II.5. Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R})$ such that μ_1 and μ_2 are atomless, and let $\rho \in \Pi(\mu_1, \mu_2, \mu_3)$ be a 3-plan such that $\rho_{1,2} \in \Pi^*(\mu_1, \mu_2)$ and $\rho_{1,3} \in \Pi^*(\mu_1, \mu_3)$. Then $\rho_{2,3} \in \Pi^*(\mu_2, \mu_3)$.

Proof. For $i \in \{1, 2, 3\}$, introduce the c.d.f. F_i of μ_i. Take $(X_1, X_2, X_3) \sim \rho$. By [San15, Theorem 2.9], since μ_1 is atomless and (X_1, X_3) is optimal, we have almost-surely $X_3 = F_3^{[-1]} \circ F_1(X_1)$. Likewise, since μ_2 is atomless and (X_2, X_1) is optimal, we have almost-surely $X_1 = F_1^{[-1]} \circ F_2(X_2)$. Combining these equalities yields almost-surely:

$$X_3 = F_3^{[-1]} \circ F_1(X_1) = F_3^{[-1]} \circ F_1 \circ F_1^{[-1]} \circ F_2(X_2).$$

By continuity of F_1 (since μ_1 is atomless) and defining $F_1(-\infty) := 0$ and $F_1(+\infty) := 1$, we have $F_1(\mathbb{R}) \cup \{F_1(-\infty)\} \cup \{F_1(+\infty)\} = [0, 1]$, allowing us to apply [EH13, Proposition 2.3, item 4)], which yields $F_1 \circ F_1^{[-1]} = I_{[0,1]}$.

We have shown that almost-surely $X_3 = F_3^{[-1]} \circ F_2(X_2)$, which shows by [San15, Theorem 2.9] that (X_2, X_3) is the optimal coupling between μ_2 and μ_3. □

We can now show that PS_θ verifies the triangle inequality on the set of probability measures with atomless projections. Combining this statement with Proposition B.II.5 shows that PS_θ is a metric this subset of $\mathcal{P}_2(\mathbb{R}^d)$.

Proposition B.II.8. Let $\theta \in \mathbb{S}^{d-1}$ and $\mathcal{P}_{2,a}(\mathbb{R}^d, \theta)$ be the set of probability measures $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $P_\theta\#\mu$ is atomless. The quantity PS_θ is a metric on $\mathcal{P}_{2,a}(\mathbb{R}^d, \theta)$.

Proof. First, by Proposition B.II.5, it only remains to show the triangle inequality. Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_{2,a}(\mathbb{R}^d, \theta)$ and let $\omega_{1,2} \in \Omega_\theta(\mu_1, \mu_2)$ be optimal for $\mathrm{CW}_\theta(\mu_1, \mu_2)$, and likewise let $\omega_{2,3} \in \Omega_\theta(\mu_2, \mu_3)$ be optimal for $\mathrm{CW}_\theta(\mu_2, \mu_3)$. We apply the standard gluing technique (see for example [San15, Lemma 5.5]): the second marginal of $\omega_{1,2}$ and the first marginal of $\omega_{2,3}$ are both μ_2, hence we can write their disintegrations with respect to μ_2 as:

$$\omega_{1,2}(\mathrm{d}x_1, \mathrm{d}x_2) = \mu_2(\mathrm{d}x_2)\omega_{1,2}^{x_2}(\mathrm{d}x_1), \quad \omega_{2,3}(\mathrm{d}x_2, \mathrm{d}x_3) = \mu_2(\mathrm{d}x_2)\omega_{2,3}^{x_2}(\mathrm{d}x_3).$$

We now introduce the "composition" 3-plan $\rho \in \Pi(\mu_1, \mu_2, \mu_3)$ as:

$$\rho(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) := \mu_2(\mathrm{d}x_2)\omega_{1,2}^{x_2}(\mathrm{d}x_1)\omega_{2,3}^{x_2}(\mathrm{d}x_3).$$

Writing $\rho_\theta := (P_\theta, P_\theta, P_\theta)\#\rho$, by definition of $\omega_{1,2}$ and $\omega_{2,3}$, we have $[\rho_\theta]_{1,2} = \pi_\theta[\mu_1, \mu_2]$ and $[\rho_\theta]_{2,3} = \pi_\theta[\mu_2, \mu_3]$. By Lemma B.II.5, we deduce that $\rho_{1,3} = \pi_\theta[\mu_1, \mu_3]$, since each $P_\theta\#\mu_i$ is

atomless. This shows that $\rho_{1,3} \in \Omega_\theta(\mu_1, \mu_2)$. Denoting $\phi_i := (x_1, x_2, x_3) \longmapsto x_i$ for $i \in \{1, 2, 3\}$, we have:

$$\begin{aligned}\mathrm{CW}_\theta(\mu_1, \mu_3) &\leq \|\phi_1 - \phi_3\|_{L^2(\rho_{1,3})} = \|\phi_1 - \phi_3\|_{L^2(\rho)} \\ &\leq \|\phi_1 - \phi_2\|_{L^2(\rho)} + \|\phi_2 - \phi_3\|_{L^2(\rho)} = \|\phi_1 - \phi_2\|_{L^2(\omega_{1,2})} + \|\phi_2 - \phi_3\|_{L^2(\omega_{2,3})} \\ &= \mathrm{CW}_\theta(\mu_1, \mu_2) + \mathrm{CW}_\theta(\mu_2, \mu_3).\end{aligned}$$

Using Theorem B.II.2, we deduce the triangle inequality for PS_θ. □

B.II.5 A Monge Formulation of PS_θ Between Point Clouds

B.II.5.1 The Case of Non-Ambiguous Projections

A direct consequence of Theorem B.II.1 is that, in the case of point clouds with non-ambiguous projections, the computation of PS_θ can be done simply by sorting the projections and taking the associated plan between the projected measures.

Corollary B.II.1. Let $(x_1, \cdots, x_n) \in (\mathbb{R}^d)^n$ and $(y_1, \cdots, y_n) \in (\mathbb{R}^d)^n$, and $\theta \in \mathbb{S}^{d-1}$ such that the families $(P_\theta x_i)_i$ and $(P_\theta y_i)_i$ are injective. Then for the measures $\mu_1 := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\mu_2 := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, it holds

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) = \frac{1}{n}\sum_{i=1}^n \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2,$$

where PS_θ is introduced in Definition B.II.4 and where $\sigma_\theta, \tau_\theta$ are the (unique) permutations sorting $(P_\theta x_i)$ and $(P_\theta y_i)$. For injective families (x_i) and (y_i), the injectivity assumptions holds for $\mathcal{U}(\mathbb{S}^{d-1})$-almost-every $\theta \in \mathbb{S}^{d-1}$.

Proof. We begin under the injectivity assumptions, which allows us to define $\sigma_\theta, \tau_\theta$ as the (unique) permutations sorting $(P_\theta x_i)$ and $(P_\theta y_i)$ respectively, and for $i \in [\![1, n]\!]$, let $z_i := \frac{1}{2}P_\theta(x_{\sigma_\theta(i)} + y_{\tau_\theta(i)})$. We remark that the family (z_i) is increasing by construction (we provide further details at the end of the proof for almost-sure injectivity) and denote $\nu := \frac{1}{n}\sum_i \delta_{z_i}$. Let $\gamma_1 \in \Pi^*(\nu, \mu_1)$, and write

$$\gamma_1 = \sum_{i,j} A_{i,j}\delta_{(z_i, x_{\sigma_\theta(j)})}.$$

By Proposition B.II.4 and the injectivity assumptions, we have $(P_\theta, P_\theta)\#\gamma_1 = \frac{1}{n}\sum_i \delta_{(P_\theta z_i, P_\theta x_{\sigma_\theta(i)})}$, and thus injectivity allows us to identify the coefficients $A_{i,j}$, yielding $\gamma_1 = \frac{1}{n}\sum_i \delta_{(z_i, x_{\sigma_\theta(i)})}$, and in particular, for any $i \in [\![1, n]\!]$, $\gamma_1^{z_i} = \delta_{x_{\sigma_\theta(i)}}$. The same reasoning applies to $\gamma_2 \in \Pi^*(\nu, \mu_2)$, and thus Theorem B.II.1 yields

$$\mathrm{PS}_\theta^2(\mu_1, \mu_2) = \mathrm{W}_\nu^2(\mu_1, \mu_2) = \frac{1}{n}\sum_{i=1}^n \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2$$

Regarding the almost-sure injectivity claim, assume now that the families (x_i) and (y_i) are injective, and take $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$. Then $(P_\theta x_i)$ is almost-surely injective, since $\mathbb{P}(P_\theta x_i = P_\theta x_j) = \mathbb{P}(\theta \in (x_i - x_j)^\top)$. The same reasoning applies to $(P_\theta y_i)$, and the injectivity of (z_i) comes from the fact that almost-surely, for $i < j$, we have $P_\theta x_{\sigma_\theta(i)} < P_\theta x_{\sigma_\theta(j)}$, and $P_\theta y_{\tau_\theta(i)} < P_\theta x_{\tau_\theta(j)}$, hence by sum $z_i < z_j$, almost-surely. □

B.II.5.2 Problem Formulation and Reduction to Sorted Projections

The natural question that arises is the impact of projection ambiguity, i.e. non-injectivity of $(P_\theta x_i)$ or $(P_\theta y_j)$. In this section, we will start from the equality $\mathrm{PS}_\theta = \mathrm{CW}_\theta$ from Theorem B.II.2,

to provide the following Monge formulation of PS_θ between point clouds (without injectivity assumptions), that we will prove in Theorem B.II.4:

$$\mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{(\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2,$$

where $\mathfrak{S}_\theta(X,Y)$ is the set of pairs of permutations (σ,τ) such that σ sorts $(P_\theta x_i)_{i=1}^n$ and τ sorts $(P_\theta y_i)_{i=1}^n$, for given $X := (x_1,\cdots,x_n) \in \mathbb{R}^{n\times d}$ and $Y := (y_1,\cdots,y_n) \in \mathbb{R}^{n\times d}$:

$$\mathfrak{S}_\theta(X,Y) := \mathfrak{S}_\theta(X)\times\mathfrak{S}_\theta(Y),\ \mathfrak{S}_\theta(X) := \left\{\sigma\in\mathfrak{S}_n : \forall i\in[\![1,n-1]\!],\ P_\theta x_{\sigma(i)} \le P_\theta x_{\sigma(i+1)},\right\}, \tag{B.II.36}$$

with an analogous definition for $\mathfrak{S}_\theta(Y)$. We will reduce to the case where the identity permutation sorts the projections $(P_\theta x_i)_{i=1}^n$ and $(P_\theta y_i)_{i=1}^n$, which will greatly simplify notation and proofs. The following Lemma states that re-labelling the points does not change the value of CW_θ and of the Monge formulation.

Lemma B.II.6. Let $X,Y \in \mathbb{R}^{n\times d}$ and $\sigma_0,\tau_0 \in \mathfrak{S}_n$. Denote by $X\circ\sigma_0 := (x_{\sigma_0(1)},\cdots,x_{\sigma_0(n)})_{i=1}^n$ and likewise $Y\circ\tau_0 := (y_{\tau_0(1)},\cdots,y_{\tau_0(n)})_{i=1}^n$. Then we have:

$$\mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_{\sigma_0(i)}}, \frac{1}{n}\sum_{j=1}^n \delta_{y_{\tau_0(j)}}\right),$$

and for the Monge formulation, we have the following cost equality:

$$\min_{(\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)}-y_{\tau(i)}\|_2^2 = \min_{(\sigma,\tau)\in\mathfrak{S}_\theta(X\circ\sigma_0,Y\circ\tau_0)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma_0\circ\sigma(i)}-y_{\tau_0\circ\tau(i)}\|_2^2. \tag{B.II.37}$$

Proof. The first equality is simply a consequence of the equality between measures:

$$\frac{1}{n}\sum_{i=1}^n \delta_{x_i} = \frac{1}{n}\sum_{i=1}^n \delta_{x_{\sigma_0(i)}}, \quad \frac{1}{n}\sum_{j=1}^n \delta_{y_j} = \frac{1}{n}\sum_{j=1}^n \delta_{y_{\tau_0(j)}}.$$

For the second equality, notice that a permutation $\sigma\in\mathfrak{S}_n$ sorts $(P_\theta x_{\sigma_0(i)})_{i=1}^n$ if and only if $P_\theta x_{\sigma_0\circ\sigma(1)} \le \cdots \le P_\theta x_{\sigma_0\circ\sigma(n)}$ if and only if $\sigma_0\circ\sigma$ sorts $(P_\theta x_i)_{i=1}^n$, thus we obtain:

$$\mathfrak{S}_\theta(X\circ\sigma_0, Y\circ\tau_0) = \left\{(\sigma_0^{-1}\circ\sigma, \tau_0^{-1}\circ\tau),\ (\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)\right\}.$$

Eq. (B.II.37) follows by change of variables. □

Thanks to Lemma B.II.6, we can assume without loss of generality (for the cost values) that the identity permutation sorts the projections $(P_\theta x_i)_{i=1}^n$ and $(P_\theta y_i)_{i=1}^n$. We formulate this assumption as follows:

Assumption B.II.1. The points $X,Y\in\mathbb{R}^{n\times d}$ and $\theta\in\mathbb{S}^{d-1}$ are such that:

$$P_\theta x_1 \le \cdots \le P_\theta x_n \text{ and } P_\theta y_1 \le \cdots \le P_\theta y_n.$$

Assumption B.II.1 holds up to relabelling the points (x_i) and (y_j): taking $\sigma\in\mathfrak{S}_n$ sorting $(P_\theta x_i)$ and $\tau\in\mathfrak{S}_n$ sorting $(P_\theta y_j)$, the relabelled points $\widetilde{X} := (x_{\sigma(i)}) =: X\circ\sigma$ and $\widetilde{Y} := (y_{\tau(j)}) =: Y\circ\tau$ verify the condition.

B.II.5.3 A Kantorovich Formulation of CW_θ Between Point Clouds

We begin by a characterisation of $\mathfrak{S}_\theta(X,Y)$, which states that a pair (σ,τ) belongs to $\mathfrak{S}_\theta(X,Y)$ if and only if each "ambiguity" set $\{i : P_\theta x_i = t\}$ is stable by σ and likewise for τ.

Lemma B.II.7. Let $X := (x_1, \cdots, x_n) \in \mathbb{R}^{n\times d}$, $Y := (y_1, \cdots, y_n) \in \mathbb{R}^{n\times d}$ and $\theta \in \mathbb{S}^{d-1}$ verifying Assumption B.II.1. Let $A := \#\{P_\theta x_i\}_{i=1}^n$ and $B := \#\{P_\theta y_j\}_{j=1}^n$. Write $\{P_\theta x_i\}_{i=1}^n = (s_a)_{a=1}^A$ where $s_1 < \cdots < s_A$ and likewise $\{P_\theta y_j\}_{j=1}^n = (t_b)_{b=1}^B$ where $t_1 < \cdots < t_B$. Introduce the "ambiguity group" sets:

$$\forall a \in [\![1, A]\!],\ I_a := \{i \in [\![1, n]\!] : P_\theta x_i = s_a\}, \quad \forall b \in [\![1, B]\!],\ J_b := \{j \in [\![1, n]\!] : P_\theta y_j = t_b\}. \tag{B.II.38}$$

Then the set $\mathfrak{S}_\theta(X, Y)$ can be re-written as follows:

$$\mathfrak{S}_\theta(X, Y) = \left\{(\sigma, \tau) \in \mathfrak{S}_n^2 : \forall a \in [\![1, A]\!],\ \sigma(I_a) = I_a,\ \forall b \in [\![1, B]\!],\ \tau(J_b) = J_b\right\}. \tag{B.II.39}$$

Proof. We show the property $\mathfrak{S}_\theta(X) = \widetilde{\mathfrak{S}} := \{\sigma \in \mathfrak{S}_n : \forall a \in [\![1, A]\!],\ \sigma(I_a) = I_a\}$ by double inclusion. First, since $s_1 < \cdots < s_A$, the inclusion $\widetilde{\mathfrak{S}} \subset \mathfrak{S}_\theta(X)$ is clear. For the converse inclusion, take $\sigma \in \mathfrak{S}_\theta(X)$. By definition and by Assumption B.II.1, we have:

$$P_\theta x_1 \leq \cdots \leq P_\theta x_n; \quad P_\theta x_{\sigma(1)} \leq \cdots \leq P_\theta x_{\sigma(n)},$$

which implies that $\forall i \in [\![1, n]\!],\ P_\theta x_i = P_\theta x_{\sigma(i)}$. Take now a the unique element of $[\![1, A]\!]$ such that $i \in I_a$. We have $s_a = P_\theta x_i = P_\theta x_{\sigma(i)}$ and thus $\sigma(i) \in I_a$, and we conclude that $\sigma \in \widetilde{\mathfrak{S}}$. The proof for $\mathfrak{S}_\theta(Y)$ follows verbatim, and Eq. (B.II.39) follows from the definition (see Eq. (B.II.36)). □

To illustrate Lemma B.II.7, we consider an example with projection ambiguities in Fig. B.II.9.

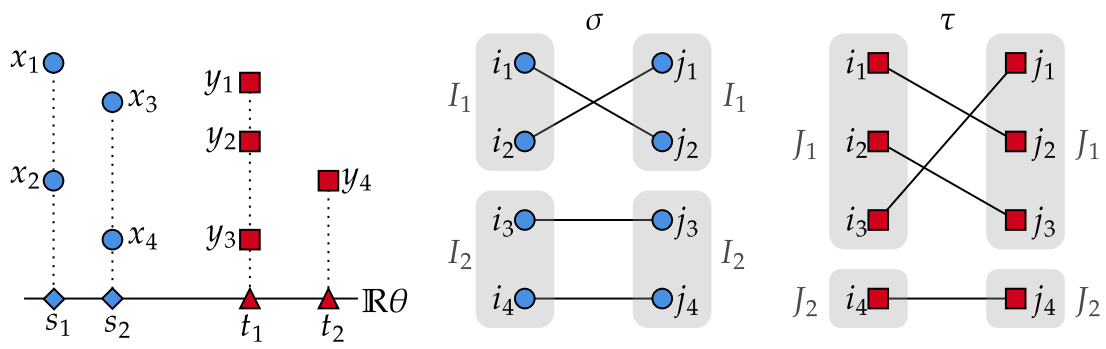


Figure B.II.9: In this example we consider two discrete uniform measures with $n := 4$ points with projection ambiguity: $s_1 := P_\theta x_1 = P_\theta x_2 < s_2 := P_\theta x_3 = P_\theta x_4$ and $t_1 := P_\theta y_1 = P_\theta y_2 = P_\theta y_3 < t_2 := P_\theta x_4$. In the notation of Lemma B.II.7, we have $A = B = 2$ and $I_1 = \{i_1, i_2\}$, $I_2 = \{i_3, i_4\}$, $J_1 = \{j_1, j_2, j_3\}$ and $J_2 = \{j_4\}$. We consider a permutation pair $(\sigma, \tau) \in \mathfrak{S}_\theta(X, Y)$, specifically $\sigma := (2, 1, 3, 4)$ and $\tau := (2, 3, 1, 4)$. We see that σ sorts the sequence $(P_\theta x_i)_{i=1}^n$ and that I_1 and I_2 are stable by σ, and likewise for τ.

We now write a discrete Kantorovich formulation of CW_θ between point clouds, whose expression we will be able to simplify later. The main idea is that transport plans P are constrained to exchange exactly as much mass between I_a and J_b as the one-dimensional OT plan π_θ between $P_\theta\#\mu$ and $P_\theta\#\nu$ sends from s_a to t_b, as illustrated in Fig. B.II.10.

Proposition B.II.9. Under Assumption B.II.1, let $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu := \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ be empirical measures. Then the CW_θ discrepancy introduced in Eq. (B.II.31) has the following expression:

$$\mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{P \in \mathbb{U} \cap \mathcal{P}_\theta(X,Y)} \sum_{i=1}^n\sum_{j=1}^n \|x_i - y_j\|_2^2 P_{i,j}, \tag{B.II.40}$$

$$\mathbb{U} := \{P \in \mathbb{R}_+^{n\times n} : P\mathbf{1} = \tfrac{1}{n}\mathbf{1},\ P^\top\mathbf{1} = \tfrac{1}{n}\mathbf{1}\}, \tag{B.II.41}$$

$$\mathcal{P}_\theta(X,Y) := \left\{ P \in \mathbb{R}^{n\times n} : \forall (a,b) \in [\![1,A]\!] \times [\![1,B]\!], \sum_{(i,j)\in I_a \times J_b} P_{i,j} = \frac{1}{n}\#(I_a \cap J_b) \right\}. \tag{B.II.42}$$

Proof. Fix $\omega \in \Omega_\theta(\mu,\nu)$ (see Eq. (B.II.32)). By the marginal constraints, we have $\operatorname{supp}\omega \subset \{(x_i,y_j)\}_{i,j}$, allowing us to define $P \in \mathbb{R}_+^{n\times n}$ by $\forall (i,j) \in [\![1,n]\!]^2$, $P_{i,j} = \omega(\{(x_i,y_j)\})$. Since $\omega \in \Pi(\mu,\nu)$, we verify immediately that $P \in \mathbb{U}$. As for the constraint $(P_\theta, P_\theta)\#\omega = \pi_\theta[\mu,\nu] =: \pi_\theta$, notice that by construction (see the notations in Lemma B.II.7), we can write by [San15, Theorem 2.9] for *any* $(\sigma,\tau) \in \mathfrak{S}_\theta(X,Y)$ that $\pi_\theta = \frac{1}{n}\sum_i \delta_{(P_\theta x_{\sigma(i)}, P_\theta y_{\tau(i)})}$. Since $\{P_\theta x_i\}_{i=1}^n = (s_a)_{a=1}^A$ and $\{P_\theta y_j\}_{j=1}^n = (t_b)_{b=1}^B$, it follows that for any $(a,b) \in [\![1,A]\!] \times [\![1,B]\!]$:

$$\pi_\theta(\{(s_a,t_b)\}) = \sum_{k=1}^n \frac{1}{n}\mathbb{1}(\sigma(k) \in I_a)\mathbb{1}(\tau(k) \in J_b) = \sum_{k=1}^n \frac{1}{n}\mathbb{1}(k \in I_a \cap J_b) = \frac{1}{n}\#(I_a \cap J_b),$$

where we have used that $\sigma(I_a) = I_a$ and $\tau(J_b) = J_b$, which holds by Lemma B.II.7. We can now show that $P \in \mathcal{P}_\theta(X,Y)$ using the constraint $(P_\theta, P_\theta)\#\omega = \pi_\theta$:

$$\frac{1}{n}\#(I_a \cap J_b) = \pi_\theta(\{(s_a,t_b)\}) = (P_\theta,P_\theta)\#\omega(\{(s_a,t_b)\}) = \sum_{(i,j)\in I_a\times J_b} P_{i,j}.$$

The cost $\int_{\mathbb{R}^{2d}} \|x-y\|_2^2 \mathrm{d}\omega(x,y)$ writes $\sum_{i,j}\|x_i - y_j\|_2^2 P_{i,j}$ by definition of P. Conversely, it can readily be checked with the same computations that for any $P \in \mathbb{U} \cap \mathcal{P}_\theta(X,Y)$, the coupling $\omega := \sum_{i,j} P_{i,j}\delta_{(x_i,y_j)}$ belongs to $\Omega_\theta(\mu,\nu)$, and yields the same transportation cost. We conclude that the equality in Eq. (B.II.40) holds. □

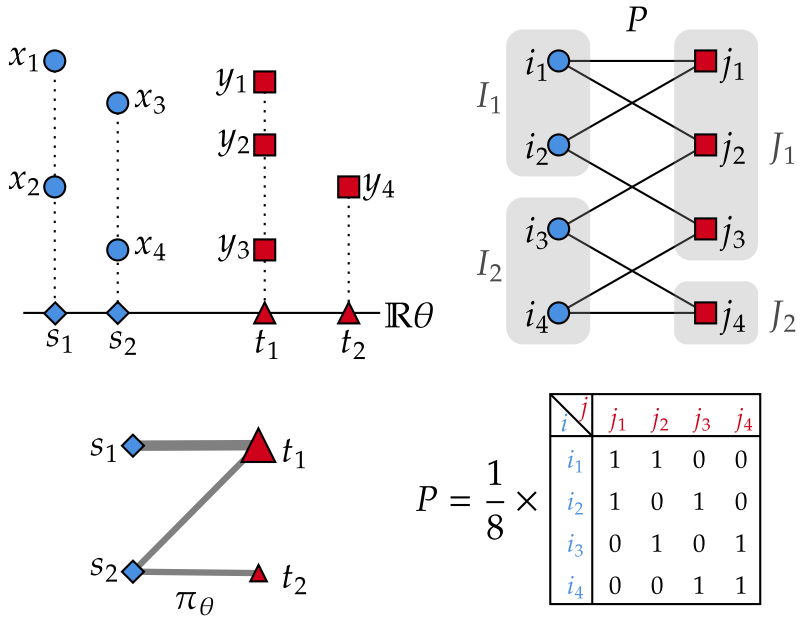


Figure B.II.10: We continue with the example from Fig. B.II.9 and illustrate the unique optimal transport plan $\pi_\theta = \frac{1}{2}\delta_{(s_1,t_1)} + \frac{1}{4}\delta_{(s_2,t_1)} + \frac{1}{4}\delta_{(s_2,t_2)}$ between $P_\theta\#\mu = \frac{1}{2}\delta_{s_1} + \frac{1}{2}\delta_{s_2}$ and $P_\theta\#\nu = \frac{3}{4}\delta_{t_1} + \frac{1}{4}\delta_{t_2}$. We show a particular transport plan $P \in \mathbb{U}\cap\mathcal{P}_\theta(X,Y)$ which is not a permutation matrix. For the constraints, notice for example that $\pi_\theta(\{(s_1,t_1)\}) = \frac{\#(I_1\cap J_1)}{4} = \frac{1}{2} = \sum_{i=1}^2\sum_{j=1}^3 P_{i,j}$.

The discrete problem in Eq. (B.II.40) can be seen as a constrained Kantorovich problem. Our goal is now to show that it admits a constrained Monge formulation, which is to say a minimisation over the constrained set of permutations $\mathfrak{S}_\theta(X,Y)$. To show this, we will adapt the proof of the Birkhoff Von Neumann Theorem [Bir46] (see also [Pey19] Theorem 2 and [Hur08] for other proofs which inspired our method). Our objective is now to build up definitions and technical

lemmas to adapt Birkhoff's Theorem, and prove the generalisation stated in Theorem B.II.3. We will consider particular elements of $\mathbb{U}$ called permutation matrices:

$$\forall(\alpha,\beta)\in\mathfrak{S}_n^2,\ P^{\alpha,\beta}:=\left[\frac{1}{n}\mathbb{1}(\alpha(i)=\beta(j))\right]_{i,j}. \tag{B.II.43}$$

This method of writing permutation matrices differs from the usual $P^{\sigma}_{i,j}:=\frac{1}{n}\mathbb{1}(\sigma(i)=j)$, and will be more convenient for our purposes. An elementary property of permutation matrices is that:

$$\forall(\alpha,\beta,\varphi)\in\mathfrak{S}_n^3,\ P^{\alpha,\beta}=P^{\varphi\circ\alpha,\varphi\circ\beta}, \tag{B.II.44}$$

since $\varphi\circ\alpha(i)=\varphi\circ\beta(j)\iff\alpha(i)=\beta(j)$. For $S\subset\mathfrak{S}_n^2$, we will write $\mathcal{P}^S:=\{P^{\alpha,\beta}:(\alpha,\beta)\in S\}$. The Birkhoff Von Neumann Theorem [Bir46] states that $\operatorname{Extr}\mathbb{U}=\mathcal{P}^{\mathfrak{S}_n^2}$, where the set of extreme points of a convex set is defined as:

Definition B.II.5. The set of extreme points $\operatorname{Extr}C$ of a convex set C is the set of points $c\in C$ that cannot be written $c=\frac{1}{2}a+\frac{1}{2}b$ for some $a,b\in C$.

Our objective is to show that $\operatorname{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y))=\mathcal{P}^{\mathfrak{S}_\theta(X,Y)}$. We begin with a Lemma showing a condition for $P^{\alpha,\beta}$ to belong to $\mathcal{P}_\theta(X,Y)$.

Lemma B.II.8. Under Assumption B.II.1, for any $(\alpha,\beta)\in\mathfrak{S}_n^2$, we have

$$P^{\alpha,\beta}\in\mathcal{P}_\theta(X,Y)\iff\exists\varphi\in\mathfrak{S}_n:(\varphi\circ\alpha,\varphi\circ\beta)\in\mathfrak{S}_\theta(X,Y).$$

In other words, $P^{\alpha,\beta}\in\mathcal{P}_\theta(X,Y)$ if and only if $P^{\alpha,\beta}=P^{\sigma,\tau}$ for some $(\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)$.

Proof. Suppose that $P^{\alpha,\beta}\in\mathcal{P}_\theta(X,Y)$. Applying the definition of $\mathcal{P}_\theta(X,Y)$ from Eq. (B.II.42), we see that (using the notation of Lemma B.II.7):

$$\forall(a,b)\in[\![1,A]\!]\times[\![1,B]\!],\ \sum_{(i,j)\in I_a\times J_b}\frac{\mathbb{1}(\alpha(i)=\beta(j))}{n}=\frac{\#(I_a\cap J_b)}{n},\ \text{thus}\ \#(\alpha(I_a)\cap\beta(J_b))=\#(I_a\cap J_b).$$

Let $E:=\{(a,b):I_a\cap J_b\neq\varnothing\}$. For any $(a,b)\in E$, we have $\#(\alpha(I_a)\cap\beta(J_b))=\#(I_a\cap J_b)$, and thus we can introduce a bijection $\varphi_{a,b}:\alpha(I_a)\cap\beta(J_b)\longrightarrow I_a\cap J_b$. We have the partition $[\![1,n]\!]=\cup_{(a,b)\in E}I_a\cap J_b$ where the union is disjoint and the elements are non-empty. Since α,β are permutations and by the property $\#(\alpha(I_a)\cap\beta(J_b))=\#(I_a\cap J_b)$, we have the partition $[\![1,n]\!]=\cup_{(a,b)\in E}\alpha(I_a)\cap\beta(J_b)$, again with disjoint unions and non-empty terms. We can define $\psi:[\![1,n]\!]\longrightarrow E$ a map such that $\forall i\in[\![1,n]\!]$, $i\in\alpha(I_a)\cap\beta(J_b)$ where $\psi(i)=(a,b)$. The map $\varphi:=i\longmapsto\varphi_{\psi(i)}(i)$ is therefore well-defined, we verify easily that it is a permutation of $[\![1,n]\!]$ using the partition $[\![1,n]\!]=\cup_{(a,b)\in E}\alpha(I_a)\cap\beta(J_b)$.

We now fix $a\in[\![1,A]\!]$ and show that $\varphi\circ\alpha(I_a)=I_a$. Let $i\in I_a$, we have $\alpha(i)\in\alpha(I_a)$, and there exists (a unique) $b\in[\![1,B]\!]$ such that $\alpha(i)\in\alpha(I_a)\cap\beta(J_b)$. By definition, we get $\psi(\alpha(i))=(a,b)$, and thus $\varphi(\alpha(i))=\varphi_{a,b}(\alpha(i))\in I_a\cap J_b\subset I_a$. We conclude that $\varphi\circ\alpha(I_a)\subset I_a$ and similarly that $\varphi\circ\beta(J_b)\subset J_b$ for any $b\in[\![1,B]\!]$. By Lemma B.II.7, we conclude that $(\varphi\circ\alpha,\varphi\circ\beta)\in\mathfrak{S}_\theta(X,Y)$, concluding the "left to right" implication.

Conversely, let $\varphi\in\mathfrak{S}_n$ and $(\sigma,\tau)\in\mathfrak{S}_\theta(X,Y)$. Notice that $P^{\sigma,\tau}=P^{\varphi\circ\sigma,\varphi\circ\tau}$ by Eq. (B.II.44). We check that $P^{\sigma,\tau}\in\mathcal{P}_\theta(X,Y)$ by applying the definition: let $(a,b)\in[\![1,A]\!]\times[\![1,B]\!]$, we have:

$$\sum_{(i,j)\in I_a\times J_b}P^{\sigma,\tau}_{i,j}=\frac{\#(\sigma(I_a)\cap\tau(J_b))}{n}=\frac{\#(I_a\cap J_b)}{n},$$

where we used that $\sigma(I_a)=I_a$ and $\tau(J_b)=J_b$, which is a consequence of Lemma B.II.7. □

B.II.5.4 Technical Lemmas on Bipartite Graphs Associated to Couplings

To study the extreme points of $\mathbb{U} \cap \mathcal{P}_\theta(X, Y)$ we will adapt the techniques from [Pey19; Hur08] and consider the bipartite graph associated to a matrix in $P \in \mathbb{R}_+^{n \times m}$, which we define in Definition B.II.6.

Definition B.II.6. The bipartite directed graph G_P associated to a matrix $P \in \mathbb{R}_+^{n \times m}$ is the graph with vertices $V_P := \{i_k,\ k \in [\![1, n]\!]\} \cup \{j_l,\ \ell \in [\![1, m]\!]\}$ and directed edges:

$$E_P := \{(i_k, j_l),\ (k, \ell) \in [\![1, n]\!] \times [\![1, m]\!],\ : P_{i_k, j_l} > 0\} \cup \{(j_l, i_k),\ (k, \ell) \in [\![1, n]\!] \times [\![1, m]\!],\ P_{i_k, j_l} > 0\}\,.$$

By slight abuse of notation, we will often write $\{i_k,\ k \in [\![1, n]\!]\}$ as simply $[\![1, n]\!]$ and $\{j_l,\ \ell \in [\![1, m]\!]\}$ as $[\![1, m]\!]$, seeing them as disjoint sets of labels. The i's will be called "left" vertices, and the j's "right" vertices. Edges (i, j) being directed from left to right, we will call them "right" edges, and likewise edges (j, i) will be referred to as "left" edges. We continue the example from Fig. B.II.10 in Fig. B.II.11 showing the bipartite graph associated to the matrix P.

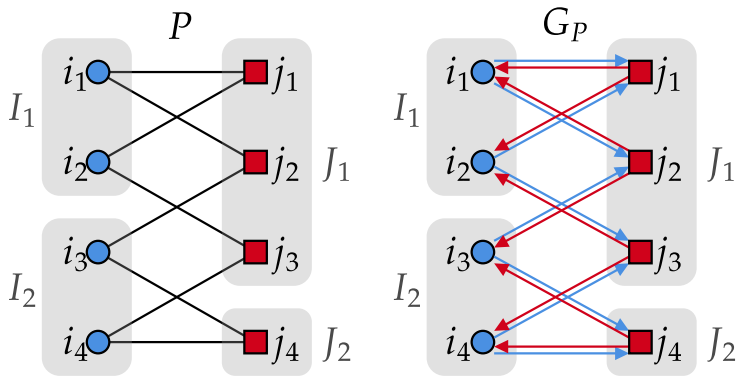


Figure B.II.11: We consider the matrix P from Fig. B.II.10 and show the associated bipartite graph G_P. The "right" edges from an $i \in [\![1, n]\!]$ on the left to a $j \in [\![1, n]\!]$ on the right are represented in blue, and the "left edges" are represented in red. Note that by construction, for each (i, j) such that $P_{i,j} > 0$, there is both a left edge (i, j) and a right edge (j, i) in G_P.

In the following, we will extract a particular cycle $i_1 \to j_1 \to i_2 \to \cdots \to i_{p+1} = i_1$ from the graph G_P of an element $P \in \mathbb{U} \setminus \mathcal{P}^{\mathfrak{S}_n^2}$. In proofs of Birkhoff's theorem, this is commonly used to show that P is not an extreme point of $\mathbb{U}$. In our setting, we will also make use of this property, in addition to strategies specific to $\mathcal{P}_\theta(X, Y)$.

Lemma B.II.9. Assume $n \geq 2$ and let $P \in \mathbb{U} \setminus \mathcal{P}^{\mathfrak{S}_n^2}$.
Then there exists a cycle $(i_1, j_1, \cdots, i_p, j_p, i_{p+1}) \in [\![1, n]\!]^{2p+1}$ in G_P with $p \geq 2$ verifying:

$$\begin{aligned} &i_{p+1} = i_1;\ (i_k)_{k=1}^p \text{ and } (j_k)_{k=1}^p \text{ are injective;} \\ &\text{and } \forall k \in [\![1, p]\!],\ P_{i_k, j_k} \in (0, \tfrac{1}{n}),\ P_{i_{k+1}, j_k} \in (0, \tfrac{1}{n}). \end{aligned} \tag{B.II.45}$$

Proof. First, we show a weaker result:

$$\begin{aligned} &\exists p \geq 2,\ \exists (i_1, j_1, \cdots, i_p, j_p, i_{p+1}) \in [\![1, n]\!]^{2p+1} \\ &\text{such that } i_{p+1} = i_1;\ \forall k \in [\![1, p]\!],\ i_k \neq i_{k+1},\ \forall k \in [\![1, p-1]\!],\ j_k \neq j_{k+1}; \\ &\text{and } \forall k \in [\![1, p]\!],\ P_{i_k, j_k} \in (0, \tfrac{1}{n}),\ P_{i_{k+1}, j_k} \in (0, \tfrac{1}{n}). \end{aligned} \tag{B.II.46}$$

Since $P \in \mathbb{U} \setminus \mathcal{P}^{\mathfrak{S}_n^2}$, there exists $(i_1, j_1) \in [\![1, n]\!]^2$ such that $P_{i_1, j_1} \in (0, \frac{1}{n})$. Since $0 < P_{i_1, j_1} < \sum_i P_{i, j_1} = \frac{1}{n}$, there exists $i_2 \neq i_1$ such that $P_{i_2, j_1} \in (0, \frac{1}{n})$. Likewise, since $0 < P_{i_2, j_1} < \sum_j P_{i_2, j} = \frac{1}{n}$, there exists $j_2 \neq j_1$ such that $P_{i_2, j_2} \in (0, \frac{1}{n})$. We continue and show the existence of $i_3 \neq i_2$ such that $P_{i_3, j_2} \in (0, \frac{1}{n})$. So far, we have built a chain $i_1 \to j_1 \to i_2 \to j_2 \to i_3$. If $i_3 = i_1$ then

we have shown Eq. (B.II.46). Otherwise we continue the process up to i_k, $k \geq 4$ while $i_k \neq i_1$, and there are two exclusive possibilities:

1. The process terminates with $(i_1, j_1, \cdots, i_p, j_p, i_{p+1})$ such that $i_{p+1} = i_1$, and by construction the cycle verifies the conditions of Eq. (B.II.46);

2. The process continues at least up to $k = n+1$, yielding $(i_1, j_1, \cdots, i_n, j_n, i_{n+1})$ verifying the conditions of Eq. (B.II.46) except $i_{n+1} \neq i_1$. Then by the pigeonhole principle, there exists $k_1 < k_2 \in [\![1, n+1]\!]^2$ such that $i_{k_1} = i_{k_2}$. Consider the cycle $(i_{k_1}, j_{k_1}, i_{k_1+1}, j_{k_1+1}, \cdots, i_{k_2})$, it verifies Eq. (B.II.46) (the length is sufficient since by construction $i_{k_1} \neq i_{k_1+1}$).

Now that we have shown Eq. (B.II.46), we deduce Eq. (B.II.45) by taking $p \geq 2$ minimal in Eq. (B.II.46). □

As an illustration, in Fig. B.II.11, by following the edges of G_P starting from the edge (i_1, j_1), we observe the cycle $(i_1, j_1, i_2, j_3, i_4, j_4, i_3, j_2, i_1)$ which satisfies the criteria of Eq. (B.II.45).

We will also require the following technical result about extracting injective cycles from (possibly) redundant cycles in a graph. For a set S and $n, m \in \mathbb{N}$, we say that two families $(s_i)_{i=1}^n \in S^n$ and $(t_j)_{j=1}^n \in S^n$ are equipotent if $n = m$ and there exists a permutation $\varphi \in \mathfrak{S}_n$ such that $\forall i \in [\![1, n]\!]$, $s_{\varphi(i)} = t_i$. We write this property $(s_i) \simeq (t_j)$. This concept is particularly useful when the families are not injective, which will sometimes be the case in the following.

Lemma B.II.10. Let $G := (V := \mathcal{A} \cup \mathcal{B}, E)$ be a directed bipartite graph, set $p \geq 1$ and consider a cycle written $(a_1, b_1, \cdots, a_p, b_p, a_{p+1}) \in V^{2p+1}$ with $\forall k \in [\![1, p]\!]$, $(a_k, b_k) \in E$, $(b_k, a_{k+1}) \in E$. Then there exists $L \geq 1$ cycles of G of the form $(a_1^\ell, b_1^\ell, \cdots, a_{p_\ell}^\ell, b_{p_\ell}^\ell, a_{p_\ell+1}^\ell)$ (with $a_1^\ell = a_{p_\ell+1}^\ell$ and each $(a_k^\ell, b_k^\ell), (b_k^\ell, a_{k+1}^\ell) \in E$) whose combined elements (without the last vertex) are exactly the elements of $(a_1, b_1, \cdots, a_p, b_p)$:

$$(a_1, b_1, \cdots, a_p, b_p) \simeq (a_1^1, b_1^1, \cdots, a_{p_1}^1, b_{p_1}^1, \cdots\cdots, a_1^L, b_1^L, \cdots, a_{p_L}^L, b_{p_L}^L), \tag{B.II.47}$$

and such that for each $\ell \in [\![1, L]\!]$, the families of edges $((a_k^\ell, b_k^\ell))_{k=1}^{p_\ell}$ and $((b_k^\ell, a_{k+1}^\ell))_{k=1}^{p}$ are injective.

Proof. Given such a cycle $\mathcal{C} := (a_1, b_1, \cdots, a_p, b_p, a_{p+1})$ we consider the two following "splitting" operators:

- Split_R takes the first pair $i < j \in [\![1, p]\!]^2$ (for the lexicographic order) such that $(a_i, b_i) = (a_j, b_j)$ if such a pair (i, j) exists (if not, $\text{Split}_R(\mathcal{C})$ returns $\mathcal{C}$). $\text{Split}_R(\mathcal{C})$ then returns the two following sub-cycles:

  $$\mathcal{C}_1 := (a_1, b_1, \cdots, a_{i-1}, b_{i-1}, a_j, b_j, \cdots, a_p, b_p, a_{p+1}),\ \mathcal{C}_2 := (a_i, b_i, \cdots, , a_{j-1}, b_{j-1}, a_j).$$

  Obviously, their concatenation without endpoints is exactly $\mathcal{C}$ without its endpoint:

  $$(a_1, b_1, \cdots, a_{i-1}, b_{i-1}, a_j, b_j, \cdots, a_p, b_p, a_i, b_i, \cdots, , a_{j-1}, b_{j-1}) \simeq (a_1, b_1, \cdots, a_p, b_p).$$

- Split_L takes the first pair $i < j \in [\![1, p]\!]^2$ such that $(b_i, a_{i+1}) = (b_j, a_{j+1})$ if such a pair (i, j) exists (if not, $\text{Split}_L(\mathcal{C})$ returns $\mathcal{C}$). $\text{Split}_L(\mathcal{C})$ then returns the two following sub-cycles, (which verify the equipotence condition):

  $$\mathcal{C}_1 := (a_1, b_1, \cdots, a_i, b_i, a_{j+1}, b_{j+1}, \cdots, a_p, b_p, a_{p+1}),\ \mathcal{C}_2 := (a_{i+1}, b_{i+1}, \cdots, a_j, b_j, a_{j+1}).$$

To split an initial $\mathcal{C}$, we construct a family $(\mathcal{C}_\ell)$ of cycles iteratively starting with $(\mathcal{C})$ by applying Split_R and Split_L to the cycles to the family $\mathcal{C}_\ell$ until no cycle can be split. This process terminates since each iteration increases the number of cycles (they are non-empty), which is bounded because $\mathcal{C}$ is finite and the concatenation of the cycles $(\mathcal{C}_\ell)$ without endpoints is exactly $\mathcal{C}$ without its endpoint. At the end of the process, the equipotence condition remains and each cycle $\mathcal{C}_\ell$ has injective edges $((a_k^\ell, b_k^\ell))_k, ((b_k^\ell, a_{k+1}^\ell))_k$ since the splitting process could not continue. □

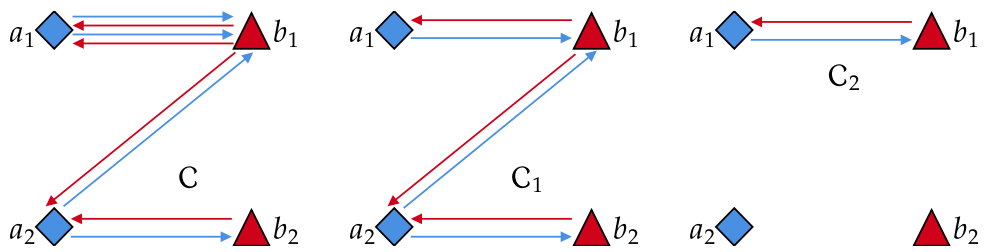


Figure B.II.12: Extracting two cycles $\mathcal{C}_1, \mathcal{C}_2$ from the cycle $\mathcal{C}$ such that each cycle $\mathcal{C}_\ell$ has distinct (directed) edges.

In Fig. B.II.12 we illustrate the splitting process of Lemma B.II.10. The cycle from Fig. B.II.12 is a cycle of $G_{\overline{P}}$ (where $\overline{P}$) is the OT plan matrix between the measures $P_\theta \# \mu$ and $P_\theta \# \nu$, constructed using P from Fig. B.II.11 with $a_1 \in [\![1, A]\!]$ where $i \in I_{a_1}$, then $b_1 \in [\![1, B]\!]$ such that $j_1 \in I_{b_1}$ and so one. This example case is paramount since it will be the use case of Lemma B.II.10 in the proof of Theorem B.II.3.

The following lemma is in essence a cyclical monotonicity property, and concerns a property of cycles in the bipartite graph associated to the matrix $\overline{P} \in \mathbb{R}_+^{A \times B}$ which is the unique optimal transport plan matrix between the one-dimensional measures $P_\theta \# \mu = \sum_{a=1}^{A} \frac{\#I_a}{n} \delta_{s_a}$ and $P_\theta \# \nu = \sum_{b=1}^{B} \frac{\#J_b}{n} \delta_{t_b}$. The idea is that by monotonicity of $\overline{P}$, no edges of $G_{\overline{P}}$ can cross one another, which constrains cycles to have a left edge (b, a) corresponding to each right edge (a, b). We remind that by assumption $s_1 < \cdots < s_A$ and $t_1 < \cdots < t_B$ (the notation was introduced in Lemma B.II.7).

Lemma B.II.11. Let $\overline{P} \in \mathbb{R}_+^{A \times B}$ be the OT matrix between $\sum_{a=1}^{A} \frac{\#I_a}{n} \delta_{s_a}$ and $\sum_{b=1}^{B} \frac{\#J_b}{n} \delta_{t_b}$. If $\mathcal{C} := (a_1, b_1, \cdots, a_p, b_p, a_{p+1})$ is a cycle in $G_{\overline{P}}$ (i.e. $(a_k)_{k=1}^{p+1} \in [\![1, A]\!]^{p+1}$, $(b_k)_{k=1}^{p} \in [\![1, B]\!]^p$, $a_{p+1} = a_1$ and $\forall k \in [\![1, p]\!]$, $\overline{P}_{a_k, b_k} > 0$, $\overline{P}_{a_{k+1}, b_k} > 0$) such that the families of edges $((a_k, b_k))_{k=1}^{p}$ and $((b_k, a_{k+1}))_{k=1}^{p}$ are injective, then $((a_k, b_k))_{k=1}^{p} \simeq ((a_{k+1}, b_k))_{k=1}^{p}$.

Proof. First, by optimality of $\overline{P}$, by [San15] Lemma 2.8, $\overline{P}$ is monotone in the sense that:

$$\forall (a, b), (a', b') \in [\![1, A]\!] \times [\![1, B]\!], \text{ such that } \overline{P}_{a,b} > 0, \overline{P}_{a',b'} > 0, \; a < a' \implies b \leq b'.$$

Note that the contrapositive yields the symmetrical property that if $b < b'$ then $a \leq a'$. Furthermore, we remind that since each (a_k, b_k) and (b_k, a_{k+1}) are edges of the graph $G_{\overline{P}}$, we have $\overline{P}_{a_k, b_k} > 0$ and $\overline{P}_{a_{k+1}, b_k} > 0$. We can understand the monotonicity property as the fact that the edges of the cycle cannot cross one another.

By injectivity of the edge families, to show the equipotence result, it suffices to show that $\forall k \in [\![1, p]\!]$, $\exists k' \in [\![1, p]\!]$ such that $(a_k, b_k) = (a_{k'+1}, b_{k'})$. Since the vertices a_k and b_k are part of the cycle $a_1 \to b_1 \to a_2 \to \cdots \to a_{p+1} = a_1$, there exists a sub-cycle $b_k \to a'_1 \to b'_1 \to \cdots \to b'_q \to a_k$, which is to say that there exists, for some $q \geq 0$, $(b_k, a'_1) \in \mathcal{C}$, $\forall k' \in [\![1, q]\!]$, $(a'_{k'}, b'_{k'}) \in \mathcal{C}$, $(b'_{k'}, a'_{k'+1}) \in \mathcal{C}$ (writing $a'_{q+1} := a_k$), and we now take $q \geq 0$ minimal. We will show that $q = 0$ by contradiction: assume $q \geq 1$, which implies that $a'_1 \neq a_k$ by minimality. Assume that $a'_1 < a_k$ (the case $a'_1 > a_k$ is analogous). By monotonicity of $\overline{P}$, we deduce $b'_1 \leq b_k$, and even $b'_1 < b_k$ since $b'_1 = b_k$ would violate the minimality of q. By monotonicity, we deduce that $a'_2 \leq a'_1$ and again, even $a'_2 < a'_1$ by minimality of a. Continuing this process we find that $a'_{q+1} < a_k$ contradicting $a'_{q+1} = a_k$. We conclude that the edge (b_k, a_k) belongs to the cycle, which is to say that there exists $k' \in [\![1, p]\!]$ such that $(a_k, b_k) = (a_{k'+1}, b_{k'})$, finishing the proof. □

In Fig. B.II.13 we illustrate the result of Lemma B.II.11 in the use case of the proof of Theorem B.II.3, regrouping the continued example from Figs. B.II.9 to B.II.12.

B.II.5.5 A Constrained Version of the Birkhoff von Neumann Theorem

We are now ready to prove a constrained version of the Birkhoff von Neumann Theorem [Bir46]. We remind that $\mathcal{P}_\theta(X, Y)$ is defined in Eq. (B.II.42), $\mathbb{U}$ in Eq. (B.II.41) and $\mathcal{P}^{\mathfrak{S}_\theta(X,Y)} =$

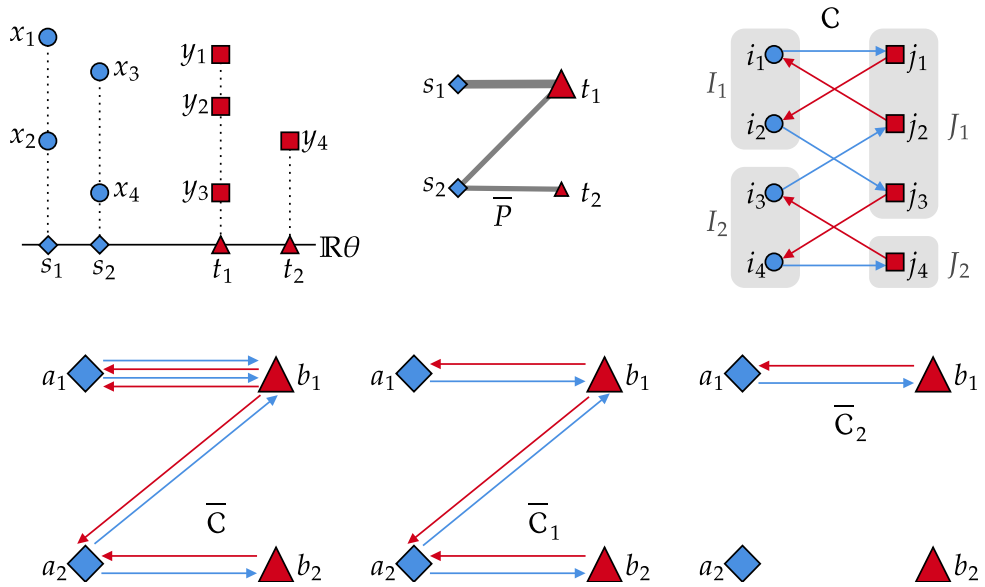


Figure B.II.13: We take two discrete uniform measures $\mu := \frac{1}{n}\sum_i \delta_{x_i}$ and $\nu := \frac{1}{n}\sum_j \delta_{y_j}$ such that $s_1 := P_\theta x_1 = P_\theta x_2 < s_2 := P_\theta x_3 = P_\theta x_4$ and $t_1 := P_\theta y_1 = P_\theta y_2 = P_\theta y_3 < t_2 := P_\theta y_4$. We represent the OT matrix $\overline{P}$ between the measures $P_\theta\#\mu$ and $P_\theta\#\nu$ and consider the bipartite graph G_P associated to a coupling $P \in \mathbb{U}\cap\mathcal{P}_\theta(X,Y)$ (presented in Fig. B.II.11). In this case the graph G_P (top-right) contains the cycle $\mathcal{C} := (i_1, j_1, i_2, j_3, i_4, j_4, i_3, j_2, i_1)$. We consider for each k the "right" edge (a_k, b_k) in $G_{\overline{P}}$ such that $i_k \in I_{a_k}$ and $j_k \in J_{b_k}$, and the "left" edge (b_k, a_{k+1}) such that $j_k \in J_{b_k}$ and $i_{k+1} \in I_{a_{k+1}}$. This defines the cycle $\overline{\mathcal{C}}$ in $G_{\overline{P}}$, that we decompose into cycles with distinct edges $(\overline{\mathcal{C}}_\ell)$ using Lemma B.II.10. Lemma B.II.11 then applies to each $\overline{\mathcal{C}}_\ell$ and we observe indeed that in $\overline{\mathcal{C}}_\ell$, each "left" edge (b,a) has a corresponding "right" edge (a,b) in the cycle.

$\{P^{\sigma,\tau},\ (\sigma,\tau) \in \mathfrak{S}_\theta(X,Y)\}$, with $P^{\sigma,\tau}$ the permutation matrix introduced in Eq. (B.II.43) and $\mathfrak{S}_\theta(X,Y)$ defined in Eq. (B.II.36). Finally, the notion of extreme points is defined in Definition B.II.5.

Theorem B.II.3. Let $(X,Y) \in \mathbb{R}^{n\times d}$ and $\theta \in \mathbb{S}^{d-1}$ verifying Assumption B.II.1. Then

$$\mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) = \mathcal{P}^{\mathfrak{S}_\theta(X,Y)}.$$

Proof. — *Step 1*: $\mathcal{P}^{\mathfrak{S}_\theta(X,Y)} \subset \mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y))$

First, for $(\sigma,\tau) \in \mathfrak{S}_\theta(X,Y)$, we have $P^{\sigma,\tau} \in \mathcal{P}_\theta(X,Y)$ by Lemma B.II.8, which shows that $P^{\sigma,\tau} \in \mathbb{U}\cap\mathcal{P}_\theta(X,Y)$. Now if $P^{\sigma,\tau} = \frac{1}{2}Q + \frac{1}{2}R$ for some $(Q,R) \in \mathbb{U}\cap\mathcal{P}_\theta(X,Y)$, then for any $(i,j) \in [\![1,n]\!]^2$, we have $P^{\sigma,\tau}_{i,j} \in \{0, \frac{1}{n}\}$, thus $\frac{1}{2}Q_{i,j}+\frac{1}{2}R_{i,j} = P^{\sigma,\tau}_{i,j}$ implies that $Q_{i,j} = R_{i,j}$ since $Q_{i,j}$ and $R_{i,j}$ are both in $[0, \frac{1}{n}]$ (since they belong to $\mathbb{U}$). This shows that $P^{\sigma,\tau} \in \mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y))$.

— *Step 2*: Writing $P \in (\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) \setminus \mathcal{P}^{\mathfrak{S}_\theta(X,Y)}$ as $P = (Q+R)/2$ with $Q,R \in [0,\frac{1}{n}]^{n\times n}$

To show $\mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) \subset \mathcal{P}^{\mathfrak{S}_\theta(X,Y)}$, we will show that

$$(\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) \setminus \mathcal{P}^{\mathfrak{S}_\theta(X,Y)} \subset (\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) \setminus \mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y)).$$

Note that when $n = 1$ the entire Theorem is trivial, and in the following we assume $n \geq 2$. We take $P \in (\mathbb{U}\cap\mathcal{P}_\theta(X,Y)) \setminus \mathcal{P}^{\mathfrak{S}_\theta(X,Y)}$ and apply Lemma B.II.9, allowing us to introduce $(i_1, j_1, \cdots, i_p, j_p, i_{p+1}) \in [\![1,n]\!]^{2p+1}$ such that $i_{p+1} = i_1$, the families $(i_k)_{k=1}^p$ and $(j_k)_{k=1}^p$ are injective and $\forall k \in [\![1,p]\!]$, $P_{i_k,j_k} \in (0,\frac{1}{n})$, $P_{i_{k+1},j_k} \in (0,\frac{1}{n})$. We consider the set of "right edges" $E_R := ((i_k,j_k))_{k=1}^p$ and "left edges": $E_L := ((i_{k+1},j_k))_{k=1}^p$. By injectivity, we have $E_R \cap E_L = \varnothing$ and that E_R and E_L are themselves injective. Note that our cycle construction is illustrated on an example in Fig. B.II.13. We take the smallest margin $\varepsilon > 0$ that P has to be in $\mathbb{U}$ (within the cycle):

$$\varepsilon := \min_{k\in[\![1,p]\!]} \{P_{i_k,j_k}, 1-P_{i_k,j_k}, P_{i_{k+1},j_k}, 1-P_{i_{k+1},j_k}\} \in (0,\tfrac{1}{n}),$$

and introduce the matrices $Q, R \in \mathbb{R}^{n\times n}$ defined as, for $(i,j) \in [\![1,n]\!]^2$:

$$Q_{i,j} := \begin{cases} P_{i,j} & \text{if} \quad (i,j) \notin E_R \cup E_L \\ P_{i,j} + \varepsilon & \text{if} \quad (i,j) \in E_R \\ P_{i,j} - \varepsilon & \text{if} \quad (i,j) \in E_L \end{cases}, \quad R_{i,j} := \begin{cases} P_{i,j} & \text{if} \quad (i,j) \notin E_R \cup E_L \\ P_{i,j} - \varepsilon & \text{if} \quad (i,j) \in E_R \\ P_{i,j} + \varepsilon & \text{if} \quad (i,j) \in E_L \end{cases}.$$

For visualisation purposes, in the example of Fig. B.II.13, we represent "left" edges of the cycle in blue (on these edges, we add $+\varepsilon$ in Q and $-\varepsilon$ in R) and "right" edges in red (on which we do the opposite). By definition of ε, we have $Q, R \in [0, \frac{1}{n}]^{n\times n}$. By construction, we also have $P = \frac{1}{2}(Q+R)$.

— *Step 3*: Showing that $Q, R \in \mathbb{U}$

Fix $j \in [\![1,n]\!]$, we show that $\sum_i Q_{i,j} = \frac{1}{n}$. Since E_L and E_R are injective and disjoint, we compute

$$\sum_{i=1}^{n} Q_{i,j} = \sum_{i:(i,j)\notin E_R\cup E_L} P_{i,j} + \sum_{i:(i,j)\in E_R} (P_{i,j}+\varepsilon) + \sum_{i:(i,j)\in E_L} (P_{i,j}-\varepsilon) = \frac{1}{n} + \varepsilon(\#I_R(j) - \#I_L(j)),$$

where $I_R(j) := \{i \in [\![1,n]\!] : (i,j) \in E_R\}$ and $I_L(j) := \{i \in [\![1,n]\!] : (i,j) \in E_L\}$. Since E_R and E_L are injective, we deduce that I_R and I_L are also injective. Take $i \in I_R(j)$ and write $(i,j) = (i_k, j_k) \in E_L$ for some $k \in [\![1,p]\!]$. We notice that $(i_k, j_{k-1}) \in E_R$ where if $k = 1$ we write $j_{k-1} := j_p$. We conclude that $\#I_R(j) = \#I_L(j)$ and thus that $\sum_i Q_{i,j} = \frac{1}{n}$. The same reasoning shows that $\sum_j Q_{i,j} = \frac{1}{n}$ for all $i \in [\![1,n]\!]$, and we conclude that $Q \in \mathbb{U}$. The same computations show that $R \in \mathbb{U}$ as well.

— *Step 4*: Showing that $Q, R \in \mathcal{P}_\theta(X,Y)$

We now show that $Q, R \in \mathcal{P}_\theta(X,Y)$ using the definition (Eq. (B.II.42)). Take $(a,b) \in [\![1,A]\!] \times [\![1,B]\!]$. We have:

$$\begin{aligned} \sum_{(i,j)\in I_a\times J_b} Q_{i,j} &= \sum_{(i,j)\in (I_a\times J_b)\cap E_R^c\cap E_L^c} P_{i,j} + \sum_{(i,j)\in (I_a\times J_b)\cap E_R} (P_{i,j}+\varepsilon) + \sum_{(i,j)\in (I_a\times J_b)\cap E_L} (P_{i,j}-\varepsilon) \\ &= \frac{\#I_a \cap J_b}{n} + \varepsilon\left(\#((I_a\times J_b)\cap E_R) - \#((I_a\times J_b)\cap E_L)\right). \end{aligned} \tag{B.II.48}$$

Let $\overline{P} \in \mathbb{R}_+^{A\times B}$ be the OT matrix between $\sum_{a=1}^A \frac{\#I_a}{n}\delta_{s_a}$ and $\sum_{b=1}^B \frac{\#J_b}{n}\delta_{t_b}$. Consider the family $\overline{\mathcal{C}} := (a_1, b_1, \cdots, a_p, b_p, a_{p+1})$ defined by the condition $\forall k \in [\![1,p]\!]$, $i_k \in I_{a_k}$, $j_k \in J_{b_k}$ and $a_{p+1} := a_1$. Since $\mathcal{C} := (i_1, i_2, \cdots, i_p, j_p, i_{p+1})$ is a cycle in G_P, it follows that $\overline{\mathcal{C}}$ is a cycle in $G_{\overline{P}}$ since the condition $P \in \mathcal{P}_\theta(X,Y)$ implies:

$$\forall (a,b) \in [\![1,A]\!] \times [\![1,B]\!], \quad \sum_{(i,j)\in I_a\times J_b} P_{i,j} = \overline{P}_{a,b},$$

thus if $P_{i,j} > 0$ for some $(i,j) \in I_a \times J_b$ then $P_{a,b} > 0$. See also Fig. B.II.13 for an example. We now apply Lemma B.II.10 to show that $\overline{\mathcal{C}}$ is the "concatenation" of $L \geq 1$ cycles $\overline{\mathcal{C}}_\ell$ of $G_{\overline{P}}$ of the form:

$$\overline{\mathcal{C}}_\ell := (a_1^\ell, b_1^\ell, \cdots, a_{p_\ell}^\ell, b_{p_\ell}^\ell, a_{p_\ell+1}^\ell),$$

where "concatenation" means that Eq. (B.II.47) holds with the same notation, and where each $\overline{\mathcal{C}}_\ell$ is such that the edge families $((a_k^\ell, b_k^\ell))_{k=1}^{p_\ell}$ and $((b_k^\ell, a_{k+1}^\ell))_{k=1}^{p_\ell}$ are injective. For each $\ell \in [\![1,L]\!]$, we apply Lemma B.II.11, which shows in particular that for any $(a,b) \in [\![1,A]\!] \times [\![1,B]\!]$:

$$\#\left\{k \in [\![1,p_\ell]\!] : (a_k^\ell, b_k^\ell) = (a,b)\right\} = \#\left\{k \in [\![1,p_\ell]\!] : (a_{k+1}^\ell, b_k^\ell) = (a,b)\right\}. \tag{B.II.49}$$

We understand Eq. (B.II.49) as the fact that for any left edge from group a to group b in $\mathcal{C}_\ell$, there corresponds exactly as many right edges from group b to group a. This will allow us to show that the terms in $+\varepsilon$ and $-\varepsilon$ are in the same number. We now re-write the sets from the

condition on Q (Eq. (B.II.48)) for a fixed $(a,b) \in [\![1,A]\!] \times [\![1,B]\!]$:

$$\begin{aligned}
\#\left((I_a \times J_b) \cap E_R\right) &= \{(i_k, j_k),\ k \in [\![1,p]\!],\ (i_k,j_k) \in I_a \times J_b\} \\
&= \#\left((a_k, b_k),\ k \in [\![1,p]\!],\ (a_k,b_k) = (a,b)\right) \\
&= \sum_{\ell=1}^{L} \#\left((a_k^\ell, b_k^\ell),\ k \in [\![1,p_\ell]\!],\ (a_k^\ell, b_k^\ell) = (a,b)\right) \\
&= \sum_{\ell=1}^{L} \#\left((b_k^\ell, a_{k+1}^\ell),\ k \in [\![1,p_\ell]\!],\ (a_{k+1}^\ell, b_k^\ell) = (a,b)\right) \\
&= \#\left((I_a \times J_b) \cap E_L\right),
\end{aligned}$$

where the first equality uses the definition of E_R, the second inequality comes from associating to each pair (i_k, j_k) its group pair (a_k, b_k) and counting the group pairs *with repetition*, the third equality the concatenation property of the cycles $\mathcal{C}_\ell$ (Eq. (B.II.47)), the fourth equality from Eq. (B.II.49) and the last inequality from the definition of E_L (doing the same computations as for E_R in reverse order).

Combining with Eq. (B.II.48) shows that $\sum_{(i,j)\in I_a \times J_b} Q_{i,j} = \frac{\#(I_a \cap J_b)}{n}$ and thus that $Q \in \mathcal{P}_\theta(X,Y)$. Likewise we show $R \in \mathcal{P}_\theta(X,Y)$ and thus we have found $Q, R \in \mathbb{U} \cap \mathcal{P}_\theta(X,Y)$ such that $P = \frac{1}{2}(Q+R)$, and we conclude that P does not belong to $\mathrm{Extr}(\mathbb{U} \cap \mathcal{P}_\theta(X,Y))$, finishing the proof. □

From Theorem B.II.3 we deduce the following theorem, which is a Monge formulation of the constrained Kantorovich problem in CW_θ:

Theorem B.II.4. Let $(X,Y) \in \mathbb{R}^{n\times d}$ and $\theta \in \mathbb{S}^{d-1}$, we have:

$$\mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{(\sigma,\tau)\in \mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2, \tag{B.II.50}$$

Proof. Beginning under Assumption B.II.1, we combine Theorem B.II.3 with the expression of CW_θ^2 from Proposition B.II.9:

$$\mathrm{CW}_\theta^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \min_{P\in \mathbb{U}\cap\mathcal{P}_\theta(X,Y)} \sum_{i,j} \|x_i - y_j\|_2^2 P_{i,j} = \min_{P\in \mathrm{Extr}(\mathbb{U}\cap\mathcal{P}_\theta(X,Y))} \sum_{i,j} \|x_i - y_j\|_2^2 P_{i,j},$$

since the solution of a linear program over a non-empty convex compact set is attained at an extreme point ([BT97] Theorem 2.7), and we conclude that the expression in Eq. (B.II.50) holds thanks to Theorem B.II.3. For the general case without Assumption B.II.1, we use Lemma B.II.6. □

B.II.6 Min-Pivot Sliced

B.II.6.1 Min-Pivot Sliced Discrepancy: Definition

A specificity of the Pivot Sliced Wasserstein discrepancy is the dependence on the axis $\theta \in \mathbb{S}^{d-1}$, which can overly constrain the choice of transport plans. In this section, we study the Min-Pivot Sliced Discrepancy which minimises PS_θ over $\theta \in \mathbb{S}^{d-1}$. This object was first introduced in [Mah+23] on the set of discrete uniform measures with n points.

$$\min \mathrm{PS}^2(\mu_1,\mu_2) := \min_{\theta\in\mathbb{S}^{d-1}} \mathrm{PS}_\theta^2(\mu_1,\mu_2) = \min_{\substack{\theta\in\mathbb{S}^{d-1} \\ \omega\in\Omega_\theta(\mu_1,\mu_2)}} \int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\omega(x_1,x_2), \tag{B.II.51}$$

where we used the notation $\Omega_\theta(\mu_1,\mu_2)$ defined in Eq. (B.II.32), and Theorem B.II.2. We show below that the infimum is attained:

Proposition B.II.10. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. Then the minimum in Eq. (B.II.51) is attained.

Proof. Take a sequence $(\theta_n)_{n\in\mathbb{N}} \in \mathbb{S}^{d-1}$ such that $\mathrm{PS}_{\theta_n}(\mu_1,\mu_2) \xrightarrow[n\longrightarrow+\infty]{} \min \mathrm{PS}(\mu_1,\mu_2)$. By compactness of $\mathbb{S}^{d-1}$, we can extract a converging subsequence of (θ_n): up to extraction we can assume that $\theta_n \xrightarrow[n\longrightarrow+\infty]{} \theta \in \mathbb{S}^{d-1}$. Denoting $\mu_{\theta_n} := \mu_{\theta_n}[\mu_1,\mu_2]$, by Proposition B.II.1, for each $n \in \mathbb{N}$ we can choose $\rho_n \in \Gamma(\mu_{\theta_n},\mu_1,\mu_2)$ optimal for $\mathrm{PS}_{\theta_n}(\mu_1,\mu_2)$. By Proposition B.II.6, we have $\mu_{\theta_n} \xrightarrow[n\longrightarrow+\infty]{w} \mu_\theta$. Using Lemma B.II.1 item 1) we obtain that the set of $\Gamma(\{\mu_{\theta_n}\},\mu_1,\mu_2)$ is tight in $\mathcal{P}_2(\mathbb{R}^{3d})$, and since $(\rho_n) \in \Gamma(\{\mu_{\theta_n}\},\mu_1,\mu_2)^{\mathbb{N}}$, there exists an extraction α such that $\rho_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \rho \in \mathcal{P}_2(\mathbb{R}^{3d})$. Applying Lemma B.II.1 item 2) shows that $\rho \in \Gamma(\mu_\theta,\mu_1,\mu_2)$.

The cost function $J := \rho \in \mathcal{P}_2(\mathbb{R}^{3d}) \longmapsto \int_{\mathbb{R}^{3d}} \|x_1-x_2\|_2^2 \mathrm{d}\rho_{1,2}(y,x_1,x_2)$ is lower semi-continuous by [San15, Lemma 1.6], which provides the following inequality:

$$\mathrm{PS}^2_\theta(\mu_1,\mu_2) \leq J(\rho) \leq \liminf_{n\longrightarrow+\infty} J(\rho_{\alpha(n)}) = \lim_{n\longrightarrow+\infty} S^2_{\theta_{\alpha(n)}}(\mu_1,\mu_2) = \min \mathrm{PS}^2(\mu_1,\mu_2),$$

where the first inequality holds by the property $\rho \in \Gamma(\mu_\theta,\mu_1,\mu_2)$, the second inequality comes from the lower semi-continuity of J and the first equality comes from the fact that $\forall n \in \mathbb{N}$, $J(\rho_n) = \mathrm{PS}^2_{\theta_n}(\mu_1,\mu_2)$. We conclude that $\min \mathrm{PS}(\mu_1,\mu_2) = \mathrm{PS}_\theta(\mu_1,\mu_2)$ and thus the infimum is attained. □

From Proposition B.II.10, we conclude that the properties of PS_θ stated in Proposition B.II.5 are inherited by $\min \mathrm{PS}$. In Example B.II.5, we show an example which numerically contradicts the triangle inequality.

Example B.II.5 ($\min \mathrm{PS}$ does not verify the triangle inequality)**.** We consider a setting with three measures $\mu_1,\mu_2,\mu_3 \in \mathcal{P}(\mathbb{R}^2)$ with 10 points each, obtained with five rotations of the example from Example B.II.3, which we represent in Fig. B.II.14. Extensive numerical approximation with $L := 10^5$ directions yields the following violation of the triangle inequality:

$$\min \mathrm{PS}(\mu_1,\mu_3) + \min \mathrm{PS}(\mu_3,\mu_2) - \min \mathrm{PS}(\mu_1,\mu_2) \approx -0.612.$$

While the expression are not tractable in closed form, this numerical experiment strongly suggests that the triangle inequality does not hold for $\min \mathrm{PS}$.

B.II.6.2 Equality with the Wasserstein Distance for Certain Discrete Measures

In [Mah+23, Proposition 3.2], the authors show (proof in [Mah+23, Section 11.1]) that the Min-Sliced Discrepancy $\min \mathrm{PS}$ equals W_2 on the set $\mathcal{P}^n(\mathbb{R}^d)$ of uniform discrete measures with n points under a condition on n and d. Their proof relies on an application of [Cov67] which requires the points to be in general position (see Definition B.II.7), however the condition is not stated in [Mah+23]. For the sake of clarity, we restate the result and provide a detailed proof. First, we remind the notion of points of $\mathbb{R}^d$ in general position in Definition B.II.7.

Definition B.II.7. Let $x_1,\cdots,x_n \in \mathbb{R}^d$. We say that the points are in general position if for all $k \in [\![1,d]\!]$, there is no subset $I \subset [\![1,n]\!]$ with $k+2$ elements such that $\{x_i\}_{i\in I}$ is contained in a k-dimensional affine subspace of $\mathbb{R}^d$.

Proposition B.II.11. Let $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ such that the union of

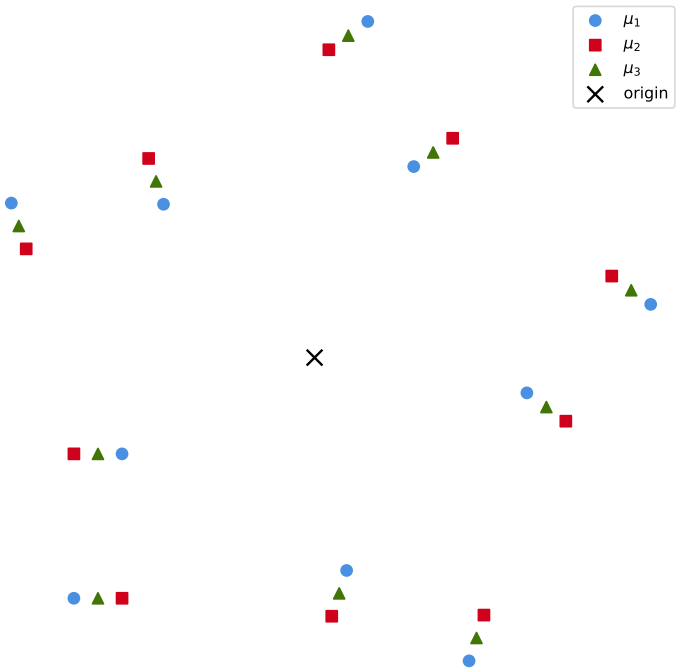


Figure B.II.14: Counter-example from Example B.II.5 to the triangle inequality for min PS.

supports $(x_i) \cup (y_j)$ is in general position. If $2n \leq d+1$, then:

$$\left\{(\sigma,\tau) \in \mathfrak{S}_n^2 : \exists \theta \in \mathbb{S}^{d-1} : \forall i \in [\![1, n-1]\!],\ P_\theta x_{\sigma(i)} < P_\theta x_{\sigma(i+1)},\ P_\theta y_{\tau(i)} < P_\theta y_{\tau(i+1)}\right\} = \mathfrak{S}_n^2. \tag{B.II.52}$$

As a result, $\min \mathrm{PS}(\mu,\nu) = \mathrm{W}_2(\mu,\nu)$.

Proof. By the Theorem in [Cov67, Section 2] and [Cov67, Equation (12) in Section 3], since $(x_1, \cdots, x_n, y_1, \cdots, y_n)$ are in general position and $d \geq 2n-1$, we have

$$\left\{\alpha \in \mathfrak{S}_{2n} : \exists \theta \in \mathbb{S}^{d-1} : \forall k \in [\![1, 2n-1]\!],\ P_\theta z_{\alpha(i)} < P_\theta z_{\alpha(i+1)}\right\} = \mathfrak{S}_{2n}, \tag{B.II.53}$$

where $(z_1, \cdots, z_{2n}) := (x_1, \cdots, x_n, y_1, \cdots, y_n)$. Take now $(\sigma,\tau) \in \mathfrak{S}_n^2$ and define $\alpha \in \mathfrak{S}_{2n}$ by:

$$\forall i \in [\![1,n]\!],\ \alpha(i) := \sigma(i), \quad \forall j \in [\![1,n]\!],\ \alpha(n+j) := \tau(j).$$

By Eq. (B.II.53), there exists $\theta \in \mathbb{S}^{d-1}$ such that $\forall k \in [\![1, 2n-1]\!],\ P_\theta z_{\alpha(i)} < P_\theta z_{\alpha(i+1)}$, showing Eq. (B.II.52).

Now by definition of $\mathfrak{S}_\theta(X,Y)$, the right hand-side term of Eq. (B.II.52) is a subset of $\cup_{\theta\in\mathbb{S}^{d-1}} \mathfrak{S}_\theta(X,Y)$, which shows that $\cup_{\theta\in\mathbb{S}^{d-1}} \mathfrak{S}_\theta(X,Y) = \mathfrak{S}_n^2$. Using Theorem B.II.2 and Theorem B.II.4 we conclude:

$$\min \mathrm{PS}(\mu,\nu) = \min_{(\sigma,\tau)\in \cup_{\theta\in\mathbb{S}^{d-1}} \mathfrak{S}_\theta(X,Y)} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2 = \min_{(\sigma,\tau)\in\mathfrak{S}_n^2} \frac{1}{n}\sum_{i=1}^n \|x_{\sigma(i)} - y_{\tau(i)}\|_2^2 = \mathrm{W}_2^2(\mu,\nu),$$

where the last equality comes from the Monge formulation of W_2 in the case of uniform measures with the same number of points (see [PC19b, Proposition 2.1] for instance). □

B.II.7 Expected Sliced Wasserstein

In [Liu+24], Liu et al. present a variant of the Sliced Wasserstein distance, consisting in taking the transport cost of a coupling that is an average of lifted sliced couplings. In this section, we will explain how to define these notions for general measures of $\mathcal{P}_2(\mathbb{R}^d)$ instead of discrete ones.

B.II.7.1 Lifting Sliced Plans

To lift a 1D transport plan onto $\mathbb{R}^d$, we will require the notion of disintegration of measures with respect to a map reminded in Section B.II.2.4. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$. Consider the disintegration of μ_1 with respect to $P_\theta := x \longmapsto x \cdot \theta$ as in Definition B.II.10: $\mu_1(\mathrm{d}x) = (P_\theta \# \mu_1)(P_\theta \mathrm{d}x)\mu_1^{P_\theta x}(\mathrm{d}x)$. The kernel $\mu_1^{P\theta x}$ is a measure on $\mathbb{R}^d$ supported on the slice $\{x' \in \mathbb{R}^d \mid x' \cdot \theta = x \cdot \theta\} = x + \theta^\perp$. Denoting similarly the disintegration of μ_2 by $\mu_2(\mathrm{d}y) = (P_\theta \# \mu_2)(P_\theta \mathrm{d}y)\mu_2^{P_\theta y}(\mathrm{d}y)$, we first notice that the disintegration of $\mu_1 \otimes \mu_2$ with respect to (P_θ, P_θ) writes simply as a product:

$$\mu_1 \otimes \mu_2(\mathrm{d}x, \mathrm{d}y) = (P_\theta \# \mu_1)(P_\theta \mathrm{d}x)(P_\theta \# \mu_2)(P_\theta \mathrm{d}y)\mu_1^{P_\theta x}(\mathrm{d}x)\mu_2^{P_\theta y}(\mathrm{d}y), \tag{B.II.54}$$

noticing that $(P_\theta, P_\theta)\#(\mu_1 \otimes \mu_2) = (P_\theta \# \mu_1) \otimes (P_\theta \# \mu_2)$.

Take now the 1D OT plan $\pi_\theta := \pi_\theta[\mu_1, \mu_2] \in \Pi^*(P_\theta \# \mu_1, P_\theta \# \mu_2)$, the idea behind the lift is to replace the independent coupling $(P_\theta, P_\theta)\#(\mu_1 \otimes \mu_2) = (P_\theta \# \mu_1) \otimes (P_\theta \# \mu_2)$ in Eq. (B.II.54) by π_θ: we define the lifted plan through its disintegration as:

$$\gamma_\theta(\mathrm{d}x, \mathrm{d}y) := \pi_\theta(P_\theta \mathrm{d}x, P_\theta \mathrm{d}y)\mu_1^{P_\theta x}(\mathrm{d}x)\mu_2^{P_\theta y}(\mathrm{d}y). \tag{B.II.55}$$

More formally, we can define γ_θ using test functions $\phi \in \mathcal{C}_b^0(\mathbb{R}^d \times \mathbb{R}^d)$:

$$\int_{\mathbb{R}^{2d}} \phi(x,y)\mathrm{d}\gamma_\theta(x,y) = \int_{\mathbb{R}^2} \left(\int_{P_\theta^{-1}(s) \times P_\theta^{-1}(t)} \phi(x,y)\mathrm{d}\mu_1^s(x)\mathrm{d}\mu_2^t(y) \right) \mathrm{d}\pi_\theta(s,t). \tag{B.II.56}$$

We illustrate the definition of the lifted plan on a simple example in Fig. B.II.15.

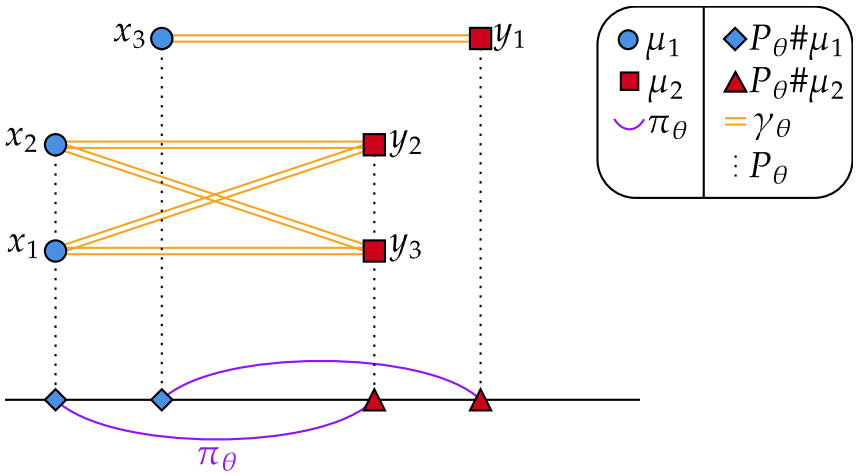


Figure B.II.15: Example of the lifted plan γ_θ between two measures μ_1 and μ_2. In this case, we notice that $P_\theta x_1 = P_\theta x_2$ and $P_\theta y_2 = P_\theta y_3$. As a result, the optimal plan π_θ between $P_\theta \# \mu_1$ and $P_\theta \# \mu_2$ does not allow us to deduce an assignment between (x_1, x_2) and (y_2, y_3). The lifted coupling γ_θ chooses the independent coupling: x_1 is assigned uniformly to (y_2, y_3) and likewise for x_2. As for x_3 and y_1, the coupling π_θ assigns $P_\theta x_3$ to $P_\theta y_1$ which imposes that γ_θ send x_3 to y_1.

In Proposition B.II.12 we show that the lifted plan is a valid coupling between μ_1 and μ_2. We also provide an explicit expression for discrete measures μ_1, μ_2, which coincides with the expression in [Liu+24, Equation 9], which serves as their definition of lifted plans.

Proposition B.II.12. Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, $\theta \in \mathbb{S}^{d-1}$ and $\gamma_\theta := \gamma_\theta[\mu_1, \mu_2]$ the lifted plan defined in Eq. (B.II.56). Then:

1. $\gamma_\theta \in \Pi(\mu_1, \mu_2)$.
2. If $\mu_1 = \sum_{i=1}^n a_i \delta_{x_i}$ and $\mu_2 = \sum_{j=1}^m b_j \delta_{y_j}$, let $\pi_\theta \in \Pi^*(P_\theta \# \mu_1, P_\theta \# \mu_2)$. For $(i,j) \in$

$[\![1,n]\!] \times [\![1,m]\!]$, we define $Q_{i,j} := \pi_\theta(\{(P_\theta x_i, P_\theta y_j)\})$, which allows us to see π_θ as a matrix of size $n \times m$.

$$\gamma_\theta = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{a_i b_j}{A_i B_j} Q_{i,j}\delta_{(x_i,y_j)}, \quad A_i := \sum_{i':x_{i'}\cdot\theta = x_i\cdot\theta} a_{i'}, \; B_j := \sum_{j':y_{j'}\cdot\theta = y_j\cdot\theta} b_{j'}. \tag{B.II.57}$$

Proof. For 1. we verify the property using a test function $\phi \in \mathcal{C}_b^0(\mathbb{R}^d)$ with Eq. (B.II.56):

$$\begin{aligned}
\int_{\mathbb{R}^{2d}} \phi(x)\mathrm{d}\gamma_\theta(x,y) &= \int_{\mathbb{R}^2} \left(\int_{P_\theta^{-1}(s)\times P_\theta^{-1}(t)} \phi(x)\mathrm{d}\mu_1^s(x)\mathrm{d}\mu_2^t(y) \right) \mathrm{d}\pi_\theta(s,t) \\
&= \int_{\mathbb{R}^2} \left(\int_{P_\theta^{-1}(s)} \phi(x)\mathrm{d}\mu_1^s(x) \right) \underbrace{\left(\int_{P_\theta^{-1}(t)} \mathrm{d}\mu_2^t(y) \right)}_{=1} \mathrm{d}\pi_\theta(s,t) \\
&= \int_{\mathbb{R}} \left(\int_{P_\theta^{-1}(s)} \phi(x)\mathrm{d}\mu_1^s(x) \right) \mathrm{d}(P_\theta\#\mu_1)(s) \\
&= \int_{\mathbb{R}^d} \phi(x)\mathrm{d}\mu_1(x),
\end{aligned}$$

where we use the fact that the first marginal of π_θ is $P_\theta\#\mu_1$, and then used disintegration of μ_1 with respect to P_θ. The same method shows that the second marginal of γ_θ is μ_2, concluding $\gamma_\theta \in \Pi(\mu_1,\mu_2)$.

For 2. we begin by writing explicitly the disintegration of μ_1 with respect to P_θ. For $\mu_1 = \sum_{i=1}^n a_i\delta_{x_i}$, we have $P_\theta\#\mu = \sum_i a_i\delta_{P_\theta x_i}$, and for $s = P_\theta x_i \in \mathrm{supp}(P_\theta\#\mu)$, we have $\mu_1^s = A_i^{-1}\sum_{i':x_{i'}\cdot\theta=s} a_{i'}\delta_{x_{i'}}$. We establish Eq. (B.II.57) by testing on $\phi \in \mathcal{C}_b^0(\mathbb{R}^{2d})$. The support of π_θ is (at most) the family of pairs $((x_i\cdot\theta, y_j\cdot\theta))_{i,j}$. We choose $I \subset [\![1,n]\!]\times[\![1,m]\!]$ such that the support of π_θ is the injective family $((x_i\cdot\theta, y_j\cdot\theta))_{(i,j)\in I}$. We then have:

$$\begin{aligned}
\int_{\mathbb{R}^{2d}} \phi(x,y)\mathrm{d}\gamma_\theta(x,y) &= \int_{\mathbb{R}^2} \left(\int_{P_\theta^{-1}(s)\times P_\theta^{-1}(t)} \phi(x,y)\mathrm{d}\mu_1^s(x)\mathrm{d}\mu_2^t(y) \right) \mathrm{d}\pi_\theta(s,t) \\
&= \sum_{(i,j)\in I} \left(\sum_{\substack{i':P_\theta x_{i'}=P_\theta x_i \\ j':P_\theta y_{j'}=P_\theta y_j}} \phi(x_{i'},y_{j'}) \frac{a_{i'}b_{j'}}{A_i B_j} \right) \pi_\theta(\{(P_\theta x_i, P_\theta y_j)\}) \\
&= \sum_{i=1}^n\sum_{j=1}^m \phi(x_i,y_j)\frac{a_i b_j}{A_i B_j} Q_{i,j},
\end{aligned}$$

where we use the fact that for $(i,j)\in I$ and (i',j') such that $P_\theta x_{i'} = P_\theta x_i$ and $P_\theta y_{j'} = P_\theta y_j$, it holds that $A_i = A_{i'}$, $B_j = B_{j'}$ and $Q_{i,j} = Q_{i',j'}$. □

The discrete expression in Eq. (B.II.57) shows that the definition of listed plans in Eq. (B.II.56) is a generalisation of the plan lift from [Liu+24, Equation 9]. We now study the transport cost associated to the lifted plan γ_θ:

Definition B.II.8. For $\theta \in \mathbb{S}^{d-1}$ and $\mu_1,\mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$. With $\gamma_\theta[\mu_1,\mu_2]$ the lifted plan defined in Eq. (B.II.56), we define the lifted cost as:

$$\mathrm{LS}_\theta^2(\mu_1,\mu_2) := \int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\gamma_\theta[\mu_1,\mu_2](x_1,x_2).$$

We will see that LS_θ defines a discrepancy on $\mathcal{P}_2(\mathbb{R}^d)$ that is almost a distance.

Proposition B.II.13. Fix $\theta \in \mathbb{S}^{d-1}$. The quantity LS_θ is non-negative, symmetric, verifies the triangle inequality, and if $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ verify $\mathrm{LS}_\theta(\mu_1, \mu_2) = 0$ then $\mu_1 = \mu_2$. Furthermore, we have the inequality $\mathrm{LS}_\theta \geq \mathrm{W}_2$.

Proof. Non-negativity and symmetry are immediate. For $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, by Proposition B.II.12, we have $\gamma_\theta[\mu_1, \mu_2] \in \Pi(\mu_1, \mu_2)$, hence $\mathrm{LS}_\theta(\mu_1, \mu_2) \geq \mathrm{W}_2(\mu_1, \mu_2)$. Suppose now that $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ are such that $\mathrm{LS}_\theta(\mu_1, \mu_2) = 0$, then $\mathrm{W}_2(\mu_1, \mu_2) = 0$ and therefore $\mu_1 = \mu_2$.

We now show the triangle inequality: let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$. We consider the sliced 3-plan η_θ defined by:

$$\eta_\theta := \left(F^{[-1]}_{P_\theta \# \mu_1}, F^{[-1]}_{P_\theta \# \mu_2}, F^{[-1]}_{P_\theta \# \mu_3}\right) \# \mathscr{L}_{[0,1]}.$$

For $i < j \in \{1, 2, 3\}$, introduce $\pi^{(i,j)}_\theta$ the unique optimal transport plan between $P_\theta \# \mu_i$ and $P_\theta \# \mu_j$. By [San15, Theorem 2.9], we see that $[\eta_\theta]_{i,j} = \pi^{(i,j)}_\theta$. We now lift the sliced plan η_θ in the same manner as in Eq. (B.II.56), defining a plan $\rho_\theta \in \mathcal{P}_2(\mathbb{R}^{3d})$ by disintegration:

$$\rho_\theta(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) := \eta_\theta(P_\theta \mathrm{d}x_1, P_\theta \mathrm{d}x_2, P_\theta \mathrm{d}x_3) \mu_1^{P_\theta x_1}(\mathrm{d}x_1) \mu_2^{P_\theta x_2}(\mathrm{d}x_2) \mu_3^{P_\theta x_3}(\mathrm{d}x_3).$$

By computing the expectation against test functions, $\forall i < j \in \{1, 2, 3\}$, $[\rho_\theta]_{i,j} = \gamma_\theta[\mu_i, \mu_j]$. We now use the classical gluing method (as in [San15, Lemma 5.5]) to show the triangle inequality, introducing the functions $\phi_i := (x_1, x_2, x_3) \longmapsto x_i$ for $i \in \{1, 2, 3\}$:

$$\begin{aligned}
\mathrm{LS}_\theta(\mu_1, \mu_3) &= \sqrt{\int_{\mathbb{R}^{2d}} \|x_1 - x_3\|_2^2 \mathrm{d}\gamma_\theta[\mu_1, \mu_3](x_1, x_3)} \\
&= \sqrt{\int_{\mathbb{R}^{3d}} \|x_1 - x_3\|_2^2 \mathrm{d}\rho_\theta(x_1, x_2, x_3)} \\
&= \|\phi_1 - \phi_3\|_{L^2(\rho_\theta)} \\
&\leq \|\phi_1 - \phi_2\|_{L^2(\rho_\theta)} + \|\phi_2 - \phi_3\|_{L^2(\rho_\theta)} \\
&= \sqrt{\int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\gamma_\theta[\mu_1, \mu_2](x_1, x_2)} + \sqrt{\int_{\mathbb{R}^{2d}} \|x_2 - x_3\|_2^2 \mathrm{d}\gamma_\theta[\mu_2, \mu_3](x_1, x_2)} \\
&= \mathrm{LS}_\theta(\mu_1, \mu_2) + \mathrm{LS}_\theta(\mu_2, \mu_3).
\end{aligned}$$

□

The discrepancy LS_θ is not a distance on $\mathcal{P}_2(\mathbb{R}^d)$: in Example B.II.6, we introduce a particular case in dimension two where $\mathrm{LS}_\theta(\mu, \mu) > 0$.

Example B.II.6 ($\mathrm{LS}_\theta(\mu, \mu)$ can be non-zero)**.** Take $\theta := (1, 0)$ and $\mu := \frac{1}{2}(\delta_{x_0} + \delta_{x_1})$, $x_0 := (0, 0)$, $x_1 := (0, 1)$. We have $P_\theta \# \mu = \delta_0$ and thus $\gamma_\theta[\mu, \mu] = \mu \otimes \mu$. The lifted cost is then:

$$\mathrm{LS}^2_\theta(\mu, \mu) = \frac{1}{4}\left(\|x_0 - x_0\|_2^2 + \|x_0 - x_1\|_2^2 + \|x_1 - x_0\|_2^2 + \|x_1 - x_1\|_2^2\right) = \frac{1}{2} > 0.$$

For probability measures μ with countable support, for almost-every $\theta \in \mathbb{S}^{d-1}$, there is no ambiguity in the projections, and thus $\mathrm{LS}_\theta(\mu, \mu) = 0$, as shown in Proposition B.II.14. To state the result, we introduce the following notation for the set of probability measures with countable[3] support:

$$\mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d) := \left\{\mu = \sum_{x \in X} a_x \delta_x : X \subset \mathbb{R}^d \text{ countable}, \ (a_x)_{x \in X} \in (0, 1]^d, \ \sum_{x \in X} a_x = 1\right\}. \quad \text{(B.II.58)}$$

Proposition B.II.14. Consider $\mu \in \mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d)$, then for almost-every $\theta \in \mathbb{S}^{d-1}$, we have $\mathrm{LS}_\theta(\mu, \mu) = 0$.

[3] by "countable", we mean a set that is either finite or equipotent to $\mathbb{N}$.

Proof. Denoting $\mathfrak{o}_u$ the uniform measure on the unit sphere $\mathbb{S}^{d-1}$, we have by countable additivity $\mathbb{P}_{\theta\sim\mathfrak{o}_u}(\exists x \neq y \in X^2 : P_\theta x = P_\theta y) \leq \sum_{x\neq y\in X^2} \mathfrak{o}_u((x-y)^\perp) = 0$. We now fix $\Theta \subset \mathbb{S}^{d-1}$ the set $\Theta := \{\theta \in \mathbb{S}^{d-1} : \forall x \neq y \in X^2,\ \theta \notin (x-y)^\perp\}$, we have shown that $\mathfrak{o}_u(\Theta) = 1$. For $\theta \in \Theta$, the family $(P_\theta x)_{x\in X}$ is injective, and the disintegration kernel μ with respect to P_θ at $P_\theta x$ is simply $\delta_{Q_{\theta^\perp} x}$, therefore the lifted plan $\gamma_\theta[\mu,\mu]$ is $\sum_{x\in X} a_x \delta_{(x,x)}$. It follows from the definition that $\mathrm{LS}_\theta(\mu,\mu) = 0$ for any $\theta \in \Theta$, concluding the proof. □

B.II.7.2 Averaging Lifted Plans

Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$. We have constructed a lifted plan $\gamma_\theta \in \Pi(\mu_1,\mu_2)$ (see Eq. (B.II.56) and Proposition B.II.12). We now define the expected lifted plan as the "average" of lifted plans over all directions $\theta \in \mathbb{S}^{d-1}$ through a probability measure $\mathfrak{o} \in \mathcal{P}(\mathbb{S}^{d-1})$. We define $\overline{\gamma}[\mu_1,\mu_2,\mathfrak{o}]$ by duality on test functions $\phi \in \mathcal{C}_b^0(\mathbb{R}^d \times \mathbb{R}^d)$:

$$\int_{\mathbb{R}^{2d}} \phi(x,y)\mathrm{d}\overline{\gamma}[\mu_1,\mu_2,\mathfrak{o}](x,y) := \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^{2d}} \phi(x,y)\mathrm{d}\gamma_\theta[\mu_1,\mu_2](x,y)\mathrm{d}\mathfrak{o}(\theta). \tag{B.II.59}$$

Having defined the expected lifted plan, we can now define the expected sliced discrepancy:

Definition B.II.9. Let $\mu_1,\mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathfrak{o} \in \mathcal{P}(\mathbb{S}^{d-1})$. The expected sliced discrepancy between μ_1 and μ_2 is defined as:

$$\begin{aligned}\mathrm{ES}^2_\mathfrak{o}(\mu_1,\mu_2) &:= \int_{\mathbb{R}^{2d}} \|x-y\|_2^2 \mathrm{d}\overline{\gamma}[\mu_1,\mu_2,\mathfrak{o}](x,y)\\ &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}^{2d}} \|x-y\|_2^2 \mathrm{d}\gamma_\theta[\mu_1,\mu_2](x,y)\mathrm{d}\mathfrak{o}(\theta)\\ &= \int_{\mathbb{S}^{d-1}} \mathrm{LS}^2_\theta(\mu_1,\mu_2)\mathrm{d}\mathfrak{o}(\theta).\end{aligned}$$

where $\overline{\gamma}[\mu_1,\mu_2,\mathfrak{o}]$ is the expected lifted plan between μ_1 and μ_2 for the measure $\mathfrak{o}$ on $\mathbb{S}^{d-1}$, defined in Eq. (B.II.59), and $\gamma_\theta[\mu_1,\mu_2]$ is the lifted plan defined in Eq. (B.II.56).

The properties of LS_θ are passed on to $\mathrm{ES}_\mathfrak{o}$ by integration.

Corollary B.II.2. For any probability measure $\mathfrak{o}$ on $\mathbb{S}^{d-1}$, the quantity $\mathrm{ES}_\mathfrak{o}$ is non-negative, symmetric, verifies the triangle inequality, and if $\mu_1,\mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$ verify $\mathrm{ES}_\mathfrak{o}(\mu_1,\mu_2) = 0$ then $\mu_1 = \mu_2$. Furthermore, we have the inequality $\mathrm{ES}_\mathfrak{o} \geq \mathrm{W}_2$.

Proof. Non-negativity, symmetry and the property $\mathrm{ES}_\mathfrak{o} \geq \mathrm{W}_2$ are immediate by applying the definition and Proposition B.II.13. For the triangle inequality, let $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$ and for $i < j \in \{1,2,3\}$, introduce $f_{i,j} := \theta \longmapsto \mathrm{LS}_\theta(\mu_i,\mu_j)$. By Proposition B.II.13 we have $0 \leq f_{1,3} \leq f_{1,2} + f_{2,3}$. We write:

$$\mathrm{ES}_\mathfrak{o}(\mu_1,\mu_3) = \|f_{1,3}\|_{L^2(\mathfrak{o})} \leq \|f_{1,2}+f_{2,3}\|_{L^2(\mathfrak{o})} \leq \|f_{1,2}\|_{L^2(\mathfrak{o})} + \|f_{2,3}\|_{L^2(\mathfrak{o})} = \mathrm{ES}_\mathfrak{o}(\mu_1,\mu_2) + \mathrm{ES}_\mathfrak{o}(\mu_2,\mu_3).$$

□

The discrepancy $\mathrm{ES}_\mathfrak{o}$ is not a distance on $\mathcal{P}_2(\mathbb{R}^d)$. First, if $d = 2$ and $\mathfrak{o} = \delta_{(1,0)}$, $\mathrm{ES}_\mathfrak{o} = \mathrm{LS}_\theta$ and the counter-example from Example B.II.6 earlier with $\mu := \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(0,1)}$ yields $\mathrm{ES}_\mathfrak{o}(\mu,\mu) > 0$.

Even for probability measures $\mathfrak{o}$ that are absolutely continuous with respect to the uniform measure on $\mathbb{S}^{d-1}$, we can find examples where $\mathrm{ES}_\mathfrak{o}(\mu,\mu) > 0$, as presented in Example B.II.7.

Example B.II.7 (Case where $\mathrm{ES}_\mathfrak{o}(\mu,\mu) > 0$ for any $\mathfrak{o}$)**.** Take $\mathfrak{o}$ any probability measure on $\mathbb{S}^1$ and $\mu := \mathcal{U}(B_{\mathbb{R}^2}(0,1))$ the uniform measure on the Euclidean unit ball of $\mathbb{R}^2$. We have for any $\theta \in \mathbb{S}^1$, $P_\theta\mu = \nu$, where $\nu(\mathrm{d}t) = \frac{2\sqrt{1-t^2}}{\pi}\mathbb{1}_{[-1,1]}(t)\mathrm{d}t$. The disintegration of μ with respect to P_θ is covariant with respect to θ, and the disintegration kernel at $t = P_\theta x$ is $\mu^t = \mathcal{U}\left(\left\{t\theta + v\theta_\perp,\ v \in [-\sqrt{1-t^2},\sqrt{1-t^2}]\right\}\right)$, where we have fixed $\theta_\perp$ a unit

orthogonal vector to θ. The disintegration kernel μ^t is the uniform measure on the ball slice of $B_{\mathbb{R}^2}(0,1) \cap (t\theta + \theta^\perp)$ (with $\theta^\perp := \{x \in \mathbb{R}^d : \theta \cdot x = 0\}$), and can simply be understood as the uniform measure on the segment $[-\sqrt{1-t^2}, \sqrt{1-t^2}]$, cast into $\mathbb{R}^2$. The optimal transport plan between ν and itself is $\pi_\theta := (I, I)\#\nu$, and it follows that the lifted plan between μ and itself is (denoting $t := P_\theta x_1$ for legibility):

$$\gamma_\theta(\mathrm{d}x_1, \mathrm{d}x_2) = \delta_{P_\theta x_1 = P_\theta x_2}(\mathrm{d}P_\theta x_1, \mathrm{d}P_\theta x_2)\ \nu(\mathrm{d}P_\theta x_1) \otimes \mathcal{U}\left(\left\{(t\theta + v_1\theta_\perp, t\theta + v_2\theta_\perp),\ (v_1, v_2) \in [-\sqrt{1-t^2}, \sqrt{1-t^2}]^2\right\}\right)(\mathrm{d}x_1, \mathrm{d}x_2).$$

We provide a visualisation of the disintegration μ^t and the coupling γ_θ in Fig. B.II.16. By symmetry $\mathrm{LS}_\theta(\mu, \mu)$ does not depend on θ, we compute it for $\theta := (1, 0)$:

$$\begin{aligned}
\mathrm{LS}^2_\theta(\mu,\mu) &= \int_{\mathbb{R}^3} \|(u, v_1) - (u, v_2)\|_2^2 \mathbb{1}_{[-1,1]}(u) \frac{2\sqrt{1-u^2}}{\pi} \mathbb{1}_{[-\sqrt{1-u^2}, \sqrt{1-u^2}]^2}(v_1, v_2) \left(\frac{1}{2\sqrt{1-u^2}}\right)^2 \mathrm{d}v_1\mathrm{d}v_2\mathrm{d}u \\
&= \int_{u=-1}^{u=1} \frac{\sqrt{1-u^2}}{2\pi} \int_{v_1=-\sqrt{1-u^2}}^{v_1=\sqrt{1-u^2}} \int_{v_2=-\sqrt{1-u^2}}^{v_2=\sqrt{1-u^2}} (v_1 - v_2)^2 \mathrm{d}v_1\mathrm{d}v_2\mathrm{d}u \\
&= \frac{5\pi}{12} > 0.
\end{aligned}$$

We conclude that $\mathrm{ES}_\sigma(\mu, \mu) > 0$ (for any σ), and thus ES_σ is not a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

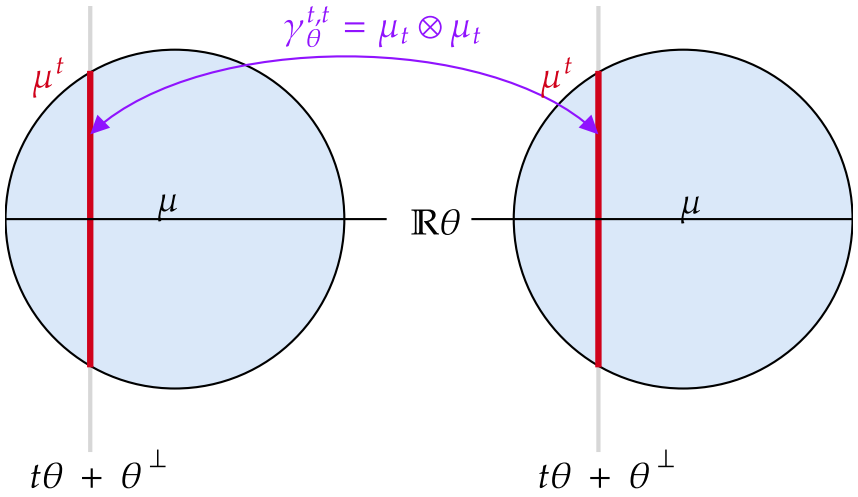


Figure B.II.16: Illustration of the lifted plan γ_θ from Example B.II.7 between μ the uniform measure on the unit Euclidean ball of $\mathbb{R}^2$, and itself. The plan is defined by disintegration: the coupling between $P_\theta\#\mu$ and $P_\theta\#\mu$ is simply (I, I), the coupling induced by the identity map. As for the orthogonal part, the disintegration kernel of μ at $t\theta$ is μ^t, the uniform measure on the ball slice $B_{\mathbb{R}^2}(0,1) \cap (t\theta + \theta^\perp)$, represented as a thick red vertical line. The lifted plan couples the disintegration kernel μ^t with itself with the independent coupling: writing $\gamma_\theta^{t,t}$ as the disintegration kernel of γ_θ at $(t, t) \in [-1, 1]^2$, we have $\gamma_\theta^{(t,t)} = \mu_t \otimes \mu_t$, which corresponds to the uniform measure on the square $\{(t\theta + v_1\theta_\perp, t\theta + v_2\theta_\perp),\ (v_1, v_2) \in [-\sqrt{1-t^2}, \sqrt{1-t^2}]^2\} \subset \mathbb{R}^4$.

Using Proposition B.II.14, we can show that the expected sliced distance is a distance on the set of "countably discrete" probability measures defined in Eq. (B.II.58).

Corollary B.II.3. For any σ a probability measure on $\mathbb{S}^{d-1}$ that is absolutely continuous with respect to σ_u, the quantity ES_σ is a distance on $\mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d)$.

Proof. Thanks to Corollary B.II.2, the only axiom to verify to show that ES_σ is a distance on $\mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d)$ is to show that $\forall \mu \in \mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d)$, $\mathrm{ES}_\sigma(\mu, \mu) = 0$. We now fix $\mu \in \mathcal{P}_{\mathrm{DC}}(\mathbb{R}^d)$. Since

$\varpi \ll \varpi_u$, we have by Proposition B.II.14 that for ϖ-almost-every $\theta \in \mathbb{S}^{d-1}$, $\mathrm{LS}_\theta(\mu,\mu) = 0$. We conclude $\mathrm{ES}^2_\varpi(\mu,\mu) = \int_{\mathbb{S}^{d-1}} \mathrm{LS}^2_\theta(\mu,\mu)\mathrm{d}\varpi(\theta) = 0$. □

B.II.8 Numerics

In this section, we evaluate the efficiency and practicability of the sliced-based transport plans, namely min-Pivot Sliced Wasserstein (min PS) and expected Sliced Wasserstein (ES), in both synthetic and real-world scenarios. We begin by presenting quantitative and qualitative results on toy datasets, evaluating their ability to generate meaningful transport plans and costs across various settings. We continue with a colour transfer task, which is simple to assess qualitatively yet can be computationally challenging in classic OT due to the large sample size ($n \geq 500^2$). We finish with a more complex task that involves large-scale datasets where a transport plan is required, namely point cloud registration. For these experiments, we employ the POT toolbox [Fla+21]. Note that we report experimental results only in the context of distributions with the same number of samples but that the results can be easily extended to the case of different number of samples. All experiments were run on CPU on a MacBook Pro with an M1 chip.

B.II.8.1 Evaluation of the Transport Losses and Plans

B.II.8.1.1 Gradient Flows

We perform a gradient flow on the support of a discrete source distribution μ, aiming to minimise the (Sliced) Wasserstein distance with respect to a discrete target distribution ν: $\min_\mu\{\mathcal{F}^\nu(\mu)\}$, following the setting of [CTV25]. This procedure yields a flow $(\mu_t)_t$ that decreases the functional $\mathcal{F}^\nu(\mu)$ over time $0 \leq t \leq 1$. We consider here several functionals: the Wasserstein distance W_2^2, the Sliced Wasserstein distance SW_2^2, $\min \mathrm{PS}^2$ and ES^2. For $\min \mathrm{PS}$ and ES, at each step we draw randomly L directions $\theta_\ell \in \mathbb{S}^{d-1}$ and compute $\min \mathrm{PS}^2 \approx \min_\ell \mathrm{PS}^2_{\theta_\ell}$ and $\mathrm{ES} \approx \frac{1}{L}\sum_{\ell=1}^L \mathrm{LS}^2_\theta$. Moreover, for $\min \mathrm{PS}$, we use an optimisation scheme described in [CTV25] to obtain an approximation $\hat{\theta}^\star$ of an optimal direction $\theta^\star \in \operatorname{argmin}_{\theta\in\mathbb{S}^{d-1}} \mathrm{PS}^2_\theta$. In what follows, it is denoted $\mathrm{PS}^2_{\hat{\theta}^\star}$.

We consider several target distributions of $n = 50$ samples, shown in the first and third columns of Fig. B.II.17: a Gaussian distribution (in 2 and 500 dimensions), a spiral, two moons, a circle and eight Gaussians of different means. The source distribution is chosen to be a uniform distribution. We use Adam as an optimisation scheme, with a learning rate of 0.02 for all methods, and consider $L = 50$ directions for the sliced approaches. We report the 2-Wasserstein distance between μ_t and ν at each iteration of the optimisation procedure, and repeat each experiment 10 times.

One can observe that the Expected Sliced discrepancy does not converge in any setting. This finding is consistent with the one of [Liu+24, Section 3.4]. In contrast, all other methods enable convergence to the target distribution, i.e. $\mu_t \to \nu$ as $t \to 1$, when working in two dimensions. When considering a 500-dimensional Gaussian distribution, only Wasserstein and $\mathrm{PS}_{\hat{\theta}^\star}$ achieve convergence: with a fixed number of samples n, we suspect that the required number of directions to obtain a good approximation must grow exponentially with the dimension, making $\min \mathrm{PS}$ (with $L = 50$ directions), ES and Sliced Wasserstein inadequate for this context. Using optimisation techniques in $\min \mathrm{PS}$ provides a single meaningful direction $\hat{\theta}^\star$, even when n is small compared to the dimension. One can notice that Wasserstein and $\mathrm{PS}_{\hat{\theta}^\star}$ have very similar behaviours, which is backed by Proposition B.II.11 which states that $\mathrm{PS}_{\hat{\theta}^\star}$ is equal to the 2-Wasserstein distance when $\hat{\theta}^\star$ is an optimal direction and when $d \geq 2n - 1$. This encourages the use of the minimisation method proposed in [CTV25], which outperforms the search over L random projections.

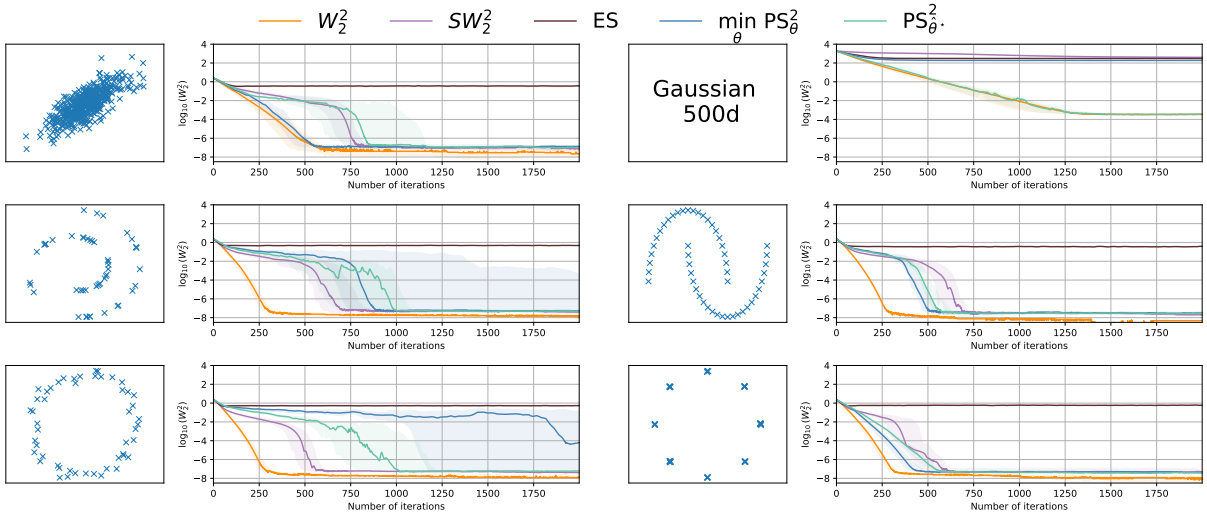


Figure B.II.17: Log 2-Wasserstein distance measured between a source and different target distributions as a function of number of iterations. Plain lines represent the median over 10 iterations while shaded regions indicate 0.25 and 0.75 quantiles.

B.II.8.1.2 Comparison of Transport Plans and Discrepancies

We now provide a quantitative assessment of the transport plans that can be estimated from sliced-based methods.

Qualitative assessment of the transport plans.. We illustrate some transport plans in several two-dimensional settings. The first one corresponds to transporting samples of a source Gaussian distribution to samples of a target Gaussian distribution with different parameters. The second one considers two distributions sampled on circles of the same centre and different radii, with $n = 24$ samples. The last one considers a more challenging and non-linear setting, in which the source distribution is composed of 8 Gaussians of several means and the target is composed of two moons.

Fig. B.II.18 presents the plans obtained with 2-Wasserstein, Expected Sliced and min-Pivot Sliced, together with the associated discrepancy. We choose $L = 50$ directions, and fix $n = 10$ samples for the first scenario and $n = 24$ otherwise. On can notice that, in the simple case of 2 Gaussians as source and target distributions (first line), the transport cost is close to the 2-Wasserstein one. Min-Pivot Sliced provides a plan that is close to the OT one; Expected Sliced provides a highly non-deterministic coupling, associated each source point to numerous targets. When it comes to non-linear settings (third and fifth lines), one can notice that the sliced estimated costs deviate from their OT counterpart: as min PS and ES rely on plans obtained by projecting on a line then lifted to the original space, and because none of these projections capture the true matching, the approximation is quite poor, with spurious matchings between the two parts of the moon. Dedicated variants of Sliced Wasserstein have been proposed in this non-linear setting, for instance *generalised* versions in which the data are projected onto a non linear surface, e.g. [Kol+19a], and *augmented* ones [CYL20] that first embeds the data into a higher dimensional space in which a linear surface better captures the distances. These variants are out of the scope of this chapter, but note that a non-linear variant of min PS has been proposed in [CTV25].

Comparing plans obtained by flows.. To avoid relying on one single direction and to better take into account the non linearities on the distributions, we propose here to build on *flows*, for which different directions can be chosen at each iteration. The second, fourth and sixth lines of Fig. B.II.18 present trajectories obtained when considering such flows, with an SGD optimiser and a fixed learning rate equal to 2 (as recommended by [Bon+15a, under Equation 44], we take a learning rate equal to the dimension). If the flow has converged after

200 steps (that is to say, when the Wasserstein distance between two consecutive step is less than 10^{-6}), we infer a transport plan as the map linking the source and the target sample reached by the flow. This strategy also allows considering Sliced Wasserstein to obtain a plan, as proposed in [Rab+12, Section 3.3]. One can notice that, as expected, 2-Wasserstein flows plan recover the transport plan and that Sliced Wasserstein based plan is close to the actual one. As observed in Section B.II.8.1.1, even in the simple case of 2 Gaussians, Expected Sliced does not converge. When considering min PS, flow-based transport allows enhancing the approximation of the plan, avoiding spurious couplings between the two moons. One further notices that this strategy comes with an extra computational cost as several iterations for computing the flow are needed to obtain the approximation. We present this method to highlight the benefits of stochastic algorithms when using sliced-based methods.

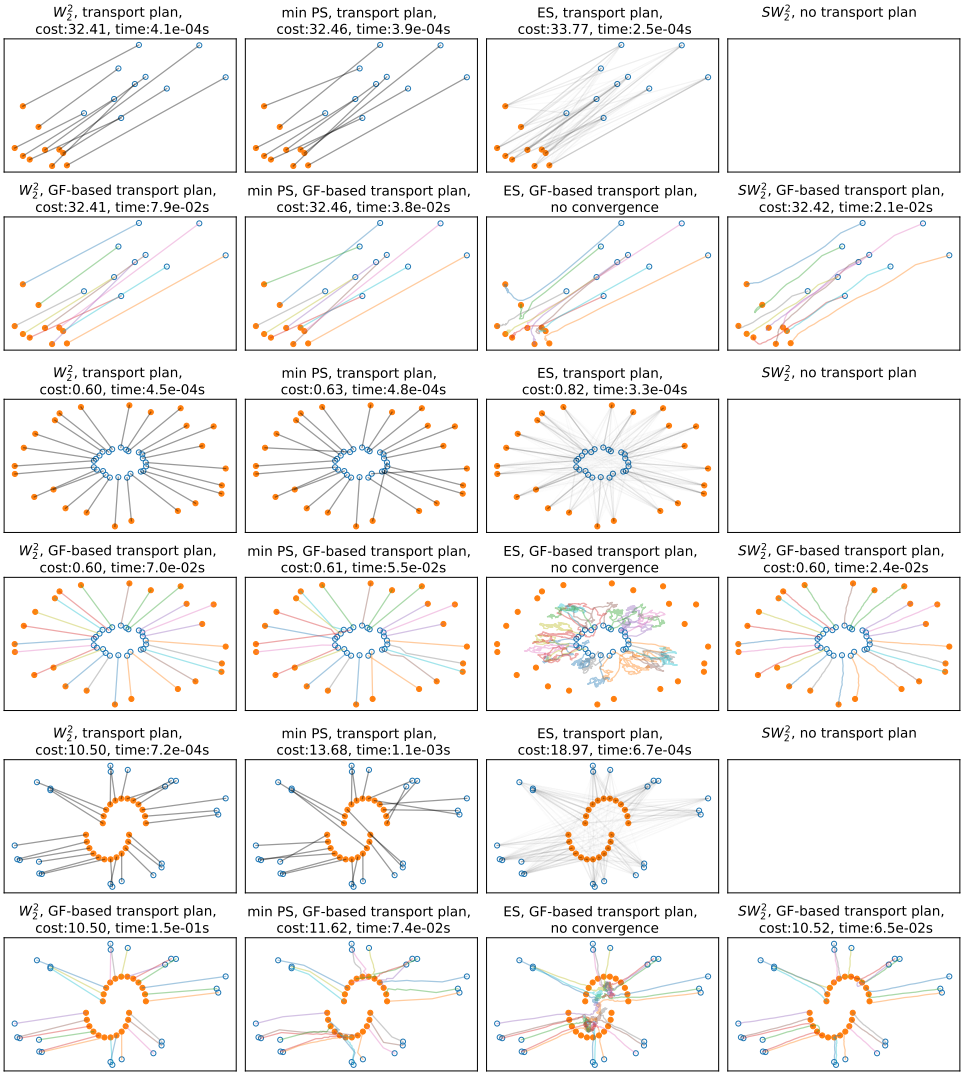


Figure B.II.18: Comparison of the plans obtained by sliced plans methods and 2-Wasserstein between a source (blue samples) and a target (orange samples) distributions. First, third and fifth lines: transport plans obtained by solving Wasserstein, min-Pivot Sliced and Expected Sliced. Second, fourth and sixth lines: trajectories obtained by solving a gradient flow for Wasserstein, min-Pivot Sliced, Expected Sliced and Sliced Wasserstein. In that case, the associated cost is computed by mapping the source sample to the target sample that is reached by the flow.

Timings. We report some timings for the different methods, in order to assess their computational efficiency. We consider the same settings as the first scenario of the previous section (two Gaussians as a source and target distribution). For the Sliced Wasserstein flow, we perform 10 steps, with an extra complexity linear with the number of steps. We vary the number of samples from $n = 10$ to $n = 10^7$, and present the results in Fig. B.II.19. One can notice that

sliced-based method are significantly faster to compute when n grows. Note that Wasserstein fails to be computed for $n \geq 10^5$ due to memory issues, as it requires to store the full cost matrix $C \in \mathbb{R}^{n \times n}$ of size n^2; there is no need to store C for sliced-based methods, that require a memory of $2n$ for min PS and at most of $2Ln$ for ES. The time complexities of all flow variants are proportional to the number of flow steps, and we notice that all sliced methods have comparable complexities in $\mathcal{O}(Lnd + Ln\log(n))$, which is substantially advantageous compared to the $\mathcal{O}(n^3 \log n)$ complexity of standard OT.

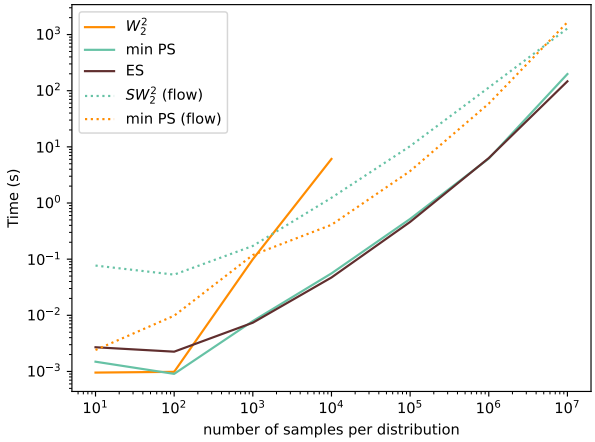


Figure B.II.19: Running time comparisons of different methods for varying number of samples n.

B.II.8.2 Illustration on Colour Transfer

Colour transfer consists in transferring the colour distribution of a source image onto a target image, while preserving the structure of the source. We see an RGB image $I \in \mathbb{R}^{w \times h \times 3}$ as the uniform measure of its pixels in the RGB space $\mu_I := \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \delta_{I_{i,j,\cdot}} \in \mathcal{P}(\mathbb{R}^3)$. Given a source image I and a target image J of same size, our objective is to match (in a certain sense) each pixel (i, j) of I to a pixel (i', j') of J. We consider three different approaches: first, we compute a permutation which is (approximately) optimal for the min PS discrepancy, approximated by searching over $L = 50$ directions. Using this permutation, we replace each pixel of I by its corresponding pixel in J. Second, we approximate the Expected Sliced plan by averaging over $L = 50$ directions. Since this does not yield a permutation but only a transport plan $\overline{\gamma}$, we use the barycentric projection (i.e. conditional expectation) of $\overline{\gamma}$, which provides only an approximate matching to μ_J. Finally, we compare these methods with the Sliced Wasserstein (SW) flow proposed in [Rab+12], which operates 10 steps of Stochastic Gradient Descent with a learning rate of 1 on $X \longmapsto \mathrm{SW}_2^2(\mu_X, \mu_J)$ initialised at $X_0 := I$ and sampling a batch of 3 orthonormal directions at each step. Note that while the final iteration is expected to verify $\mu_X \approx \mu_J$, it may not be the case in practice depending on the hyperparameter choices. We report our results on three different image pairs in Figs. B.II.20 to B.II.22.

Figure B.II.20: Colour transfer example on images of size 1000×669.

Figure B.II.21: Colour transfer example on images of size 1280×1024.

Figure B.II.22: Colour transfer example on images of size 500×500.

In Fig. B.II.20, the source and target images are relatively monochrome, which makes the colour transfer task easier. We observe that the Pivot-Sliced and SW methods are comparable, while the Expected Sliced results in duller colours. Contrastingly, in Fig. B.II.21, the colour palettes are more diverse and Pivot-Sliced yields a visually worse result than SW, while SW matches the colour distributions less faithfully, with some artifacts in the sky. As for Expected Sliced, the results are again duller and quite different to the target colour distribution. Finally, in Fig. B.II.22, only the SW method produces visually consistent results, the matching provided by min PS and ES fail to preserve sufficient spatial structure, in particular in the green colours. Overall, while the plan associated to min PS can suffice in practice, it appears that iterative methods such as the SW flow are better suited for this task. Our experiments suggest that the barycentric projection of the Expected Sliced plan does not provide a sound transportation.

B.II.8.3 Experiments on a Shape Registration Task

We now consider a shape registration task, with a rigid transformation that involves a translation and a rotation. Most approaches to solve this problem are concerned with finding the right correspondences between the points. For instance, the Iterative Closest Point (ICP) algorithm [BM92] relies on nearest neighbour correspondences, considering the Euclidean distance between points. Optimal transport is now a workhorse for this task, as it provides a principled way to find correspondences between two point clouds; see [BD23] for a review of OT-based methods for point cloud registration. We here evaluate the performance of the sliced-based methods, namely min PS and ES, in this context. We compare them to the 2-Wasserstein distance, which is a standard benchmark for point cloud registration, and also to Sliced Wasserstein, using a gradient flow as described in Section B.II.8.1.1 to get an approximated transport plan. Note that Expected Sliced does not provide a one-to-one correspondence but they can be inferred from the blurred transport plan [Sol+15].

We consider two point clouds of 3D shapes, which are subsampled from the *bunny* and *armadillo* shapes of the *open3d* library [ZPK18]. We first subsample both shapes with $n = 500$ points, and then we apply 10 different rigid transformations to the source shape to get the target shape. We then run the ICP algorithm, with several alignment methods: the nearest neighbour correspondence, Wasserstein, Sliced Wasserstein, min-Pivot Sliced and Expected Sliced, to realign the two shapes. Fig. B.II.23 presents the two shapes, subsampled with $n = 2000$ points for visualization purposes. The second line presents the Wasserstein distance between the (registered) source and target point clouds along the iterations. One can notice that Min-Pivot Sliced yields the best registration among all methods: this conclusion was also reached by [Mah+23]

who conjecture that it allows exiting local minima of the ICP algorithm by finding an approximated matching.

We also consider the case where the shapes are not subsampled, which is a more computationally challenging setup, especially for the armadillo shape. One can drawn similar conclusions, with min PS yielding the best registration, with little variation around the different repetitions of the experiment.

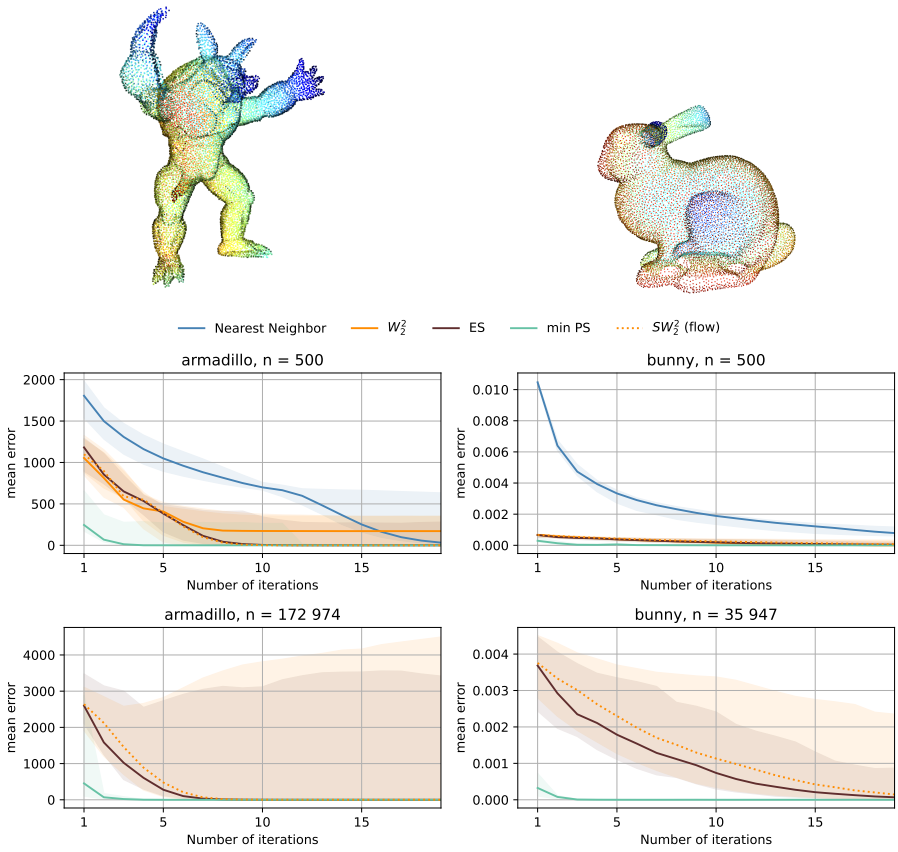


Figure B.II.23: Evolution of the loss along the iterations of the ICP algorithm. The loss is computed as the mean square distance between each target sample and the registered source. The first column corresponds to the results for the armadillo shape, while the second column corresponds to the bunny shape.

Acknowledgements

We would like to thank Nathaël Gozlan and Agnès Desolneux for their insights on technical aspects of Section B.II.2.

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

B.II.9 Appendix

B.II.9.1 Ambiguity in SWGG from [Mah+23]

Let $\mu_1 = \frac{1}{n}\sum_i \delta_{x_i}$, $\mu_2 = \frac{1}{n}\sum_i \delta_{y_i}$, and $\theta \in \mathbb{S}^{d-1}$. Consider σ_θ a permutation which sorts $(\theta^\top x_i)_{i=1}^n$ and τ_θ sorting $(\theta^\top y_i)_{i=1}^n$. The Sliced Wasserstein Generalised Geodesic distance ([Mah+23, Equation 8]) is defined as

$$\mathrm{SWGG}_2^2(\mu_1, \mu_2, \theta) := \frac{1}{n}\sum_{i=1}^n \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2. \tag{B.II.60}$$

We illustrate the coupling induced by $\mathrm{SWGG}_2^2(\mu_1, \mu_2, \theta)$ in Fig. B.II.24:

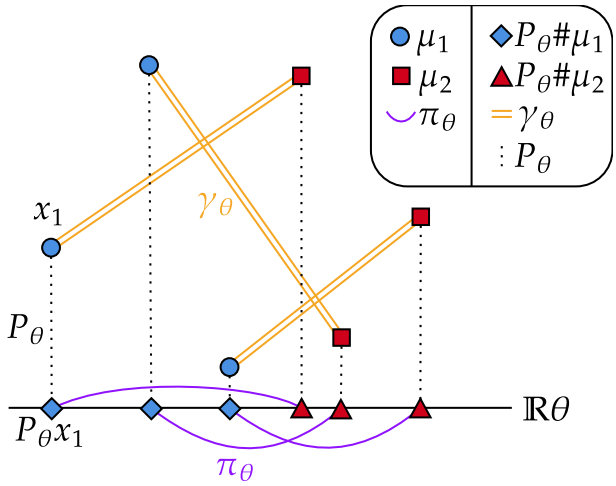


Figure B.II.24: Coupling $\gamma_\theta \in \Pi(\mu_1, \mu_2)$ induced by $\mathrm{SWGG}_2^2(\mu_1, \mu_2, \theta)$ for $d = 2$, $n = 3$, $\theta = (1, 0)$. The support of the measure μ_1 is represented by blue circles, and the support of μ_2 with red squares. The projected measures $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$ are represented by the blue diamonds and triangles respectively. The optimal coupling between $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$ is drawn with purple curves, and the associated coupling γ_θ between μ_1 and μ_2 is represented by the orange double lines. In this example, the projections of the points of the support of μ_1 are distinct (as for μ_2), thus the coupling π_θ determines uniquely the coupling γ_θ, there is no ambiguity.

Unfortunately, the right hand-side quantity in Eq. (B.II.60) depends on the choice of the permutations, rendering the quantity ill-defined, as showcased in Example B.II.8.

Example B.II.8 (Ambiguity in SWGG)**.** Consider $d = 2$, $n = 2$, the points $x_1 = (0, 1)$, $x_2 = (0, 0)$, $y_1 = (0, 0)$, $y_2 = (0, 1)$, the line $\theta = (1, 0)$ and the measures $\mu_1 = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$, $\mu_2 = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$. We have $\mu_1 = \mu_2$, and $\theta^\top u = 0$ for all points $u \in \{x_1, x_2, y_1, y_2\}$, hence any choice of permutations $(\sigma_\theta, \tau_\theta)$ sorts the respective points $(\theta^\top x_i)$ and $(\theta^\top y_i)$. Choosing $(\sigma_\theta, \tau_\theta) = (I, I)$, we obtain

$$\mathrm{SWGG}_2^2(\mu_1, \mu_2, \theta) = \tfrac{1}{2}(\|x_1 - y_1\|_2^2 + \|x_2 - y_2\|_2^2) = 1,$$

which in particular in non-zero, which shows that $\mathrm{SWGG}_2(\cdot, \cdot, \theta)$ is not a distance. Another possible choice $(\sigma_\theta, \tau_\theta) = ((2, 1), (2, 1))$ yields a value of 0.

One could consider the following "fix" to the permutation choice issue:

$$\widetilde{\mathrm{SWGG}}_2^2(\mu_1, \mu_2, \theta) := \min_{(\sigma_\theta, \tau_\theta) \in \mathfrak{S}_\theta(X, Y)} \frac{1}{n} \sum_{i=1}^{n} \|x_{\sigma_\theta(i)} - y_{\tau_\theta(i)}\|_2^2, \tag{B.II.61}$$

where $\mathfrak{S}_\theta(X, Y)$ is the set of pairs of permutations $(\sigma_\theta, \tau_\theta)$ that sort $(\theta^\top x_i)_{i=1}^n$ and $(\theta^\top y_i)_{i=1}^n$ respectively. We illustrate this idea in Fig. B.II.25.

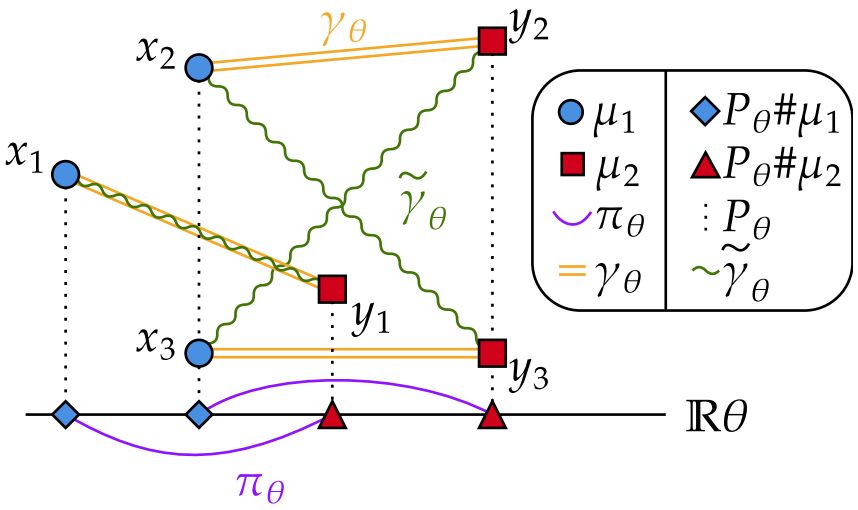


Figure B.II.25: In this example, the projections sometimes coincide, and the optimal coupling π_θ between $P_\theta\#\mu_1$ and $P_\theta\#\mu_2$ does not determine the coupling between (x_2, x_3) and (y_2, y_3). In terms of permutations, there are two possibilities: $\gamma_\theta := \frac{1}{3}(\delta_{x_1\otimes y_1} + \delta_{x_2\otimes y_2} + \delta_{x_3\otimes y_3})$ displayed with orange double lines, and $\widetilde{\gamma}_\theta := \frac{1}{3}(\delta_{x_1\otimes y_1} + \delta_{x_2\otimes y_3} + \delta_{x_3\otimes y_2})$ represented by green squiggly lines. Here, the cost of γ_θ is lower, so we would choose it.

B.II.9.2 Midpoints are Geodesic Middles

In the following, we remind a well-known simple result about geodesic spaces, which we apply to show that Wasserstein means are middles of Wasserstein geodesics (see Proposition B.II.3). We consider a geodesic space $(\mathcal{X}, d)$, which is to say that d is a distance on $\mathcal{X}$ such that for any $(x_1, x_2) \in \mathcal{X}^2$ there exists a curve $\gamma : [0,1] \longrightarrow \mathcal{X}$ with $\gamma(0) = x_1$ and $\gamma(1) = x_2$ such that $d(\gamma(t), \gamma(s)) = |t - s| d(x_1, d_2)$. Such a curve is called a geodesic between x_1 and x_2.

Lemma B.II.12. Let $(\mathcal{X}, d)$ be a geodesic space, let $x_1, x_2 \in \mathcal{X}$ and consider the set $\mathrm{M}(x_1, x_2)$ of Midpoints:

$$\mathrm{M}(x_1, x_2) = \operatorname*{argmin}_{y\in\mathcal{X}} d(x_1, y)^2 + d(y, x_2)^2. \tag{B.II.62}$$

This set is in fact exactly the set of middles of geodesics:

$$\mathrm{M}(x_1, x_2) = \left\{\gamma(\tfrac{1}{2}) \mid \gamma \text{ is a geodesic between } x_1 \text{ and } x_2\right\}. \tag{B.II.63}$$

Proof. Denote by $\mathrm{M}'(x_1, x_2)$ the right hand-side of Eq. (B.II.63), first we show $\mathrm{M}'(x_1, x_2) \subset \mathrm{M}(x_1, x_2)$ and compute the optimal value of Eq. (B.II.62). Let γ a constant-speed geodesic between x_1 and x_2, we have

$$d(x_1, \gamma(\tfrac{1}{2}))^2 + d(\gamma(\tfrac{1}{2}), x_2)^2 = d(\gamma(0), \gamma(\tfrac{1}{2}))^2 + d(\gamma(\tfrac{1}{2}), \gamma(1))^2 = d(x_1, x_2)^2/2.$$

Now take any $y \in \mathcal{X}$, we have (by convexity of $t \longmapsto t^2$, then by the triangle inequality for d)

$$\begin{aligned} d(x_1, y)^2 + d(y, x_2)^2 &= 2(d(x_1, y)^2/2 + d(y, x_2)^2/2) && \text{(B.II.64)}\\ &\geq 2(d(x_1, y)/2 + d(y, x_2)/2)^2 && \text{(B.II.65)}\\ &\geq d(x_1, x_2)^2/2. && \text{(B.II.66)} \end{aligned}$$

This shows that any such $\gamma(\frac{1}{2})$ is solution of the optimisation problem which defines $\mathrm{M}(x_1, x_2)$, and thus $\mathrm{M}'(x_1, x_2) \subset \mathrm{M}(x_1, x_2)$. The value of the minimisation problem from Eq. (B.II.62) is $d(x_1, x_2)^2/2$.

Let $y^* \in \mathrm{M}(x_1, x_2)$, we now show that $d(x_1, y^*) = d(y^*, x_2) = d(x_1, x_2)/2$. Since y^* is optimal and that the optimal value is $d(x_1, x_2)^2$, the inequalities Eq. (B.II.65) and Eq. (B.II.66) are

equalities for $y := y^*$. First, Eq. (B.II.65) yields $d(x_1, y^*) = d(y^*, x_2)$, then Eq. (B.II.66) yields $d(x_1, y^*) = d(x_1, x_2)/2$.

We now show that $\mathrm{M}(x_1, x_2) \subset \mathrm{M}'(x_1, x_2)$: let $y^* \in \mathrm{M}(x_1, x_2)$, consider γ_1 a geodesic from x_1 to y^*, and γ_2 a geodesic from y^* to x_2. We introduce the curve

$$\gamma : \left\{ \begin{array}{ccc} [0,1] & \longrightarrow & \mathcal{X} \\ t & \longmapsto & \left\{ \begin{array}{ll} \gamma_1(2t) & \text{if } t \in [0, \frac{1}{2}]; \\ \gamma_2(2t-1) & \text{if } t \in [\frac{1}{2}, 1]. \end{array} \right. \end{array} \right.$$

Our objective is to show that γ is a geodesic from x_1 to x_2 (since $\gamma(\frac{1}{2}) = y^*$, this will show that $y^* \in \mathrm{M}'(x_1, x_2)$). By construction $\gamma(0) = x_1$, $\gamma(1) = x_2$. Let $(t, s) \in [0, 1]^2$ with $t \leq s$, we want to prove $d(\gamma(t), \gamma(s)) = |s - t| d(x_1, x_2)$.

Firstly, we consider the case $(t, s) \in [0, \frac{1}{2}]^2$. In this case,

$$d(\gamma(t), \gamma(s)) = d(\gamma_1(2t), \gamma_1(2s)) = (2s - 2t) d(x_1, y^*) = (s - t) d(x_1, x_2),$$

where we used $d(x_1, y^*) = d(x_1, x_2)/2$, which we proved earlier for any optimal y^*. The case $(t, s) \in [\frac{1}{2}, 1]^2$ can be treated similarly.

Secondly, we assume $t \in [0, \frac{1}{2}]$ and $s \in [\frac{1}{2}, 1]$. We first prove $d(\gamma(t), \gamma(s)) \leq (s - t) d(x_1, x_2)$ using the triangle inequality and $d(x_i, y^*) = d(x_1, x_2)/2$ for $i \in \{1, 2\}$:

$$\begin{aligned} d(\gamma(t), \gamma(s)) &\leq d(\gamma(t), y^*) + d(y^*, \gamma(s)) \\ &= d(\gamma_1(2t), \gamma_1(1)) + d(\gamma_2(0), \gamma_2(2s-1)) \\ &= (1 - 2t) d(x_1, y^*) + (2s - 1) d(y^*, x_2) \\ &= (s - t) d(x_1, x_2). \end{aligned}$$

For the converse inequality $d(\gamma(t), \gamma(s)) \geq (s - t) d(x_1, x_2)$, we apply the triangle inequality:

$$d(x_1, x_2) \leq d(x_1, \gamma(t)) + d(\gamma(t), \gamma(s)) + d(\gamma(s), x_2),$$

which yields:

$$\begin{aligned} d(\gamma(t), \gamma(s)) &\geq d(x_1, x_2) - d(\gamma_1(0), \gamma_1(2t)) - d(\gamma_2(2s-1), x_2) \\ &= (1 - t - (1 - s)) d(x_1, x_2) = (s - t) d(x_1, x_2). \end{aligned}$$

The case $s \in [0, \frac{1}{2}]$ and $t \in [\frac{1}{2}, 1]$ is done symmetrically and thus $d(\gamma(t), \gamma(s)) = |s - t| d(x_1, x_2)$, which shows that $y^* \in \mathrm{M}'(x_1, x_2)$. We conclude that $\mathrm{M}'(x_1, x_2) = \mathrm{M}(x_1, x_2)$. □

B.II.9.3 Reminders on Disintegration of Measures

In Definition B.II.10, we recall the definition of disintegration of measures with respect to a map (taken from [AGS05, Theorem 5.3.1]). By slight abuse of notation, we will write $P^{-1}(y) := P^{-1}(\{y\})$ for a map $P : \mathcal{X} \longrightarrow \mathcal{Y}$ that need not be injective and $y \in \mathcal{Y}$.

Definition B.II.10. Consider a Borel map $P : \mathcal{X} \longrightarrow \mathcal{Y}$ between Polish spaces $\mathcal{X}, \mathcal{Y}$ and $\mu \in \mathcal{P}(\mathcal{X})$. There exists a $P\#\mu$-almost-everywhere unique Borel family $(\mu^y)_{y \in \mathcal{Y}} \subset \mathcal{P}(\mathcal{X})$ of measures verifying $\mu^y(\mathcal{X} \setminus P^{-1}(y)) = 0$, and verifying the following identity against test functions $\phi \in \mathcal{C}_b^0(\mathcal{X})$:

$$\int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) = \int_{\mathcal{Y}} \left(\int_{P^{-1}(y)} \phi(x) \mathrm{d}\mu^y(x) \right) \mathrm{d}(P\#\mu)(y). \tag{B.II.67}$$

We will write Eq. (B.II.67) symbolically as:

$$\mu(\mathrm{d}x) = (P\#\mu)\,(P(\mathrm{d}x))\ \ \mu^{P(x)}(\mathrm{d}x). \tag{B.II.68}$$

For example, in the case $\mathcal{X} = \mathbb{R}^d \times \mathbb{R}^d$ and $P(y,x) = y$, the disintegration corresponds to the disintegration with respect to the first marginal ν of a coupling $\gamma \in \Pi(\nu,\mu)$. In this case, each measure γ^y is a measure of $\mathcal{P}(\mathbb{R}^{2d})$ concentrated on the slice $\{y\} \times \mathbb{R}^d$, which is routinely identified as a measure on $\mathbb{R}^d$ in literature. This disintegration is written symbolically as $\gamma(\mathrm{d}y, \mathrm{d}x) = \nu(\mathrm{d}y)\gamma^y(\mathrm{d}x)$.

B.II.9.4 Proof of the Disintegration Formula for ν-based Wasserstein

In this section, we provide a proof to Theorem B.II.1, and use the notation from the statement. Let $\rho \in \Gamma(\nu,\mu_1,\mu_2)$ (see Eq. (B.II.7)), we have

$$\begin{aligned}\int_{\mathbb{R}^{3d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho(y,x_1,x_2) &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^{2d}} \|x_1 - x_2\|_2^2 \mathrm{d}\rho^y(x_1,x_2) \right) \mathrm{d}\nu(y) \\ &\geq \int_{\mathbb{R}^d} \mathrm{W}_2^2(P_1\#\rho^y, P_2\#\rho^y)\mathrm{d}\nu(y), \end{aligned} \tag{B.II.69}$$

where we wrote the disintegration $\rho(\mathrm{d}y, \mathrm{d}x_1, \mathrm{d}x_2) = \nu(\mathrm{d}y)\rho^y(\mathrm{d}x_1, \mathrm{d}x_2)$. Note that by [AGS05, Lemma 12.4.7], the map $y \longmapsto \mathrm{W}_2^2(P_1\#\rho^y, P_2\#\rho^y)$ is Borel.

Now since $\rho \in \Gamma(\nu,\mu_1,\mu_2)$, we can write $P_{1,2}\#\rho =: \gamma_1 \in \Pi^*(\nu,\mu_1)$ and $P_{1,3}\#\rho =: \gamma_2 \in \Pi^*(\nu,\mu_2)$. It follows that for ν-almost every $y \in \mathbb{R}^d$, we have for $i \in \{1,2\}$ that $P_i\#\rho^y = \gamma_i^y$, where we disintegrated $\gamma_i(\mathrm{d}y,\mathrm{d}x) = \nu(\mathrm{d}y)\gamma_i^y(\mathrm{d}x)$ (for example by [AGS05, Lemma 5.3.2]). Taking the infimum on ρ on both sides yields

$$\mathrm{W}_\nu^2(\mu_1,\mu_2) \geq \inf_{\gamma_i \in \Pi^*(\nu,\mu_i),\, i\in\{1,2\}} \int_{\mathbb{R}^d} \mathrm{W}_2^2(\gamma_1^y, \gamma_2^y)\mathrm{d}\nu(y). \tag{B.II.70}$$

Fixing $\gamma_i \in \Pi^*(\nu,\mu_i)$ for $i \in \{1,2\}$, we now construct a 3-plan $\rho \in \Gamma(\nu,\mu_1,\mu_2)$ which attains the lower bound in Eq. (B.II.69). Consider the disintegrations $\gamma_i(\mathrm{d}y,\mathrm{d}x) = \nu(\mathrm{d}y)\gamma_i^y(\mathrm{d}x)$ for $i \in \{1,2\}$. The two families $(\gamma_i^y)_{y\in\mathbb{R}^d}$ are Borel in $\mathcal{P}_2(\mathbb{R}^d)$, hence by [AGS05, Lemma 12.4.7], there exists a Borel family $(\rho^y)_{y\in\mathbb{R}^d}$ in $\mathcal{P}_2(\mathbb{R}^{2d})$ such that for all $y \in \mathbb{R}^d$, $\rho^y \in \Pi^*(\gamma_1^y, \gamma_2^y)$. Setting $\rho(\mathrm{d}y, \mathrm{d}x_1, \mathrm{d}x_2) := \nu(\mathrm{d}y)\rho^y(\mathrm{d}x_1, \mathrm{d}x_2)$ yields the desired 3-plan, since for ν-almost every $y \in \mathbb{R}^d$, ρ^y is an optimal transport plan between γ_1^y and γ_2^y. We have shown that

$$\forall \gamma_i \in \Pi^*(\nu,\mu_i),\ i \in \{1,2\},\ \mathrm{W}_\nu^2(\mu_1,\mu_2) \leq \int_{\mathbb{R}^d} \mathrm{W}_2^2(\gamma_1^y, \gamma_2^y)\mathrm{d}\nu(y), \tag{B.II.71}$$

which shows the equality in Eq. (B.II.19).

We finish by showing that the infimum in Eq. (B.II.19) is indeed attained. Note that having the weak convergence of plans $(\gamma_n) \in \Pi(\nu,\mu_1)$ does not yield the ν-almost-everywhere convergence of the disintegrations γ_n^y in general. Thankfully, we can leverage the existence of a solution of the original formulation from Eq. (B.II.12) by Proposition B.II.1. Using the fact that the two problems have the same value, we can take a solution ρ of Eq. (B.II.12) and construct a solution of Eq. (B.II.19) by disintegration.

B.III

Sliced Gromov Wasserstein

B.III.1 Purpose and Motivation . . . 251
B.III.2 Frank-Wolfe for the GW Problem . . . 252
B.III.3 Bi-Convex Relaxation and Nested Wasserstein Formulation . . . 253
B.III.4 Expression of the BCR Between Point Clouds . . . 253
B.III.5 Expression of the BCR Between Discrete Measures for 2-Gromov . . . 254
B.III.6 Heuristic of Sliced Permutations for the BCR Between Point Clouds . . . 256

Abstract

In this chapter, we will consider an application of Sliced Optimal Transport Plans which computes a heuristic approximation of the Gromov-Wasserstein (GW) distance. We begin by explaining the state-of-the-art Frank-Wolfe (FW) solver for GW, and reformulate a well-known biconvex relaxation of the GW problem. Replacing the OT plans in this relaxation with Sliced OT plans, we obtain a new heuristic approximation of the GW distance, which we study numerically and from a complexity viewpoint. As it stands, this method appears not be be competitive with the FW method.

Based on a joint project with Laetitia Chapel, Julie Delon and Nicolas Courty.

B.III.1 Purpose and Motivation

Introduction to GW. The Gromov-Wasserstein (GW) distance [Mém11] is an optimisation problem which allows the comparison of two probability measures μ and ν on different metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$. To compare points within the respective spaces $\mathcal{X}$ and $\mathcal{Y}$, we consider cost functions $c_{\mathcal{X}} : \mathcal{X}^2 \longrightarrow \mathbb{R}$ and $c_{\mathcal{Y}} : \mathcal{Y}^2 \longrightarrow \mathbb{R}$ on $\mathcal{X}$ and $\mathcal{Y}$. Two pairs of points $(x, x') \in \mathcal{X}^2$ and $(y, y') \in \mathcal{Y}^2$ are considered similar if their respective costs $c_{\mathcal{X}}(x, x')$ and $c_{\mathcal{Y}}(y, y')$ are close. This idea is lifted to compare measures μ and ν: the p-GW problem looks for a transport plan $\pi \in \Pi(\mu, \nu)$ such that pairs (x, y) and (x', y') in the support of π are similar in the sense that they minimise $|c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y')|$:

$$\mathrm{GW}_p^p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{(\mathcal{X}\times\mathcal{Y})^2} |c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y')|^p \mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y'). \tag{B.III.1}$$

For discrete measures $\mu := \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ and $\nu := \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}(\mathcal{Y})$, Eq. (B.III.1) becomes:

$$\mathrm{GW}_p^p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = \min_{P \in \Pi(a,b)} \sum_{i,j,k,\ell} |c_{\mathcal{X}}(x_i, x_k) - c_{\mathcal{Y}}(y_j, y_\ell)|^p P_{i,j} P_{k,\ell}, \tag{B.III.2}$$

where $\Pi(a,b)$ is the set of transport plans matrices $P \in \mathbb{R}_+^{n\times m}$ such that $P\mathbf{1}_m = a$ and $P^\top \mathbf{1}_n = b$. Without considering underlying spaces, measures and costs, the discrete GW problem Eq. (B.III.2) can be viewed as a (non-convex) quadratic program:

$$\mathrm{G}_p(a,b,C,D) := \min_{P\in\Pi(a,b)} \sum_{i,j,k,\ell} |C(i,k) - D(j,\ell)|^p P_{i,j} P_{k,\ell}, \tag{B.III.3}$$

which we define for weights $a \in \triangle_n$, $b \in \triangle_m$ and cost matrices $C \in \mathbb{R}^{n\times n}, D \in \mathbb{R}^{m\times m}$. Obviously, Eq. (B.III.2) is a particular case of Eq. (B.III.3), as can be seen by introducing $C(i,k) := c_{\mathcal{X}}(x_i, x_k)$ and $D(j,\ell) := c_{\mathcal{Y}}(y_j, y_\ell)$ are the cost matrices associated to the measures μ and ν.

Computational Challenges. To begin with, as a non-convex Quadratic Problem, the discrete GW problem Eq. (B.III.3) is NP-hard [SG76; Loi+07]. Furthermore, naive methods require the computation (and storage in memory) of the cost tensor $\mathbf{C}_{i,j,k,\ell} = |C(i,k) - D(j,\ell)|^p$ which is in $\mathcal{O}(n^2m^2)$. Several approximation methods have been proposed, starting with Entropic regularisation [PCS16; Sol+16; CM19; Xu+19]. [Li+] solve a relaxation of the GW problem using the Bregman Alternated Projected Gradient method. A Conditional Gradient (a.k.a. Frank Wolfe) solver was popularised by [Vay+20] wherein it was introduced for a generalisation of GW. Finally, [Vay+19] propose a slicing method for GW between measures on (different) Euclidean spaces, relying on a 1D formula for GW (that holds only sometimes [BHS23a]).

Purpose of the Chapter. We begin with a short reminder on the Frank-Wolfe algorithm for GW Section B.III.2. In Section B.III.3, we consider a relaxation of the GW problem called the Bi-Convex Relaxation (BCR), which was considered in [Mém11], and remind a "Nested Wasserstein" formulation that was pointed out in [Vay20] (Equation 2.60). We formulate more amenable expressions in the case of point clouds in Section B.III.4 and of general discrete measures for a GW order $p = 2$ in Section B.III.5. In Section B.III.6, we investigate a heuristic for the computation of the BCR expression from Section B.III.4. Unfortunately, the numerical results are not conclusive.

B.III.2 Frank-Wolfe for the GW Problem

In this section, we remind the Frank-Wolfe algorithm [FW+56] for the GW problem. This technique was popularised by [Vay+20] for a generalisation of the GW coined the Fused Gromov-Wasserstein (FGW) problem. First, we remind the discrete 2-GW cost that we wish to minimise over $P \in \Pi(a,b)$:

$$J_{\mathrm{GW}}(P) := \sum_{i,j,k,\ell} \left(C(i,k) - D(j,\ell)\right)^2 P_{i,j} P_{k,\ell}. \tag{B.III.4}$$

Noticing that J_{GW} is $\mathcal{C}^1$ we compute its (Euclidean) gradient $\nabla J_{\mathrm{GW}}(P) \in \mathbb{R}^{n\times m}$:

$$\nabla J_{\mathrm{GW}}(P) = \left[\sum_{k,\ell} \left(C(i,k) - D(j,\ell)\right)^2 P_{k,\ell} + \sum_{k,\ell} \left(C(k,i) - D(\ell,j)\right)^2 P_{k,\ell} \right]_{i,j}. \tag{B.III.5}$$

The idea of the Frank-Wolfe algorithm [FW+56] is to iterate minimisation of a linearisation of the cost function J_{GW} at the current point P_t, as formalised in Algorithm B.III.1.

Algorithm B.III.1: Frank-Wolfe for the GW Problem

Input: Cost matrices $C \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, weights $a \in \triangle_n, b \in \triangle_m$, number of iterations T.

Initialisation: $P_0 \leftarrow a \otimes b$;
for $t \in [\![0, T-1]\!]$ **do**
  Compute $M_{t+1} \to \nabla J_{\mathrm{GW}}(P_t)$;
  Solve $Q_{t+1} \in \mathrm{argmin}_{Q\in\Pi(a,b)} \langle M_{t+1}, Q\rangle$;
  Solve $\tau_{t+1} \in \mathrm{argmin}_{\tau\in[0,1]} \; J_{\mathrm{GW}}\left((1-\tau)P_t + \tau Q_{t+1}\right)$;
  $P_{t+1} \to (1-\tau_{t+1})P_t + \tau_{t+1} Q_{t+1}$
Return P_T.

Details for the practical computation of the line search step (Line 5) can be found in [Vay+20, Algorithm 2]. Converge to a local stationary point of J_{GW} is guaranteed by [Lac16, Theorem 1], since ∇J_{GW} is $L_{J_{\mathrm{GW}}} := \max_{i,j,k,\ell}(C(i,k) - D(k,\ell))^2$-Lipschitz continuous. The convergence rate is shown to be in $\mathcal{O}(1/\sqrt{t})$, which is to say that it takes at most $\mathcal{O}(\varepsilon^{-2})$ iterations to reach an ε-stationary point. Note that the constant in the $\mathcal{O}$ notation depends on the initialisation P_0 and on the Lipschitz constant $L_{J_{\mathrm{GW}}}$.

Regarding time complexity, the computation of the gradient (Line 3) can be done in $\mathcal{O}(n^2m + nm^2)$ using the trick from [PCS16, Proposition 1], which is specific to the order $p = 2$. The resolution of the discrete OT problem in Line 4 is in $\mathcal{O}((n+m)nm\log(n+m)\log((n+m)\|M_{t+1}\|_\infty))$ [Tar97] with the network simplex algorithm. The line-search step (Line 5) is in $\mathcal{O}(n^2m + nm^2)$ using the computations of [Vay+20, Algorithm 2].

B.III.3 Bi-Convex Relaxation and Nested Wasserstein Formulation

We consider the following relaxation of the GW problem, called the Bi-Convex Relaxation (BCR) [Mém11]:

$$\mathrm{BCR}_p(a,b,C,D) := \min_{P,Q\in\Pi(a,b)} \sum_{i,j,k,\ell} |C(i,k) - D(j,\ell)|^p P_{k,\ell} Q_{i,j} =: J_{\mathrm{BCR}}(P,Q), \tag{B.III.6}$$

where for convenience wee will often omit the index sets $(i,j,k,\ell) \in [\![1,n]\!]\times[\![1,m]\!]\times[\![1,n]\!]\times[\![1,m]\!]$ We focus on the "inner" optimisation, that is, with P fixed. As in [Vay20, Equation 2.60], we consider for $(k,\ell) \in [\![1,n]\!] \times [\![1,m]\!]$ the following one-dimensional measures:

$$\alpha_k := \sum_{i=1}^n a_i \delta_{C(i,k)}, \quad \beta_\ell := \sum_{j=1}^m b_j \delta_{D(j,\ell)}.$$

We observe the following expression of the inner optimisation:

$$\min_{P\in\Pi(a,b)} \sum_{i,j} |C(i,k) - D(j,\ell)|^p Q_{i,j} = \mathrm{W}_p^p(\alpha_k, \beta_\ell),$$

resulting in the "Nested Wasserstein" formulation:

$$\mathrm{BCR}_p(a,b,C,D) = \min_{P\in\Pi(a,b)} \sum_{k,\ell} \mathrm{W}_p^p(\alpha_k,\beta_\ell) P_{k,\ell}. \tag{B.III.7}$$

We recognise in Eq. (B.III.7) a Kantorovich problem with the cost matrix $M := (\mathrm{W}_p^p(\alpha_k,\beta_\ell))_{k,\ell}$. Given that the measures α_k, β_ℓ are one-dimensional, each cost $\mathrm{W}_p^p(\alpha_k,\beta_\ell)$ can be computed in $\mathcal{O}(n\log(n) + m\log(m))$, yielding a total cost of $\mathcal{O}(n^2m\log(n) + nm^2\log(m))$ for the cost matrix. The resolution of the this outer Kantorovich problem can be done in $\mathcal{O}((n+m)nm\log(n+m)\log((n+m)\|M\|_\infty))$ [Tar97] with the network simplex algorithm.

B.III.4 Expression of the BCR Between Point Clouds

We now focus on the case of point clouds, which is to say that the weights a, b are supposed uniform and $n = m$. We denote by $\mathbb{U} := \left\{P \in \mathbb{R}_+^{n\times n} : P^\top \mathbb{1} = P\mathbb{1} = \frac{1}{n}\mathbb{1}\right\}$ the set of valid transport plans between the uniform discrete measures. We are interested in the Monge case, which is to say the case where we look for transport plans that are permutations.

Fixing Q. To study the alternate optimisation scheme, we now fix $Q = Q^\tau$, a plan associated to a permutation τ, which is to say that $Q^\tau_{i,j} = \frac{1}{n}\mathbb{1}(\tau(i) = j)$. As before we omit the index sets,

we remind that $(i,j,k,\ell) \in [\![1,n]\!] \times [\![1,m]\!] \times [\![1,n]\!] \times [\![1,m]\!]$.

$$\begin{aligned}\min_{P\in\mathbb{U}} J_{\mathrm{BCR}}(P,Q^\tau) &= \min_{P\in\mathbb{U}} \sum_{i,j,k,\ell} |C(i,k)-D(j,\ell)|^p \frac{1}{n}\mathbb{1}(\tau(i)=j)P_{k,\ell} \\ &= \frac{1}{n}\min_{P\in\mathbb{U}} \sum_{(k,\ell)\in[\![1,n]\!]^2} \|C(\cdot,k)-D(\cdot,\ell)\circ\tau\|_p^p\, P_{k,\ell} \\ &= \frac{1}{n}\mathrm{W}_p^p(\alpha,\beta),\end{aligned}$$

where

$$\alpha := \frac{1}{n}\sum_{k=1}^n \delta_{C(\cdot,k)}, \quad \beta := \frac{1}{n}\sum_{\ell=1}^n \delta_{D(\cdot,\ell)\circ\tau}.$$

Here we see C as a discrete measure of n points in $\mathbb{R}^n$, with each column $C(\cdot,k)$ as a point. For D, we first permute the lines of the matrix D with the permutation τ, then see the columns of D as points in $\mathbb{R}^n$.

Notice an optimal plan P^* obtained in the minimisation of the Wasserstein distance can be chosen of the form P^σ for some permutation σ, since the measures α and β are uniform and have the same amount of points n (see [PC19b, Proposition 2.1]).

Fixing P. Regarding the other step in the alternate optimisation scheme, we will permute the columns instead of the lines, and the the matrices as discrete measures with the lines as points in $\mathbb{R}^n$: fix $P = P^\sigma$,

$$\begin{aligned}\min_{Q\in\mathbb{U}} J_{\mathrm{BCR}}(P^\sigma,Q) &= \min_{Q\in\mathbb{U}} \sum_{i,j,k,\ell} |C(i,k)-D(j,\ell)|^p Q_{i,j}\frac{1}{n}\mathbb{1}(\sigma(k)=\ell) \\ &= \frac{1}{n}\min_{Q\in\mathbb{U}} \sum_{i,j} \|C(i,\cdot)-D(j,\cdot)\circ\sigma\|_p^p\, Q_{i,j} \\ &= \frac{1}{n}\mathrm{W}_p^p(\alpha',\beta'),\end{aligned}$$

where

$$\alpha' := \frac{1}{n}\sum_{i=1}^n \delta_{C(i,\cdot)}, \quad \beta' := \frac{1}{n}\sum_{j=1}^n \delta_{D(j,\cdot)\circ\sigma}.$$

B.III.5 Expression of the BCR Between Discrete Measures for 2-Gromov

We now consider the general discrete case where $a \in \triangle_n, b \in \triangle_m$ and $C \in \mathbb{R}^{n\times n}, D \in \mathbb{R}^{m\times m}$, and focus on the case $p := 2$. For $P, Q \in \Pi(a,b)$, we remind the expression of the biconvex cost:

$$J_{\mathrm{BCR}}(P,Q) := \sum_{i,j,k,\ell} (C(i,k)-D(j,\ell))^2 P_{k,\ell}Q_{i,j},$$

where we omit the index sets, we remind that $(i,j,k,\ell) \in [\![1,n]\!] \times [\![1,m]\!] \times [\![1,n]\!] \times [\![1,m]\!]$.

Fixing Q. Fix $Q \in \Pi(a,b)$, we study $\operatorname{argmin}_P J_{\mathrm{BCR}}(P,Q)$. We have:

$$J_{\mathrm{BCR}}(P,Q) = \sum_{k,\ell}\left(\sum_{i,j}(C(i,k)-D(j,\ell))^2 Q_{i,j}\right) P_{k,\ell}. \tag{B.III.8}$$

We re-write the inner term, leveraging the marginal conditions imposed by $Q \in \Pi(a, b)$.

$$\begin{aligned}\sum_{i,j}(C(i,k) - D(j,\ell))^2 Q_{i,j} &= \sum_{i,j}(C(i,k)^2 + D(j,\ell)^2 - 2C(i,k)D(j,\ell))Q_{i,j} \\ &= \sum_i C(i,k)^2 a_i + \sum_j D(j,\ell)^2 b_j - 2\sum_{i,j} C(i,k)D(j,\ell)Q_{i,j} \\ &= \sum_i \left(C(i,k) - \frac{1}{a_i}\sum_j Q_{i,j}D(j,\ell)\right)^2 a_i \\ &\quad + \sum_j D(j,\ell)^2 b_j - \sum_i \frac{1}{a_i}\left(\sum_j Q_{i,j}D(j,\ell)\right)^2.\end{aligned}$$

We introduce

$$K(\ell, Q) := \sum_j D(j,\ell)^2 b_j - \sum_i \frac{1}{a_i}\left(\sum_j Q_{i,j}D(j,\ell)\right)^2,$$

which we isolate because it does not depend on k, and we have shown the following expression:

$$\sum_{i,j}(C(i,k) - D(j,\ell))^2 Q_{i,j} = \sum_i \left(C(i,k) - \frac{1}{a_i}\sum_j Q_{i,j}D(j,\ell)\right)^2 a_i + K(\ell, Q). \tag{B.III.9}$$

Note that one may understand the term

$$\frac{1}{a_i}\sum_j Q_{i,j}D(j,\ell)$$

with a barycentric projection: consider the measure $\pi^{(k,\ell)} \in \mathcal{P}(\mathbb{R}^2)$ defined by $\pi^{(k,\ell)} := \sum_{i,j} Q_{i,j}\delta_{(C(i,k),D(j,\ell))}$. Then the barycentric projection of $\pi^{(k,\ell)}$ at the point $C(i,k)$ is given by:

$$\overline{\pi^{(k,\ell)}} = \frac{1}{a_i}\sum_j Q_{i,j}D(j,\ell).$$

To see $\operatorname{argmin}_P J_{\mathrm{BCR}}(P, Q)$ as a Kantorovich problem, we introduce the matrices $C' \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{n\times m}$:

$$C'(i,j) = \sqrt{a_i}C(i,j),\ E(i,\ell) = \frac{1}{\sqrt{a_i}}\sum_{j=1}^m Q_{i,j}D(j,\ell)$$

Combining Eqs. (B.III.8) and (B.III.9), we obtain

$$J_{\mathrm{BCR}}(P, Q) = \sum_{k,\ell}\|C'(\cdot,k) - E(\cdot,\ell)\|_2^2 P_{k,\ell} + \sum_{k,\ell} K(\ell, Q)P_{k,\ell},$$

and since $K(\ell, Q)$ does not depend on k, we can use the marginal constraint on P:

$$\sum_{k,\ell} K(\ell, Q)P_{k,\ell} = \sum_\ell K(\ell, Q)b_\ell,$$

which does not depend on the variable P, hence

$$\operatorname*{argmin}_{P\in\Pi(a,b)} J_{\mathrm{BCR}}(P, Q) = \operatorname*{argmin}_{P\in\Pi(a,b)} \sum_{k,\ell}\|C'(\cdot,k) - E(\cdot,\ell)\|_2^2 P_{k,\ell}. \tag{B.III.10}$$

We see that the reformulation from Eq. (B.III.10) is a Kantorovich problem with cost matrix $M := (\|C'(\cdot,k) - E(\cdot,\ell)\|_2^2)_{k,\ell}$.

Fixing P. If we now fix $P \in \Pi(a,b)$, we first re-write the complete cost:

$$J_{\mathrm{BCR}}(P,Q) = \sum_{i,j}\left(\sum_{k,\ell}(C(i,k)-D(j,\ell))^2 P_{k,\ell}\right) Q_{i,j}. \tag{B.III.11}$$

Focusing on the inner term, we have:

$$\begin{aligned}\sum_{k,\ell}(C(i,k)-D(j,\ell))^2 P_{k,\ell} &= \sum_{k,\ell}(C(i,k)^2 + D(j,\ell)^2 - 2C(i,k)D(j,\ell))P_{k,\ell}\\ &= \sum_k C(i,k)^2 a_k + \sum_\ell D(j,\ell)^2 b_\ell - 2\sum_{k,\ell} C(i,k)D(j,\ell)P_{k,\ell}\\ &= \sum_k \left(C(i,k) - \frac{1}{a_k}\sum_\ell P_{k,\ell}D(j,\ell)\right)^2 a_k\\ &+ \sum_\ell D(j,\ell)^2 b_\ell - \sum_k \frac{1}{a_k}\left(\sum_\ell P_{k,\ell}D(j,\ell)\right)^2.\end{aligned}$$

Again, we can isolate the term

$$K'(j,P) := \sum_\ell D(j,\ell)^2 b_\ell - \sum_k \frac{1}{a_k}\left(\sum_\ell P_{k,\ell}D(j,\ell)\right)^2,$$

which does not depend on i. We can now re-write the inner term:

$$\sum_{k,\ell}(C(i,k)-D(j,\ell))^2 P_{k,\ell} = \sum_k \left(C(i,k) - \frac{1}{a_k}\sum_\ell P_{k,\ell}D(j,\ell)\right)^2 a_k + K'(j,P). \tag{B.III.12}$$

Again, there is a barycentric projection at play: introducing the coupling $\pi^{(i,j)} := \sum_{k,\ell} P_{k,\ell}\delta_{(C(i,k),D(j,\ell))}$, we have:

$$\overline{\pi^{(i,j)}}(C(i,k)) = \frac{1}{a_k}\sum_\ell P_{k,\ell}D(j,\ell).$$

Similarly to the previous case, we introduce $C'' \in \mathbb{R}^{n\times n}$ and $F \in \mathbb{R}^{n\times m}$ and verifying:

$$C''(i,k) = \sqrt{a_k}C(i,k),\ F(k,j) = \frac{1}{\sqrt{a_k}}\sum_\ell P_{k,\ell}D(j,\ell),$$

and combine Eqs. (B.III.11) and (B.III.12):

$$J_{\mathrm{BCR}}(P,Q) = \sum_{i,j}\|C''(i,\cdot) - F(j,\cdot)\|_2^2 Q_{i,j} + \sum_{i,j} K'(j,P)Q_{i,j},$$

and since $K'(j,P)$ does not depend on i, we can use the marginal constraint on Q:

$$\sum_{i,j} K'(j,P)Q_{i,j} = \sum_j K'(j,P)b_j,$$

which does not depend on the variable Q, and we obtain the following Kantorovich reformulation:

$$\operatorname*{argmin}_{Q\in\Pi(a,b)} J_{\mathrm{BCR}}(P,Q) = \operatorname*{argmin}_{Q\in\Pi(a,b)} \sum_{i,j}\|C''(i,\cdot) - F(j,\cdot)\|_2^2 Q_{i,j}, \tag{B.III.13}$$

B.III.6 Heuristic of Sliced Permutations for the BCR Between Point Clouds ....

In this section, we propose the heuristic of replacing the permutation obtained from solving W_p^p in Section B.III.4 between the lines or columns of the cost matrices C and D with a sliced permutation.

For $Q = Q^\tau$ fixed, we have seen in Section B.III.4 that the optimisation over P writes:

$$\min_{P\in\mathbb{U}} J_{\mathrm{BCR}}(P, Q^\tau) = \frac{1}{n}\min_{P\in\mathbb{U}} \sum_{(k,\ell)\in[\![1,n]\!]^2} \|C(\cdot,k) - D(\cdot,\ell)\circ\tau\|_p^p P_{k,\ell},$$

which we propose to replace with a discretised min-Pivot-Sliced plan (a.k.a. SWGG [Mah+23], we disregard potential projection ambiguities):

$$\min_{P\in\mathbb{U}} J_{\mathrm{BCR}}(P, Q^\tau) \approx \frac{1}{n^2}\min_{u\in[\![1,U]\!]} \sum_{k=1}^n \|C(\cdot,k) - D(\cdot,\sigma_u(k))\circ\tau\|_p^p,$$

where $\sigma_u = \psi_u\circ\varphi_u^{-1}$, with $\varphi_u \in \mathfrak{S}_n$ sorting $(\theta_u^\top C(\cdot,k))_{k=1}^n$ and $\psi_u \in \mathfrak{S}_n$ sorting $(\theta_u^\top D(\cdot,\ell)\circ\tau)_{\ell=1}^n$. To recap, for $Q = Q^\tau$ fixed, the lines of D are permuted with τ, and we compare the distributions of the columns of C and D (whose lines are permuted) using the min-Pivot-Sliced discrepancy, which yields a plan P induced by a permutation σ_u. The approximation is an upper-bound since the optimisation is done over a subset of all permutations.

For $P = P^\sigma$ fixed, by Section B.III.4, we have:

$$\min_{Q\in\mathbb{U}} J_{\mathrm{BCR}}(P^\sigma, Q) = \frac{1}{n}\min_{Q\in\mathbb{U}} \sum_{i,j} \|C(i,\cdot) - D(j,\cdot)\circ\sigma\|_p^p Q_{i,j},$$

which again we replace with a min-Sliced permutation:

$$\min_{Q\in\mathbb{U}} J_{\mathrm{BCR}}(P^\sigma, Q) \approx \frac{1}{n^2}\min_{u\in[\![1,U]\!]} \sum_{i=1}^n \|C(i,\cdot) - D(\tau_u(i),\cdot)\circ\sigma\|_p^p,$$

where $\tau_u = \psi_u\circ\varphi_u^{-1}$, with $\varphi_u \in \mathfrak{S}_n$ sorting $(\theta_u^\top C(i,\cdot))_{i=1}^n$ and $\psi_u \in \mathfrak{S}_n$ sorting $(\theta_u^\top D(j,\cdot)\circ\sigma)_{j=1}^n$. This time, for $P = P^\sigma$ fixed, the columns of D are permuted with σ, and we compare the distributions of the lines of C and D (whose columns are permuted) using the min-Sliced, which yields a plan Q induced by a permutation τ_u.

This heuristic leads to the alternate minimisation scheme presented in Algorithm B.III.2 which computes a permutation σ that is a heuristic solution of the BCR problem Eq. (B.III.6).

Algorithm B.III.2: Sliced Gromov-Wasserstein Permutation

Input: Cost matrices $C, D \in \mathbb{R}^{n\times n}$, number of projections U, number of iterations T.

Initialisation: $\tau \leftarrow I_n$;
for $t \in [\![1,T]\!]$ **do**
  for $u \in [\![1,U]\!]$ **do**
    Sample $\theta_u \sim \mathcal{U}(\mathbb{S}^{d-1})$;
    $\varphi_u \leftarrow \mathrm{argsort}\left[(\theta_u^\top C(\cdot,k))_{k=1}^n\right]$;
    $\psi_u \leftarrow \mathrm{argsort}\left[(\theta_u^\top D(\cdot,\ell)\circ\tau)_{\ell=1}^n\right]$;
    $\sigma_u \leftarrow \psi_u\circ\varphi_u^{-1}$;
  $\sigma \leftarrow \underset{\sigma_u,\, u\in[\![1,U]\!]}{\mathrm{argmin}} \sum_{k=1}^n \|C(\cdot,k) - D(\cdot,\sigma_u(k))\circ\tau\|_p^p$;
  for $u \in [\![1,U]\!]$ **do**
    Sample $\theta_u \sim \mathcal{U}(\mathbb{S}^{d-1})$;
    $\varphi_u \leftarrow \mathrm{argsort}\left[(\theta_u^\top C(i,\cdot))_{i=1}^n\right]$;
    $\psi_u \leftarrow \mathrm{argsort}\left[(\theta_u^\top D(j,\cdot)\circ\sigma)_{j=1}^n\right]$;
    $\tau_u \leftarrow \psi_u\circ\varphi_u^{-1}$;
  $\tau \leftarrow \underset{\tau_u,\, u\in[\![1,U]\!]}{\mathrm{argmin}} \sum_{i=1}^n \|C(i,\cdot) - D(\tau_u(i),\cdot)\circ\sigma\|_p^p$;
return τ.

From a practical standpoint, one may also check for a convergence criterion such as $\sigma = \tau$. Once the algorithm has returned a permutation σ, the associated GW cost is:

$$\mathrm{G}_p^p(C, D) \approx \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} |C(i,k) - D(\sigma(i), \sigma(k))|^p. \tag{B.III.14}$$

The min-Sliced Gromov Wasserstein cost is formally an upper bound of the Gromov-Monge problem [MN22]:

$$\mathrm{GM}_p^p(C, D) := \min_{\sigma \in \mathfrak{S}_n} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} |C(i,k) - D(\sigma(i), \sigma(k))|^p. \tag{B.III.15}$$

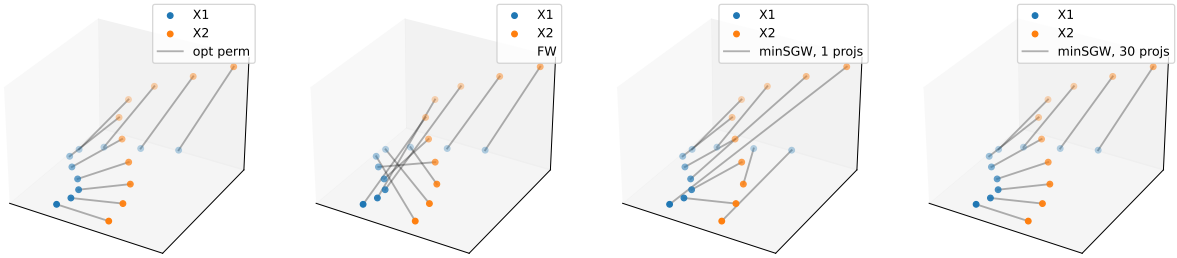
It is currently an active area of research to find a correspondence between the full Gromov-Wasserstein problem and its Monge formulation (constraining the coupling to be a permutation, or more generally a map) [ML18; Vay+19; BHS23a; DLV24].

From a complexity standpoint, the cost of computing projections is $\mathcal{O}(Un^2)$, and the cost of sorting is $\mathcal{O}(Un\log(n))$. Computing the costs to determine σ and τ is in $\mathcal{O}(Un^2)$, and this yields a total complexity of $\mathcal{O}(TUn^2)$. As a heuristic, this method is only competitive compared to Frank-Wolfe (Algorithm B.III.1) for $U \ll n$, thus in experiments we will illustrate the case $U := n/10$. It is worth emphasising that all computations from Algorithm B.III.2 can be performed in parallel on GPU (the loops over u are written this way for legibility), with a memory footprint of $\mathcal{O}(n^2 + Un)$.

We consider a toy experimental setting where the cost matrices C and D arise from points clouds $(x_1, \cdots, x_n) \in \mathbb{R}^{n\times 2}$ and $(y_1, \cdots, y_n) \in \mathbb{R}^{n\times 3}$ respectively, with the following expressions:

$$\forall k \in [\![1, n]\!],\ x_k := (t_k \cos(t_k), t_k),\ y_k := (t_k \cos(t_k), t_k, t_k),\ t_k := \tfrac{2(k-1)\pi}{n}.$$

The cost matrices are then given by $C(i,k) := \|x_i - x_k\|_2^2$ and $D(j,\ell) := \|y_j - y_\ell\|_1$. We consider the "ground truth" permutation to be the identity permutation, and the cost matrices are permuted with a random permutation before being passed to the algorithms. We illustrate this setting and the results of Algorithm B.III.2 and of the state-of-the-art Frank Wolfe algorithm in Fig. B.III.1.



(a) Ground truth permutation. (b) Frank-Wolfe plan (a permutation here). (c) minSGW permutation with $U = 1$. (d) minSGW permutation with $U = 30$.

Figure B.III.1: Illustration of the toy example with $n = 10$.

In Fig. B.III.2, we compare minSGW and FW on the same toy example for varying n and $U = \frac{n}{10}$. Unfortunately, minSGW does not seem to be competitive, but may have some value as a coarse heuristic.

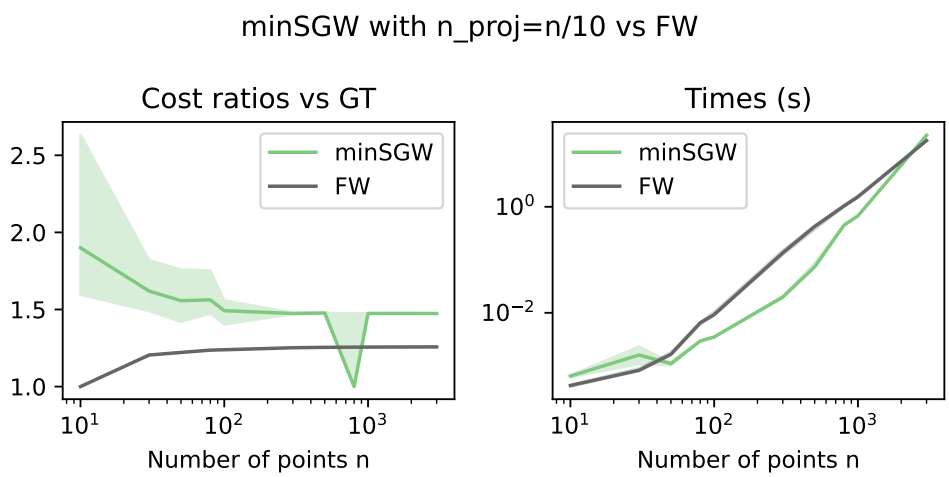


Figure B.III.2: Comparison of minSGW and FW on the toy example for varying n and $U = n/10$. On the left, we compare the cost minSGW and FW by comparing their ratios with the ground truth cost (which are larger than 1 due to being sub-optimal). On the right, we compare computation times (performed on GPU with PyTorch).

While we believe that our GW heuristic has some promise, our implementation efforts or not yet conclusive.

Part C

Optimal Transport Barycentres

Chapter C.I studies numerical resolution methods for the Generalised Wasserstein Barycentre problem [DGS21], and studies an additional extension of this problem. This paper is based on joint work with Julie Delon and Rémi Flamary.

Chapter C.II introduces a general notion of barycentres for generic transport costs, with a focus on a fixed-point method loosed grounded on linearisation principles. This chapter is based on the paper:

[TDG24] Eloi Tanguy, Julie Delon and Nathaël Gozlan.
"Computing Barycentres of Measures for Generic Transport Costs".
arxiv preprint 2501.04016 (Dec. 2024).

C.I

(Blind) Generalised Wasserstein Barycentres

C.I.1 Introduction . . . 263
C.I.1.1 Motivation and Purpose 263
C.I.1.2 Outline and Contributions 265
C.I.2 The Generalised Wasserstein Barycentre Problem . . . 266
C.I.2.1 Discrete Formulation 266
C.I.2.2 Free-Support Algorithm using the Change-of-Variables 266
C.I.2.3 Lagrangian Energy Minimisation 267
C.I.2.4 (Stochastic) Gradient Descent Algorithm 268
C.I.2.5 Block Coordinate Descent Algorithm 270
C.I.2.6 Numerical Illustration 270
C.I.3 The Blind GWB Problem . . . 273
C.I.3.1 Problem Statement and Existence of a Solution 273
C.I.3.2 Lagrangian Energy Minimisation 274
C.I.3.3 (Stochastic) Gradient Descent Algorithm 274
C.I.3.4 Block Coordinate Descent Algorithm 276
C.I.3.5 Numerical Illustration on a Reconstruction Problem 277

Abstract

Generalised Wasserstein Barycentre (GWB) [DGS21] is a problem that generalises the Wasserstein Barycentre problem by allowing the barycentre and target measures to be in different Euclidean spaces. The comparison is performed using linear maps from the barycentre space to the space of each target measure. In this chapter, we introduce three natural numeric methods to solve the GWB problem from a Lagrangian perspective, optimising the support of a measure of fixed weight and support size. We also present an extension of the GWB problem called Blind GWB (BGWB), which allows for the optimisation of the linear maps in the GWB problem. We study some elementary theoretical properties of the problem and adapt the numerical methods to this new problem.

This chapter is based on joint work with Julie Delon and Rémi Flamary.

C.I.1 Introduction

C.I.1.1 Motivation and Purpose

Optimal Transport (OT) is a powerful tool which lifts a notion of distance (or, more generally, a cost) between two points in a space to a notion of distance (or cost) between probability measures on that space. On the Euclidean space $\mathbb{R}^d$ equipped with the cost $c(x, y) = \|x - y\|_2^2$,

the 2-Wasserstein distance compares two probability measures μ and ν with a finite second moment by computing the expected cost of the least costly coupling between them:

$$\mathrm{W}_2^2(\mu,\nu) = \inf_{\substack{X\sim\mu\\ Y\sim\nu}} \mathbb{E}\left[\|X-Y\|_2^2\right] = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y), \tag{C.I.1}$$

where $\Pi(\mu,\nu)$ is the set of probability measures π on $\mathbb{R}^{2d}$ such that their marginals are μ and ν, respectively. For a complete presentation of the properties of the Wasserstein distance, we refer to the monographs [Vil09; San15]. Writing $\mathcal{P}_2(\mathbb{R}^d)$ as the set of probability measures on $\mathbb{R}^d$ with finite second moment, the 2-Wasserstein distance is a metric on $\mathcal{P}_2(\mathbb{R}^d)$. Taking inspiration from the natural mean in $\mathbb{R}^d$, [AC11] use the 2-Wasserstein metric to lift the notion of barycentre in $\mathbb{R}^d$ to the space $\mathcal{P}_2(\mathbb{R}^d)$. The idea is to start from the variational expression of Euclidean means in $\mathbb{R}^d$: given weights (λ_k) in the simplex $\triangle_K$ and points $(y_k)_{k=1}^K$ in $\mathbb{R}^d$, we have:

$$\sum_{k=1}^K \lambda_k y_k = \operatorname*{argmin}_{x\in\mathbb{R}^d} \sum_{k=1}^K \lambda_k \|y_k - x\|_2^2.$$

This expression is known as the Fréchet mean of the points $(y_k)_{k=1}^K$ with weights $(\lambda_k)_{k=1}^K$ for the cost $c(x,y) = \|x-y\|_2^2$. The idea of [AC11] is to perform the same lifting method as for the Wasserstein distance, definition 2-Wasserstein barycentres as solutions of the following minimisation problem, given measures $(\nu_k)_{k=1}^K$ in $\mathcal{P}_2(\mathbb{R}^d)$ and weights $(\lambda_k)_{k=1}^K \in \triangle_K$:

$$\operatorname*{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathop{\mathbb{E}}_{\substack{X\sim\mu\\ Y_1\sim\nu_k,\cdots,Y_K\sim\nu_K}} \left[\sum_{k=1}^K \lambda_k \|X-Y_k\|_2^2\right] = \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\mu,\nu_k).$$

Existence of a solution to this problem is guaranteed by [AC11], although unlike the Euclidean case, uniqueness does not hold in general. Note that the theoretical foundations for the work [AC11] are laid in the more general framework of [CE10], which studies a similar problem with more general costs than $c(x,y) = \|x-y\|_2^2$. In Fig. C.I.1, we illustrate the 2-Wasserstein barycentre of two measures ν_1 and ν_2 in $\mathcal{P}_2(\mathbb{R}^2)$ with equal weights $\lambda_1 = \lambda_2 = 1/2$. The considered measures are empirical measures, and the barycentre resolution is done in Python using the `POT` library [Fla+21].

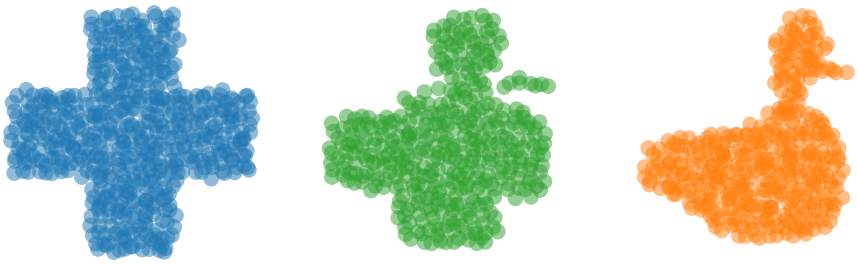


Figure C.I.1: 2-Wasserstein barycentre of the "cross" empirical measure on the left and of the "duck" empirical measure on the right.

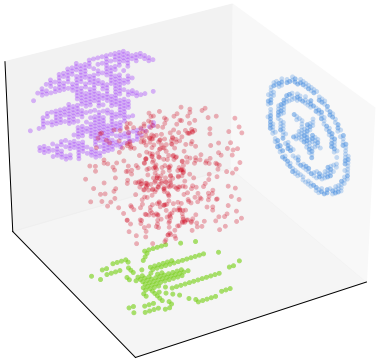
The notion of 2-Wasserstein barycentres was extended in [DGS21] to measures ν_k on different Euclidean spaces $\mathbb{R}^{d_k}$. To understand this generalisation, we begin with a notion of barycentres between points $y_k \in \mathbb{R}^{d_k}$: we seek a barycentre points $x \in \mathbb{R}^d$ whose projections $P_k x \in \mathbb{R}^{d_k}$ are as close as possible to the each respective y_k. The maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$ are taken as linear maps (not necessarily projections, although this intuition is helpful) and we ask for the matrix $A := \sum_{k=1}^K \lambda_k P_k^T P_k \in S_d^+(\mathbb{R})$ to be invertible. There is then a unique Fréchet mean for the costs $c_k(x,y) := \|P_k x - y\|_2^2$, which is our notion of barycentre between the (y_k):

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} \sum_{k=1}^K \lambda_k \|P_k x - y_k\|_2^2 = A^{-1} \sum_{k=1}^K P_k^T x_k.$$

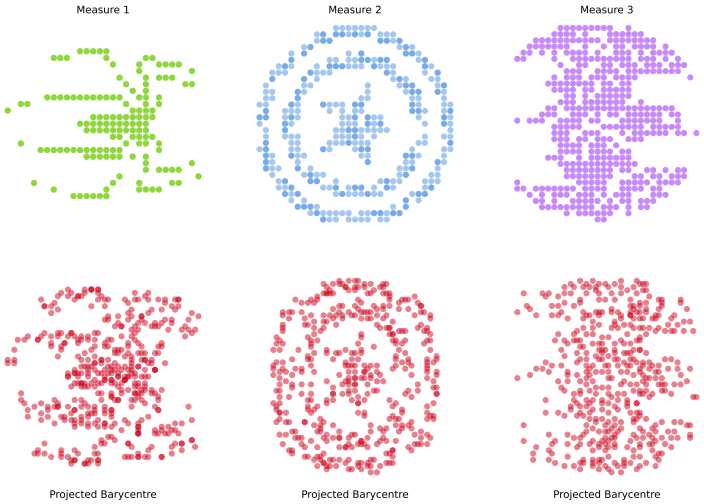
Following the same lifting method as for the 2-Wasserstein barycentre, [DGS21] define the barycentre of measures ν_k in $\mathcal{P}_2(\mathbb{R}^{d_k})$ with weights $(\lambda_k)_{k=1}^K$ and linear maps $P_k \in \mathbb{R}^{d_k \times d}$ as:

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{k=1}^{K} \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k), \tag{C.I.2}$$

where we confound the linear map P_k with its $d_k \times d$ matrix in the canonical basis, and where $P_k \# \mu$ is the pushforward of μ by P_k, i.e. for any Borel set $B \subset \mathbb{R}^{d_k}$, we have: $P_k \# \mu(B) = \mu(P_k^{-1}(B))$. The problem in Eq. (C.I.2) is called the Generalised Wasserstein Barycentre (GWB) problem, and the solutions are called GWB barycentres. In Fig. C.I.2, we illustrate a GWB barycentre of three two-dimensional measures.



(a) 3D representation of the GWB solution.



(b) Projections of the GWB solution.

Figure C.I.2: Results of a GWB which determines a barycentre in $\mathbb{R}^3$ whose projections match three given discrete measures in $\mathbb{R}^2$. The target measures are represented in blue, green and purple, while the barycentre is represented in red. On the left, we observe the barycentre in $\mathbb{R}^3$ along with embeddings of the two-dimensional measures. On the right, for each of the three target measures ν_k, we compare the target measure with the projection $P_k \# \mu$ of the barycentre μ in $\mathbb{R}^3$.

The paper [DGS21] provides a theoretical framework for the GWB problem as well as a reformulation into an ordinary 2-Wasserstein Barycentre problem, which allows the use of existing algorithms for numerical resolution. In this chapter, we recall the application of this reformulation for a numerical resolution with the "free-support" algorithm of [CD14], and present three additional algorithms for the GWB problem. We also present an extension of the GWB called the Blind GWB (BGWB), which allows for the optimisation of the linear maps P_k in the GWB problem:

$$\underset{\substack{\mu \in \mathcal{P}_2(\mathbb{R}^d) \\ P_1 \in \mathbb{R}^{d_1 \times d} \dots P_K \in \mathbb{R}^{d_K \times d}}}{\operatorname{argmin}} \sum_{k=1}^{K} \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k). \tag{C.I.3}$$

C.I.1.2 Outline and Contributions

In Section C.I.2.2, we recall the GWB problem and its reformulation as a 2-Wasserstein Barycentre problem. We also present the "free-support" algorithm of [CD14] for the numerical resolution of the GWB problem. This method looks for a barycentre measure of the from $\mu = \gamma_X := \sum_{i=1}^n a_i \delta_{x_i}$, imposing the number of points n in particular. While [CD14] include optimisation of the weights $a \in \triangle_n$ through projected sub-gradient descent, like in the `POT` implementation [Fla+21], we fix the weights a, taking a Lagrangian approach to the numerical resolution, as presented in Section C.I.2.3.

We introduce natural Gradient Descent (GD) and Stochastic Gradient Descent (SGD) methods for the GWB in Section C.I.2.4, including some convergence results. In Section C.I.2.5, we present a Block Coordinate Descent (BCD) method for the GWB problem. Through a numerical illustration in Section C.I.2.6, we illustrate the GWB problem and its resolution with GD, SGD and BCD.

In Section C.I.3, we present the BGWB problem which generalises GWB by optimising over the linear maps P_k in addition to the barycentre measure μ. After a brief theoretical study in Section C.I.3.1, we adapt the GD, SGD and BCD methods to the BGWB problem in Sections C.I.3.3 and C.I.3.4. We illustrate the BGWB problem and its resolution with GD, SGD and BCD in a toy example in Section C.I.3.5.

C.I.2 The Generalised Wasserstein Barycentre Problem

C.I.2.1 Discrete Formulation

As introduced by [DGS21], the Generalised Wasserstein Barycentre (GWB) consists in finding a barycentre in $\mathcal{P}_2(\mathbb{R}^d)$ between measure $\nu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$ using linear maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ for $k \in [\![1, K]\!]$. The problem writes as follows in general, given weights $(\lambda_k) \in \triangle_K$:

$$\operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^{K} \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k) =: F_{\mathrm{GWB}}(\mu). \tag{C.I.4}$$

The problem has solutions in $\mathcal{P}_2(\mathbb{R}^d)$, as shown in [DGS21, Proposition 3.2]. In this chapter, we will focus on the case where the measures (ν_k) are discrete, and we write them as:

$$\forall k \in [\![1, K]\!],\ \nu_k = \sum_{i=1}^{n_k} b_{k,i} \delta_{y_{k,i}},\ b_k := b_{k,\cdot} \in \triangle_{n_k},\ Y_k := y_{k,\cdot} \in \mathbb{R}^{n_k \times d_k},$$

where $b_{k,\cdot}$ denotes the vector $(b_{k,i})_{i=1}^{n_k}$ and $y_{k,\cdot}$ the matrix $(y_{k,i})_{i=1}^{n_k}$ with the $y_{k,i} \in \mathbb{R}^{d_k}$ stacked line-by-line Throughout this section, we will make the assumption that the linear maps P_k cover the whole space $\mathbb{R}^d$, which we state in Assumption C.I.1.

> **Assumption C.I.1.** The linear maps P_k are such that Span $A = \mathbb{R}^d$, where $A := \sum_{k=1}^{K} \lambda_k P_k P_k^\top$.

Note that the assumption Assumption C.I.1 requires the intuitive condition $\sum_k d_k \geq d$, which allows the linear maps P_k to collectively see the whole space $\mathbb{R}^d$. As show in in [DGS21], Assumption C.I.1 is equivalent to the condition $\cap_k \mathrm{Ker} P_k = \{0\}$.

In the particular case where the measures ν_k are themselves images of a measure $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ by the maps P_k, the problem can be seen as a reconstruction problem, aiming to find the set of measures μ that satisfy the constraints $P_k \# \mu = P_k \# \mu_0$. We have seen this problem in Chapter A.I, and have shown in Theorem A.I.2 that if $\sum_k d_k > d$, then for almost-every linear maps (P_k) (whose lines are each independently drawn from a distribution with density over $\mathbb{R}^{d_k}$), there is a unique solution to the reconstruction GWB problem, which is the measure μ_0 itself.

C.I.2.2 Free-Support Algorithm using the Change-of-Variables

In [DGS21], the authors provide a reformulation of the GWB problem into a usual 2-Wasserstein Barycentre problem. The idea is to cast the problem in $\mathcal{P}_2(\mathbb{R}^d)$ by considering the measures $\tilde{\nu}_k := A^{-1/2} P_k^\top \# \nu_k$, and minimising the following functional:

$$G_{\mathrm{GWB}} := \tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d) \longmapsto \sum_{k=1}^{K} \lambda_k \mathrm{W}_2^2(\tilde{\mu}, \tilde{\nu}_k).$$

The problem of minimising G_{GWB} over $\mathcal{P}_2(\mathbb{R}^d)$ is a 2-Wasserstein Barycentre which has solutions thanks to [AC11]. The problems of minimising F_{GWB} and G_{GWB} are equivalent, since by

[DGS21, Proposition 3.1], there exists a constant $C \in \mathbb{R}$ such that $F_{\mathrm{GWB}}(\mu) = G_{\mathrm{GWB}}(A^{1/2}\#\mu) + C$. As a result, given a solution $\tilde{\mu}^*$ minimising G, the measure $\mu^* := A^{-1/2}\#\tilde{\mu}^*$ is a solution of the GWB problem. In particular, this reformulation implies that in our discrete setting, there exists a solution μ^* of the GWB problem that is discrete with a support of size $n \leq \sum_k n_k - K + 1$, by [ABM16, Theorem 2].

A state-of-the-art algorithm for solving the discrete 2-Wasserstein Barycentre problem is the so-called "free-support" method [CD14], which is implemented in the Python library `POT` [Fla+21] without optimisation over the weights of the barycentre measure. This method imposes the size n of the support of the candidate barycentre measure, as well as its weights, which in practice are often set to be uniform. The name "free-support" is by opposition to Eulerian methods which fix a grid of points in $\mathbb{R}^d$ and optimise the weights of the barycentre. The "free-support" method optimises over the barycentre positions $X \in \mathbb{R}^{n\times d}$ by fixed-point iterations, which can be seen as a Lagrangian paradigm. In Algorithm C.I.1 we introduce the use of the change-of-variables from [DGS21] and the free-support algorithm to solve the GWB problem.

Algorithm C.I.1: Free-support solver for GWB.

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and linear maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$. Number of iterations T and weights $a \in \triangle_n$.

Result: Barycentre $\mu_T = \sum_{i=1}^n a_i \delta_{A^{-1/2} u_i^{(T)}}$.

1 **Initialisation:** Compute $V_k := Y_k P_k A^{-1/2} \in \mathbb{R}^{n_k \times d}$ for $k \in [\![1, K]\!]$.
2 Choose $\tilde{\mu}_0 = \sum_{i=1}^n a_i \delta_{u_i^{(0)}}$ with $U^{(0)} \in \mathbb{R}^{n\times d}$.
3 **for** $t \in [\![0, T-1]\!]$ **do**
4 **for** $k \in [\![1, K]\!]$ **do**
5 Solve the OT problem: $\pi^{(k)} \in \underset{\pi \in \Pi(a, b_k)}{\operatorname{argmin}} \sum_{i,j} \pi_{i,j} \|u_i^{(t)} - v_{k,j}\|_2^2$;
6 **end**
7 **for** $i \in [\![1, n]\!]$ **do**
8 Compute $u_i^{(t+1)} = \sum_{k=1}^K \frac{\lambda_k}{a_i} \sum_{j=1}^{n_k} \pi_{i,j}^{(k)} v_{k,j}$;
9 **end**
10 **end**

The computation at Line 8 of Algorithm C.I.1 states that the position of the i-th point of the barycentre at iteration $t+1$ is the mean (with weights (λ_k)) of the barycentric projections $\left(\overline{\pi^{(k)}}(u_i^{(t)})\right)_k$, where $\pi^{(k)}$ is an OT plan between $\tilde{\mu}^{(t)}$ and $\tilde{\nu}_k$. We have written the algorithm such that a fixed number of iterations T is performed, but in practice the algorithm is run until convergence. A possible stopping criterion is based on the relative change of the barycentre measure, and convergence is considered attained when the following condition is satisfied:

$$\mathrm{W}_2^2(\tilde{\mu}_{t+1}, \tilde{\mu}_t) < \alpha \|X^{(t)}\|_2^2,$$

for some criterion $\alpha > 0$. We have contributed this method to the Python library `POT`.

C.I.2.3 Lagrangian Energy Minimisation

In the following, we will consider two additional algorithms for solving the GWB problem, which are based on the minimisation of the functional F in Eq. (C.I.4) with respect to the barycentre positions $X \in \mathbb{R}^{n\times d}$, leading to the following constrained GWB problem:

$$\underset{X \in \mathbb{R}^{n\times d}}{\operatorname{argmin}} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(P_k\#\gamma_X, \nu_k) =: \mathcal{E}_{\mathrm{GWB}}(X), \tag{C.I.5}$$

where $\gamma_X := \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}_2(\mathbb{R}^d)$, for a fixed weight vector $a \in \triangle_n$.

In the particular case where the weights are uniform (i.e. $a_i = \frac{1}{n}$, $\lambda_k = \frac{1}{K}$) and the measures ν_k are one-dimensional ($d_k = 1$) projections of a measure $\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j} \in \mathcal{P}_2(\mathbb{R}^d)$, i.e. $\nu_k = \theta_k^\top \# \nu$ with $\theta_k \in \mathbb{S}^{d-1}$, we can consider the linear maps $P_k := P_\theta := \theta^\top$ (confounding linear maps and matrices by slight abuse of notation). The energy function $\mathcal{E}_{\mathrm{GWB}}$ is then equal to a Monte-Carlo approximation of the Sliced Wasserstein (SW) distance:

$$\mathrm{SW}_2^2(\gamma_X, \nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2(P_\theta \# \gamma_X, P_\theta \# \nu) \mathrm{d}\varpi(\theta) \approx \frac{1}{K}\sum_{k=1}^{K} \mathrm{W}_2^2(P_{\theta_k} \# \gamma_X, P_{\theta_k} \# \nu) = \mathcal{E}_{\mathrm{GWB}}(X).$$

The energy $\mathcal{E}_{\mathrm{GWB}}$ in this setting and its correspondence to the true SW distance $\mathrm{SW}_2^2(\gamma_X, \nu)$ is the focus of Chapter A.II. Taking inspiration from the cell decomposition principle from Section A.II.2.3, we can also write the energy $\mathcal{E}_{\mathrm{GWB}}$ as a minimum of quadratic expressions in the general case:

$$\mathcal{E}_{\mathrm{GWB}}(X) = \min_{\pi_1 \in \mathrm{Extr}\,\Pi(a,b_1), \cdots, \pi_K \in \mathrm{Extr}\,\Pi(a,b_K)} \sum_{k=1}^{K} \lambda_K \sum_{i=1}^{n}\sum_{j=1}^{n_k} \pi_{i,j}^{(k)} \|P_k x_i - y_{k,j}\|_2^2, \tag{C.I.6}$$

where $\mathrm{Extr}\,\Pi(a, b_k)$ is the (finite) set of extreme points of the compact polytope $\Pi(a, b_k)$. We conclude as in Proposition A.II.5 that $\mathcal{E}_{\mathrm{GWB}}$ is locally Lipschitz continuous, and semi-algebraic which is to say that it is piecewise polynomial with pieces defined by polynomial equations and inequalities.

Proposition C.I.1. The functional $\mathcal{E}_{\mathrm{GWB}}$ is semi-algebraic and locally Lipschitz on $\mathbb{R}^{n\times d}$.

C.I.2.4 (Stochastic) Gradient Descent Algorithm

A natural approach to minimising the functional $\mathcal{E}_{\mathrm{GWB}}$ is to use Gradient Descent (GD), or Stochastic Gradient Descent (SGD) with a random index $k \in [\![1, K]\!]$ at each iteration. We have seen in Proposition C.I.1 that $\mathcal{E}_{\mathrm{GWB}}$ is semi-algebraic and locally Lipschitz, which implies that it is differentiable almost-everywhere. Thanks to the sub-differentiability properties of the discrete Kantorovich problem (Proposition B.I.8), the following expression defines a (Clarke) sub-gradient of the functional $\mathcal{E}_{\mathrm{GWB}}$ at $X \in \mathbb{R}^{n\times d}$:

$$\begin{aligned} \varphi_{\mathcal{E}_{\mathrm{GWB}}}(X) &:= \sum_{k=1}^{K} \lambda_k \varphi_k(X), \\ \varphi_k(X) &:= 2(\mathrm{diag}(a) X A - \pi^{(k)} Y_k P_k) \in \mathbb{R}^{n\times d}, \end{aligned} \tag{C.I.7}$$

where we write $\mathrm{diag}(a)$ the diagonal matrix with a_i on the diagonal. The sub-gradient $\varphi_{\mathcal{E}_{\mathrm{GWB}}}(X)$ depends on a choice of the OT plans $\pi^{(k)} \in \Pi^*(a, b_k)$ for $k \in [\![1, K]\!]$. This choice can be done so as to ensure that the functions φ_k are semi-algebraic (for example with the lexicographic order to select within the set of optimal extreme points of each OT problem). In practice, one may conveniently automatic differentiation with Pytorch [Pas+19] to compute a sub-gradient $\mathcal{E}_{\mathrm{GWB}}$, thanks to the automatic sub-gradient computation of the OT cost with `ot.emd2` in the `POT` package [Fla+21]. We begin by formalising the GD algorithm in Algorithm C.I.2.

Algorithm C.I.2: GWB resolution with Gradient Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and linear maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$. Number of iterations T, weights $a \in \triangle_n$, learning rates (α_t).

Result: Barycentre positions $X \in \mathbb{R}^{n\times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n\times d}$;

for $t \in [\![1, T]\!]$ **do**

  Step the positions: $X_{t+1} \leftarrow X_t - \alpha_t \, \varphi_{\mathcal{E}_{\mathrm{GWB}}}(X_t)$;

end

As in Proposition B.I.10, the regularity of $\mathcal{E}_{\mathrm{GWB}}$ proved in Proposition C.I.1 provides convergence of Algorithm C.I.2 to a (Clarke) critical point of the functional $\mathcal{E}_{\mathrm{GWB}}$. We remind that for a locally Lipschitz function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, its Clarke sub-differential at $x \in \mathbb{R}^d$ is the convex hull of the limits of the gradients of f:

$$\partial_C f(x) = \mathrm{conv}\left\{ \lim_{t \longrightarrow +\infty} \nabla f(x_t) : x_t \xrightarrow[t \longrightarrow +\infty]{} x,\ x_t \in D_f \right\},$$

where D_f is the set of differentiability of f and conv denotes the convex envelope. The set of Clarke critical points of f is the set of points $x \in \mathbb{R}^d$ such that $0 \in \partial_C f(x)$.

Proposition C.I.2. Convergence of GD for GWB, appplication of [EN97, Theorem 4.1] Consider learning rates (α_t) such that $\alpha_t \to 0$ and $\sum_t \alpha_t = +\infty$. Then the iterates (X_t) of Algorithm C.I.2 are bounded and sub-sequentially converge to Clarke critical points of $\mathcal{E}_{\mathrm{GWB}}$.

Proof. By Proposition C.I.1 and Proposition B.I.8, we can apply [EN97, Theorem 4.1], any accumulation point of the iterates (X_t) is a Clarke critical point of the functional $\mathcal{E}_{\mathrm{GWB}}$, and the sequence $(\mathcal{E}_{\mathrm{GWB}}(X_t))$ converges. By coercivity of $\mathcal{E}_{\mathrm{GWB}}$, the sequence (X_t) is bounded, and subsequential limits are necessarily accumulation points, thus Clarke critical. □

We now turn our attention to the Stochastic Gradient Descent (SGD) algorithm, which consists in taking a random index $k \in [\![1, K]\!]$ with a law $\mathbb{D}_\lambda$ on $[\![1, K]\!]$ at each iteration. The probability measure $\mathbb{D}_\lambda \in \mathcal{P}([\![1, K]\!])$ is defined by $\mathbb{D}_\lambda(\{k\}) = \lambda_k$ for $k \in [\![1, K]\!]$. We state the SGD algorithm in Algorithm C.I.3.

Algorithm C.I.3: GWB resolution with Stochastic Gradient Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and linear maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$. Number of iterations T, weights $a \in \triangle_n$, learning rates (α_t).

Result: Barycentre positions $X \in \mathbb{R}^{n \times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n \times d}$;

for $t \in [\![1, T]\!]$ **do**

  Draw $k \sim \mathbb{D}_\lambda$ and step the positions: $X_{t+1} \leftarrow X_t - \alpha_t\ \varphi_k(X_t)$;

end

As in Section A.II.4.5, we can show sub-sequential convergence of the SGD algorithm to a Clarke critical point of the functional $\mathcal{E}_{\mathrm{GWB}}$, under the assumption of boundedness of iterates. The proof is very similar to that of Theorem A.II.10.

Proposition C.I.3. Convergence of SGD for GWB, application of [Dav+20, Theorem 4.2] Consider learning rates $(\alpha_t) \in \mathbb{R}_+^{\mathbb{N}}$ such that $\sum_t \alpha_t = +\infty$ and $\sum_t \alpha_t^2 < +\infty$. Take iterates (X_t) of Algorithm C.I.3, and assume that they are almost-surely bounded. Then almost-surely, any sub-sequential limit of (X_t) is a Clarke critical point of $\mathcal{E}_{\mathrm{GWB}}$.

Proof. To apply [Dav+20, Theorem 4.2], we need to verify their assumptions C.1, C.2, C.3, D.1 and D.2. The assumptions C.1 (on the learning rates) and C.2 (on the boundedness of the iterates) are assumed in the statement. For C.3, we simply take no external noise (which fits in their framework). For D.1, Proposition B.I.8 shows that the set of critical values of $\mathcal{E}_{\mathrm{GWB}}$ is finite, thus its complementary is dense. As for D.2, we use [Dav+20, Lemma 5.2] with the fact that $\mathcal{E}_{\mathrm{GWB}}$ is semi-algebraic and locally Lipschitz, implying path differentiability by [BP21, Proposition 2]. □

While we do not discuss optimisation over the weights a of the barycentre, this can be done by projected Gradient Descent on the probability simplex, as in [CD14].

C.I.2.5 Block Coordinate Descent Algorithm

We now shift our focus to the Block Coordinate Descent (BCD) algorithm, which consists in alternating the optimisation of the barycentre positions X and the OT plans $\pi^{(k)}$. For convenience we will write $\pi := (\pi^{(1)}, \cdots, \pi^{(k)})$ and introduce the following energy:

$$J_{\mathrm{GWB}}(X, \pi) := \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n} \sum_{j=1}^{n_k} \pi_{i,j}^{(k)} \|P_k x_i - y_{k,j}\|_2^2. \tag{C.I.8}$$

We notice that $\mathcal{E}_{\mathrm{GWB}}(X) = \min_\pi \ J_{\mathrm{GWB}}(X, \pi)$, minimising over $\pi \in \prod_{k=1}^{K} \Pi(a, b_k)$.

Closed-form of $\operatorname{argmin}_X J_{\mathrm{GWB}}(X, \pi)$**..** Fixing $\pi = (\pi^{(1)}, \cdots, \pi^{(k)})$ with each $\pi^{(k)} \in \Pi(a, b_k)$, we see that $J_{\mathrm{GWB}}(\cdot, \pi)$ has a quadratic expression, and we compute (like in Eq. (C.I.7)):

$$\operatorname*{argmin}_{X \in \mathbb{R}^{n \times d}} J_{\mathrm{GWB}}(X, \pi) = \mathrm{diag}(1/a) \left(\sum_{k=1}^{K} \lambda_k \pi^{(k)} Y_k P_k \right) A^{-1}, \tag{C.I.9}$$

where we remind that our convention writes $X \in \mathbb{R}^{n \times d}$ and $Y_k \in \mathbb{R}^{n_k \times d_k}$, and we have used the fact that $A := \sum_k \lambda_k P_k^\top P_k$ is invertible by Assumption C.I.1. We denote by $\mathrm{diag}(1/a)$ the diagonal matrix with $1/a_i$ on the diagonal.

The optimisation in π with X fixed corresponds to K Kantorovich problems, thus we are now ready to formulate the BCD algorithm in Algorithm C.I.4.

Algorithm C.I.4: GWB resolution with Block Coordinate Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and linear maps $P_k : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_k}$. Number of iterations T, weights $a \in \triangle_n$.
Result: Barycentre positions $X \in \mathbb{R}^{n \times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n \times d}$;
for $t \in [\![1, T]\!]$ **do**
  for $k \in [\![1, K]\!]$ **do**
    Solve the OT problem: $\pi^{(k)} \in \operatorname*{argmin}_{\pi \in \Pi(a, b_k)} \sum_{i,j} \pi_{i,j} \|P_k x_i^{(k)} - y_{k,j}\|_2^2$;
  end
  Step the positions: $X_{t+1} \leftarrow \mathrm{diag}(1/a) \left(\sum_{k=1}^{K} \lambda_k \pi^{(k)} Y_k P_k \right) A^{-1}$;
end

C.I.2.6 Numerical Illustration

We illustrate the GWB problem with a simple example, where we attempt to find a point cloud of $\mathbb{R}^3$ whose 2D projections match given 2D points clouds that represent chess pieces. The experiment setting is presented in Fig. C.I.3 which embeds the 2D points clouds into $\mathbb{R}^3$ so as to represent the desired projections of the 3D barycentre.

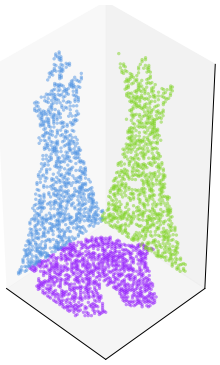
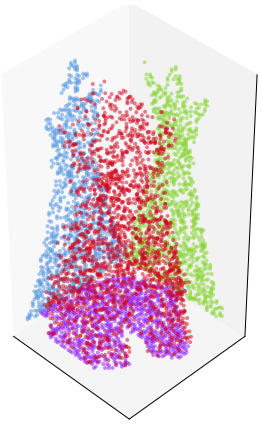
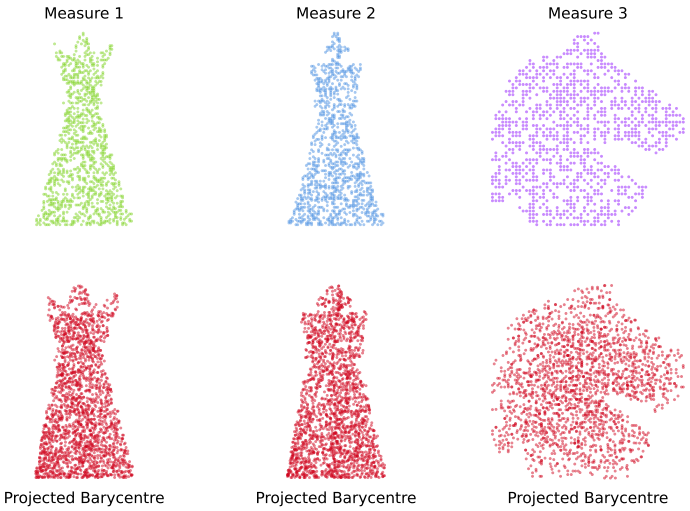


Figure C.I.3: The desired projections of the barycentre are the 2D points clouds of chess pieces. Formally, we have represented $P_k^\top \# \nu_k$ for each of the k measures. Our objective is to find a 3D chess-piece-like object whose two vertical projections (from the sides) match a queen piece (green, on the right) and a king piece (blue, on the left), and whose projection onto the ground plane matches a knight icon (purple, one the bottom).

For all our numerical experiments, we initialise with uniform noise and set $n := 2000$, each target point cloud having $n_k := 1000$ points. All code can be found on our GitHub repository. In Fig. C.I.4, we show the results of the GWB problem with the GD algorithm (Algorithm C.I.2).



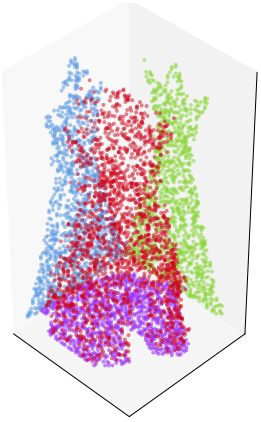
(a) 3D representation of the GD solution.



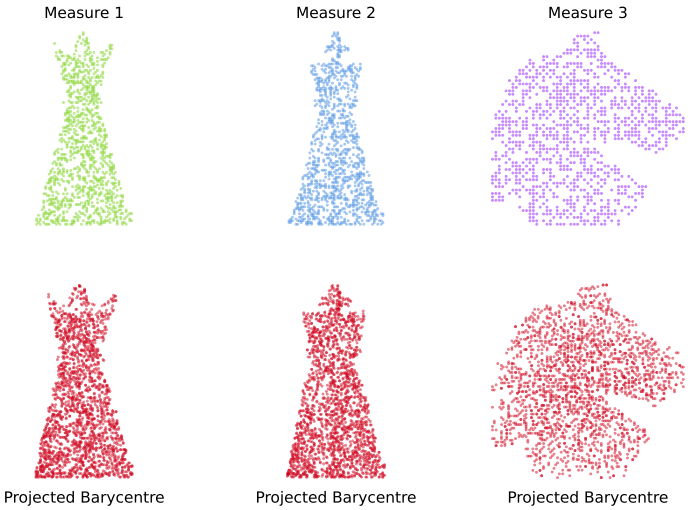
(b) Projections of the GD solution.

Figure C.I.4: Results of the GWB problem using the GD algorithm (Algorithm C.I.2).

In Fig. C.I.5, we now consider the results of the GWB problem with the SGD algorithm (Algorithm C.I.3), which visually appear slightly better, in particular concerning the precision of the cross on the knight piece.



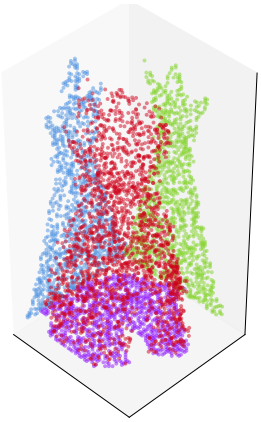
(a) 3D representation of the SGD solution.



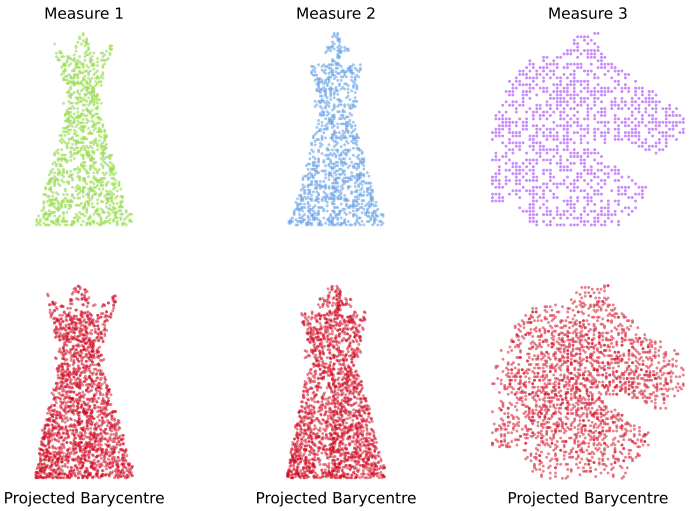
(b) Projections of the SGD solution.

Figure C.I.5: Results of the GWB problem using the SGD algorithm (Algorithm C.I.3).

Finally, we present the same visualisation with the BCD algorithm (Algorithm C.I.4), which appears to be comparable to the SGD solution.



(a) 3D representation of the BCD solution.



(b) Projections of the BCD solution.

Figure C.I.6: Results of the GWB problem using the BCD algorithm (Algorithm C.I.4).

To provide a quantitative comparison of the three algorithms on this particular example, we compute the cost $\mathcal{E}_{\mathrm{GWB}}$ at each iteration, and also plot the respective losses minimised by each algorithm (which are not directly comparable). We show these loss evolutions in Fig. C.I.7. Note that each algorithm is run until convergence to a stationary point, where the chosen criterion is $\|X_t - X_{t-1}\|_2^2 < 10^{-5}$ for each algorithm. We observe that the SGD algorithm reaches quickly a lower cost than GD, yet takes more iterations to satisfy the convergence criterion. The BCD algorithm appears to converge an order of magnitude faster than the other two algorithms, and reaches a final cost comparable to that of the SGD algorithm.

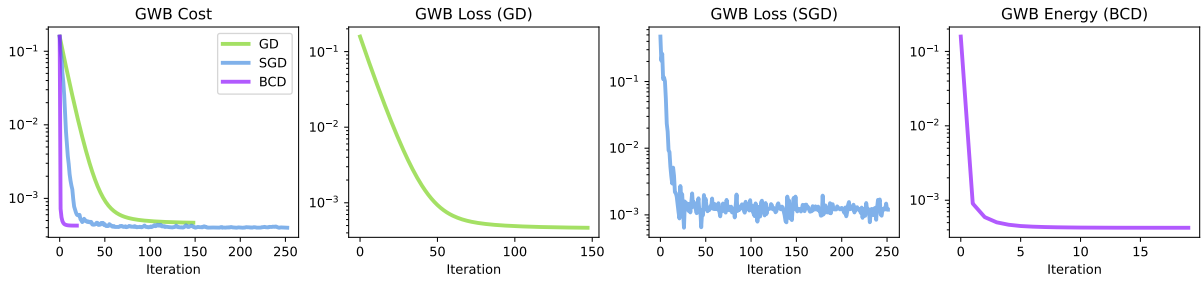


Figure C.I.7: GWB cost and respective losses of GD, SGD and BCD algorithms on a particular GWB problem.

C.I.3 The Blind GWB Problem

C.I.3.1 Problem Statement and Existence of a Solution

In this section, we consider a generalisation of the GWB problem studied in Section C.I.2, where optimisation is also performed of the linear maps $P := (P_k \in \mathbb{R}^{d_k \times d})_{k=1}^K$. We call this problem the Blind GWB (BGWB) problem, since the maps P_k, intuitively seen as projections, are not known. Following the definition of the GWB problem, we define the BGWB problem as:

$$\underset{\substack{\mu \in \mathcal{P}_2(\mathbb{R}^d) \\ \forall k \in [\![1,K]\!],\, P_k \in \mathbb{R}^{d_k \times d}}}{\operatorname{argmin}} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(P_k \# \mu, \nu_k) =: F(\mu, P). \tag{C.I.10}$$

A crucial difficulty of this problem is its ill-posedness, since for any invertible square matrix $R \in \mathbb{R}^{d \times d}$, we have $F(R\#\mu, PR^{-1}) = F(\mu, P)$, where $PR^{-1} := (P_k R^{-1})_k$. Indeed, $(P_k R^{-1})\#(R\#\mu) = \mu$. This property in inconvenient to prove existence, but in Proposition C.I.4, we reformulate the BGWB problem in a manner that allows us to prove existence similarly to the W_2 barycentre problem in [AC11, Proposition 2.3]. For convenience, we will write $\mathcal{Q} := \mathbb{R}^{d_1 \times d} \times \cdots \times \mathbb{R}^{d_K \times d}$.

Proposition C.I.4. For $k \in [\![1, K]\!]$, fix $\nu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$ and take weights $(\lambda_k) \in \triangle_K$. Then the BGWB problem formulated in Eq. (C.I.10) has a solution $(\mu^*, P^*) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{Q}$.

Proof. We introduce an equivalent variant of the BGWB which hides the invariances in a constraint set:

$$\underset{(\mu_1, \cdots, \mu_K) \in \mathcal{M}}{\operatorname{argmin}} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(\mu_k, \nu_k) =: G(\mu_1, \cdots, \mu_k), \tag{C.I.11}$$

where $\mathcal{M}$ is the set of measures $(\mu_1, \cdots, \mu_K) \in \mathcal{P}_2(\mathbb{R}^d)$ such that there exists $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $P_k \in \mathbb{R}^{d_k \times d}$ for $k \in [\![1, K]\!]$ such that $\mu_k = P_k \# \mu$. By continuity of the map $(Q, \rho) \in \mathbb{R}^{d_k \times d} \times \mathcal{P}_2(\mathbb{R}^{d_k}) \longmapsto Q\#\rho$, the set $\mathcal{M}$ is closed in $\Pi_k \mathcal{P}_2(\mathbb{R}^d)$ for the weak convergence of measures. By definition, we have that $\inf_{\mu, (P_k)} F(\mu, (P_k)) = \inf_{(\mu_k) \in \mathcal{M}} G((\mu_k))$.

Take now $((\mu_1^{(n)}, \cdots, \mu_K^{(n)}))_{n \in \mathbb{N}} \in \mathcal{M}^{\mathbb{N}}$ a minimising sequence of G. First, we fix $k \in [\![1, K]\!]$ and show that the sequence $m_n^2 := \int \|\cdot\|_2^2 \mathrm{d}\mu_k^{(n)}$ is bounded. Indeed[1], by convexity of $\|\cdot\|_2^2$, we have for any $x, y \in \mathbb{R}^{d_k}$, $\|x\|_2^2 \leq 2\|y\|_2^2 + 2\|x - y\|_2^2$. We then take $\pi \in \Pi^*(\mu_k^{(n)}, \nu_k)$ and integrate the previous inequality with respect to π:

$$m_n^2 = \int_{\mathbb{R}^{d_k}} \|x\|_2^2 \mathrm{d}\mu_k^{(n)}(x) \leq 2 \int_{\mathbb{R}^{d_k}} \|y\|_2^2 + 2\mathrm{W}_2^2(\mu_k^{(n)}, \nu_k),$$

and since $((\mu_1^{(n)}, \cdots, \mu_K^{(n)}))_{n \in \mathbb{N}}$ is a minimising sequence, the sequence $(\mathrm{W}_2^2(\mu_k^{(n)}, \nu_k))_n$ is bounded. We conclude that there exists $C > 0$ such that $\forall n \in \mathbb{N}$, $m_n^2 \leq C$.

[1] Many thanks to Nathaël Gozlan for pointing out this method.

Now, we show that the sequence $(\mu_k^{(n)})_n$ is tight using Markov's inequality (following a blog post by Djalil Chafaï). Fix $r > 0$ we denote by $B(0, \sqrt{r})$ the closed ball of radius $\sqrt{r}$ in $\mathbb{R}^{d_k}$ for the Euclidean norm. We have:

$$\sup_{n\in\mathbb{N}} \mu_k^{(n)}\left(B(0,\sqrt{r})^c\right) = \sup_{n\in\mathbb{N}} \mu_k^{(n)}\left(\left\{x \in \mathbb{R}^{d_k} : \|x\|_2^2 > r\right\}\right) \leq \frac{1}{r}\int_{\mathbb{R}^{d_k}} \|x\|_2^2 \mathrm{d}\mu_k^{(n)}(x) \leq \frac{C}{r},$$

where the first inequality is an application of Markov's inequality. For any $\varepsilon > 0$, it follows that the compact set $\mathcal{K}_\varepsilon := B(0, \sqrt{C/\varepsilon})$ is such that $\forall n \in \mathbb{N},\ \mu_k^{(n)}(\mathcal{K}_\varepsilon^c) \leq \varepsilon$. We conclude that the sequence $(\mu_k^{(n)})_n$ is tight.

By Prokhorov's theorem, we can extract a subsequence $(\mu_k^{(\alpha_k(n))})_n$ that converges weakly to a measure $\mu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$. Applying this method to consecutive sub-extractions, we can introduce an extraction $\alpha : \mathbb{N} \longrightarrow \mathbb{N}$ such that for all $k \in [\![1, K]\!]$, the sequence $(\mu_k^{(\alpha(n))})_n$ converges weakly to $\mu_k \in \mathcal{P}_2(\mathbb{R}^{d_k})$. By closedness of the set $\mathcal{M}$, we have that $(\mu_k) \in \mathcal{M}$. By Kantorovich duality (with the formulation of [San15, Theorem 1.40] for example), the map $\mu \longmapsto \mathrm{W}_2^2(\mu, \nu_k)$ is a supremum of continuous functions, hence is it lower semi-continuous on $\mathcal{P}_2(\mathbb{R}^{d_k})$ for the weak convergence of measures. We conclude that:

$$G(\mu_1, \cdots, \mu_K) \leq \liminf_{n\longrightarrow+\infty} G(\mu_1^{(\alpha(n))}, \cdots, \mu_K^{(\alpha(n))}) = \inf_{(\rho_k)\in\mathcal{M}} G(\rho_1, \cdots, \rho_K),$$

thus $(\mu_1, \cdots, \mu_K)$ is optimal for the variant of the BGWB problem introduced in Eq. (C.I.11). Since $(\mu_1, \cdots, \mu_K) \in \mathcal{M}$, there exists $P \in \mathcal{Q}$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mu_k = P_k\#\mu$ for $k \in [\![1, K]\!]$, and thus the infimum of the BGWB problem is attained. □

Note that when the measure μ is fixed, the optimisation problem in (P_k) reduces to K independent problems, each fining a linear map $P_k \in \mathbb{R}^{d_k \times d}$ that minimises the W_2 distance between $P_k\#\mu$ and ν_k. This is a particular case of the problem studied in Chapter B.I, where we minimised an OT cost $\mathcal{T}_c(g\#\mu, \nu)$ over $g \in G$, where G is a set of functions.

C.I.3.2 Lagrangian Energy Minimisation

Like in Section C.I.2.3, we focus on the case where the ν_k are discrete measures and consider the restriction of the BGWB problem to measures $\gamma_X := \sum_{i=1}^n a_i\delta_{x_i} \in \mathcal{P}_2(\mathbb{R}^d)$ with a support $X \in \mathbb{R}^{n\times d}$ of fixed size n and fixed weights $a \in \triangle_n$. Our goal is to minimise the energy:

$$\operatorname*{argmin}_{\substack{X\in\mathbb{R}^{n\times d}\\ P\in\mathcal{Q}}} \sum_{k=1}^K \lambda_k \mathrm{W}_2^2(P_k\#\gamma_X, \nu_k) =: \mathcal{E}_{\mathrm{BGWB}}(X, P), \tag{C.I.12}$$

where we remind $\mathcal{Q} := \mathbb{R}^{d_1\times d} \times \cdots \times \mathbb{R}^{d_K\times d}$. The energy $\mathcal{E}_{\mathrm{BGWB}}$ can be written as a minimum over a finite amount of quadratic expressions in X and P as in Eq. (C.I.6), and we deduce likewise the following regularity property:

> **Proposition C.I.5.** The functional $\mathcal{E}_{\mathrm{BGWB}}$ is semi-algebraic and locally Lipschitz on $\mathbb{R}^{n\times d} \times \mathcal{Q}$.

C.I.3.3 (Stochastic) Gradient Descent Algorithm

Here, we adapt the method described in Algorithm C.I.2 to the BGWB problem. As in Section C.I.2.4, we use the results of Section B.I.4.1 to exhibit Clarke (sub)-gradients of the energy $\mathcal{E}_{\mathrm{BGWB}}$ with respect to positions X and the maps P. With respect to X, the computations are identical, and we can introduce $\varphi_{\mathcal{E}_{\mathrm{BGWB}}}(X, P) := \varphi_{\mathcal{E}_{\mathrm{GWB}}}(X, P)$ as in Eq. (C.I.7), which is a Clarke sub-gradient of $\mathcal{E}_{\mathrm{BGWB}}$ with respect to X. Again, for SGD it will be convenient to introduce the gradients of each term k in the sum, which we denote by $\varphi_k(X, P)$. With respect to P, we use similar computations to obtain the following sub-gradient $\psi_k(X, P_k)$ at $X \in \mathbb{R}^{n\times d}$

with respect to $P_k \in \mathbb{R}^{d_k \times d}$, and the complete sub-gradient $\psi_{\mathcal{E}_{\mathrm{BGWB}}}$ of $\mathcal{E}_{\mathrm{BGWB}}$ with respect to P:

$$\begin{aligned} \psi_{\mathcal{E}_{\mathrm{BGWB}}}(X, P) &:= [\lambda_k \psi_k(X, P_k)]_{k \in [\![1,K]\!]}, \\ \psi_k(X, P_k) &:= 2\left(P_k C_X - Y_k^\top \left[\pi^{(k)}\right]^\top X\right) \in \mathbb{R}^{d_k \times d}, \end{aligned} \tag{C.I.13}$$

with $C_X := \sum_i a_i x_i x_i^\top = X^\top \mathrm{diag}(a) X$. As in Section C.I.2.4, our sub-gradients $\varphi_{\mathcal{E}_{\mathrm{BGWB}}}(X, P)$ and $\psi_{\mathcal{E}_{\mathrm{BGWB}}}(X, P)$ depend on a choice of OT plans $(\pi^{(k)})$. We can now formulate the GD algorithm for the BGWB problem in Algorithm C.I.5.

Algorithm C.I.5: BGWB resolution with Gradient Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$. Number of iterations T, weights $a \in \triangle_n$, learning rates (α_t).

Result: Barycentre positions $X \in \mathbb{R}^{n \times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n \times d}$ and $P_0 \in \mathcal{Q}$;

for $t \in [\![1, T]\!]$ **do**

  Step the positions: $X_{t+1} \leftarrow X_t - \alpha_t \, \varphi_{\mathcal{E}_{\mathrm{BGWB}}}(X_t, P_t)$;

  Step the maps: $P_{t+1} \leftarrow P_t - \alpha_t \, \psi_{\mathcal{E}_{\mathrm{BGWB}}}(X_t, P_t)$;

end

Concerning convergence, we can continue as in Section C.I.2.4 and use the semi-algebraicity of the energy $\mathcal{E}_{\mathrm{BGWB}}$ to show that accumulation points of the GD algorithm are Clarke critical, however we have no guarantee of boundedness of the iterates.

Proposition C.I.6. Convergence of GD for BGWB, application of [EN97, Theorem 4.1] Consider learning rates (α_t) such that $\alpha_t \to 0$ and $\sum_t \alpha_t = +\infty$. Then accumulation points of the iterates (X_t, P_t) of Algorithm C.I.5 are Clarke critical points of $\mathcal{E}_{\mathrm{GWB}}$, and the sequence $(\mathcal{E}_{\mathrm{BGWB}}(X_t, P_t))$ converges.

Proof. By Proposition C.I.5 and Proposition B.I.8, we can apply [EN97, Theorem 4.1], which is precisely the statement of the Proposition. □

We now consider the SGD variant of Algorithm C.I.5, using the same notation as in Section C.I.2.4:

Algorithm C.I.6: BGWB resolution with Stochastic Gradient Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$. Number of iterations T, weights $a \in \triangle_n$, learning rates (α_t).

Result: Barycentre positions $X \in \mathbb{R}^{n \times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n \times d}$ and $P_0 \in \mathcal{Q}$;

for $t \in [\![1, T]\!]$ **do**

  Draw $k \sim \mathbb{D}_\lambda$ and step the positions: $X_{t+1} \leftarrow X_t - \alpha_t \, \varphi_k(X_t, P_t)$;

  Step the map P_k: $P_{t+1}^{(k)} \leftarrow P_t^{(k)} - \alpha_t \, \psi_k(X_t, P_t^{(k)})$;

end

Regarding convergence, we can use the exact same arguments as in the proof of Proposition C.I.3 to show that sub-sequential limits the iterates (X_t, P_t) of Algorithm C.I.6 are almost-surely Clarke critical points of the energy $\mathcal{E}_{\mathrm{GWB}}$, assuming boundedness.

Proposition C.I.7 (Convergence of SGD for BGWB, application of [Dav+20] Theorem 4.2)**.** Consider learning rates $(\alpha_t) \in \mathbb{R}_+^{\mathbb{N}}$ such that $\sum_t \alpha_t = +\infty$ and $\sum_t \alpha_t^2 < +\infty$. Take iterates (X_t, P_t) of Algorithm C.I.3, and assume that they are almost-surely bounded. Then almost-surely, any sub-sequential limit of (X_t, P_t) is a Clarke critical point of $\mathcal{E}_{\mathrm{BGWB}}$.

We do not enter into theoretical details about natural variants of Algorithms C.I.5 and C.I.6 which heuristically enforce boundedness of iterates. Such ideas include the use of a projection step of each P_k on the Stiefel manifold $\mathbb{S}_{d_k,d}$ of unitary $d_k \times d$ matrices, forcibly normalising the rows of each P_k at each iterations, the projection of X onto a compact set, or the use of a regularisation term in the energy $\mathcal{E}_{\mathrm{BGWB}}$ which penalises the size of X and P. In practice, it is unclear whether these methods are necessary to ensure convergence, and it is equally unclear whether they improve the convergence speed and the quality of attained critical points.

C.I.3.4 Block Coordinate Descent Algorithm

We now adapt the BCD algorithm described in Section C.I.2.5 to the BGWB problem, and to this end we introduce the BGWB energy:

$$J_{\mathrm{BGWB}}(X, \pi, P) := \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n} \sum_{j=1}^{n_k} \pi_{i,j}^{(k)} \|P_k x_i - y_{k,j}\|_2^2, \tag{C.I.14}$$

where $\pi = (\pi^{(k)})_{k=1}^K$ with each $\pi^{(k)} \in \Pi(a, b_k)$ and $P = (P_1, \cdots, P_k) \in \mathcal{Q} := \Pi_k \mathbb{R}^{d_k \times d}$.

The closed-form expression of the minimisation in X computed in Eq. (C.I.9) still holds, however it requires the invertibility of $A_P := \sum_k \lambda_k P_k^\top P_k$, which at the time was guaranteed by Assumption C.I.1, but in the BGWB case, this is no longer obvious. If A_P is invertible, we can compute the minimisation in X as:

$$\operatorname*{argmin}_{X \in \mathbb{R}^{n \times d}} J_{\mathrm{BGWB}}(X, \pi, P) = \operatorname{diag}(1/a) \left(\sum_{k=1}^{K} \lambda_k \pi^{(k)} Y_k P_k \right) A_P^{-1}. \tag{C.I.15}$$

We are now interested in the minimisation over P_k We compute as in Eq. (C.I.13), separating the problem into K independent problems, and assuming that $C_X := X^\top \operatorname{diag}(a) X$ is invertible, we compute:

$$\operatorname*{argmin}_{P_k \in \mathbb{R}^{d_k \times d}} J_{\mathrm{BGWB}}(X, \pi, P) = \left(Y_k^\top \left[\pi^{(k)} \right]^\top X C_X^{-1} \right)_{k \in [\![1,K]\!]}. \tag{C.I.16}$$

A natural question is whether the invertibility conditions on A_P and C_X are maintained across BCD iterations. Concerning the step in X, assuming that A_P is invertible and that each matrix $\pi^{(k)} Y_k$ has a rank at least d_k, standard linear algebra techniques show that the minimising matrix X^* in Eq. (C.I.15) is such that C_{X^*} is invertible. This condition requires $n_k \geq d_k$, and even in this case, there is no theoretical guarantee that $\pi^{(k)} Y_k$ be of full rank in general, but this property is satisfied in practice numerically, barring engineered pathological cases. As for the step in P, when $n \geq d$ and C_X is invertible, then the minimising linear maps P^* of Eq. (C.I.16) are such that A_P is invertible provided that the $n \times n$ matrix $S := \sum_k \lambda_k \pi^{(k)} Y_k Y_k^\top [\pi^{(k)}]^\top$ is invertible. Again, this condition is not guaranteed in general, but it is satisfied in practice. To summarise, to ensure that the invertibility conditions are satisfied across BCD iterations, we need to assume that $n_k \geq d_k$ and $n \geq d$, then invertibility is numerically very likely to be verified in practice.

We can now formulate the BCD algorithm for the BGWB problem in Algorithm C.I.7, which is an extension of Algorithm C.I.4.

Algorithm C.I.7: BGWB resolution with Block Coordinate Descent

Data: Barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$ and weights $b_k \in \triangle_{n_k}$. Number of iterations T, weights $a \in \triangle_n$.

Result: Barycentre positions $X \in \mathbb{R}^{n \times d}$.

Initialisation: Draw $X_0 \in \mathbb{R}^{n \times d}$ and $P_0 \in \mathcal{Q}$;

for $t \in [\![1, T]\!]$ **do**

  for $k \in [\![1, K]\!]$ **do**

    Solve the OT problem: $\pi^{(k)} \in \underset{\pi \in \Pi(a, b_k)}{\operatorname{argmin}} \sum_{i,j} \pi_{i,j} \|P_k x_i^{(k)} - y_{k,j}\|_2^2$;

  end

  Step the positions: $X_{t+1} \leftarrow \operatorname{diag}(1/a) \left(\sum_{k=1}^K \lambda_k \pi^{(k)} Y_k P_k \right) A^{-1}$;

  Step the maps: for $k \in [\![1, K]\!]$: $P_{t+1}^{(k)} \leftarrow Y_k^\top \left[\pi^{(k)}\right]^\top X C_X^{-1}$;

end

C.I.3.5 Numerical Illustration on a Reconstruction Problem

In this section, we illustrate numerical solutions of the BGWB problem using Algorithms C.I.5 to C.I.7 on a problem where we know that an exact matching is possible: we consider measures ν_k which are of the form $\overline{P}_k \# \overline{\mu}_0$, which implies that $P = (\overline{P}_k)$ and $\mu := \overline{\mu}$ is a solution of the BGWB problem, with a cost of 0 (this is not the only solution, as the problem is ill-posed). Even in this favourable setting, the BGWB remains challenging to solve numerically, and the results are sensitive to the hyper-parameters and initialisations. As a reasonable simplification, we initialised the projections as random unitary matrices, and at each iteration of the GD and SGD algorithms, we projected each row of each P_k onto the unit sphere. This heuristic ensured the boundedness of the iterations and eased the convergence of the GD and SGD algorithms.

We represent the "ground truth" barycentre and its projections (by the "ground truth" projections) in Fig. C.I.8. To embed a projection into $\mathbb{R}^d$, we represent $\overline{P}_k^\top \overline{P}_k \# \overline{\mu}_0$, and offset the position orthogonally to the projection for legibility.

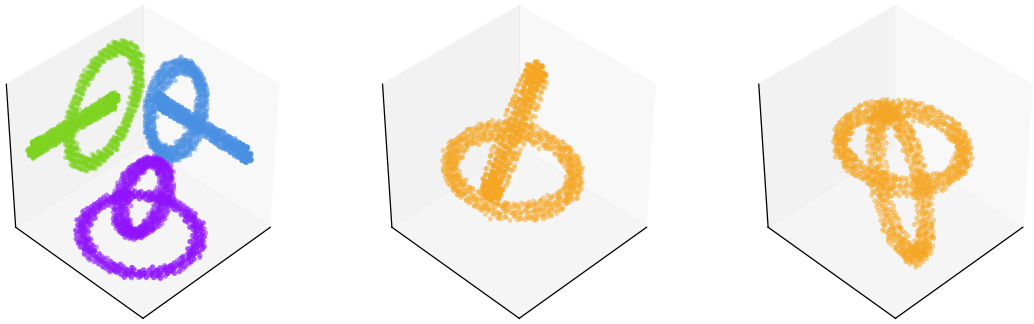


Figure C.I.8: The "ground truth" barycentre $\overline{\mu}_0$ and its projections $\overline{P}_k \# \overline{\mu}_0$ for $k = 1, 2, 3$, with different viewing angles. The projections are represented in green, blue and purple, while the ground truth barycentre is represented in orange.

We begin with the Gradient Descent algorithm (Algorithm C.I.5) and visualise the (learned) projections of the (learned) barycentre, comparing them with the target 2D measures in Fig. C.I.9. The results are visually satisfactory, even though the learned projections are very different from the ground truth, the projections $P_k \# \gamma_X$ are very close to the target 2D measures.

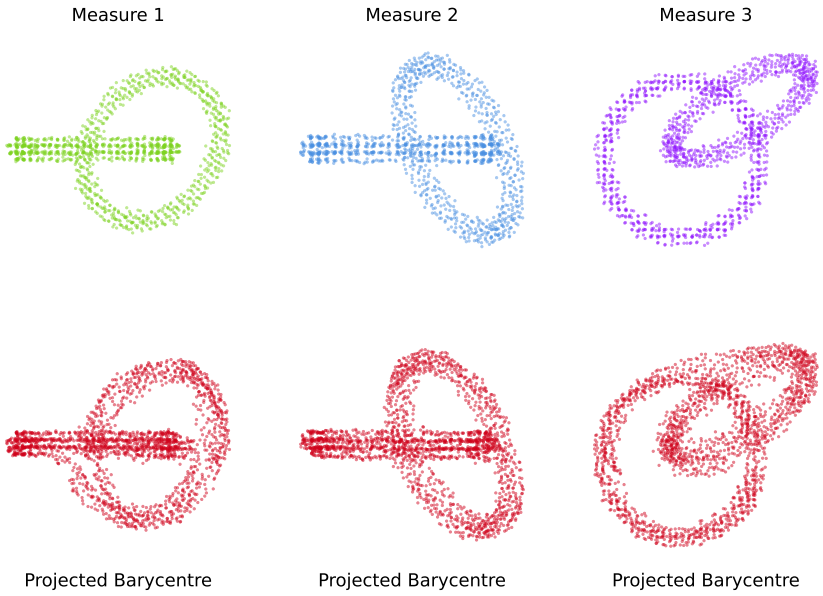


Figure C.I.9: The projections of the BGWB barycentre learned by GD (Algorithm C.I.5).

In Fig. C.I.10, we visualise the learned barycentre in $\mathbb{R}^3$ and the embedded target measures $(P_k^\top P_k \# \overline{\mu})$.

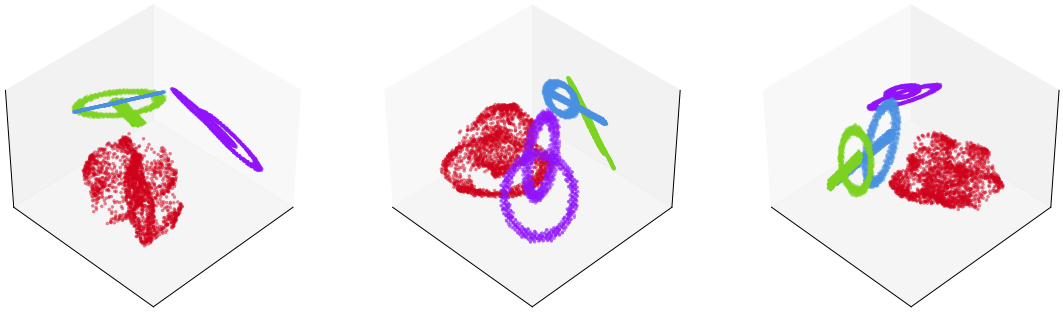


Figure C.I.10: The barycentre learned by GD (Algorithm C.I.5) in $\mathbb{R}^3$, represented in red, and the embedded target measures (in blue, green and purple).

Finally, we observe some iterations of the GD algorithm through their projections in Fig. C.I.11. We notice that the iterations stabilise halfway through the training process, however the stopping criterion ($\|X_t - X_{t-1}\|_2^2 \leq 10^{-5}$) is never met by the algorithm, which exhausts all the (300) allowed iterations.

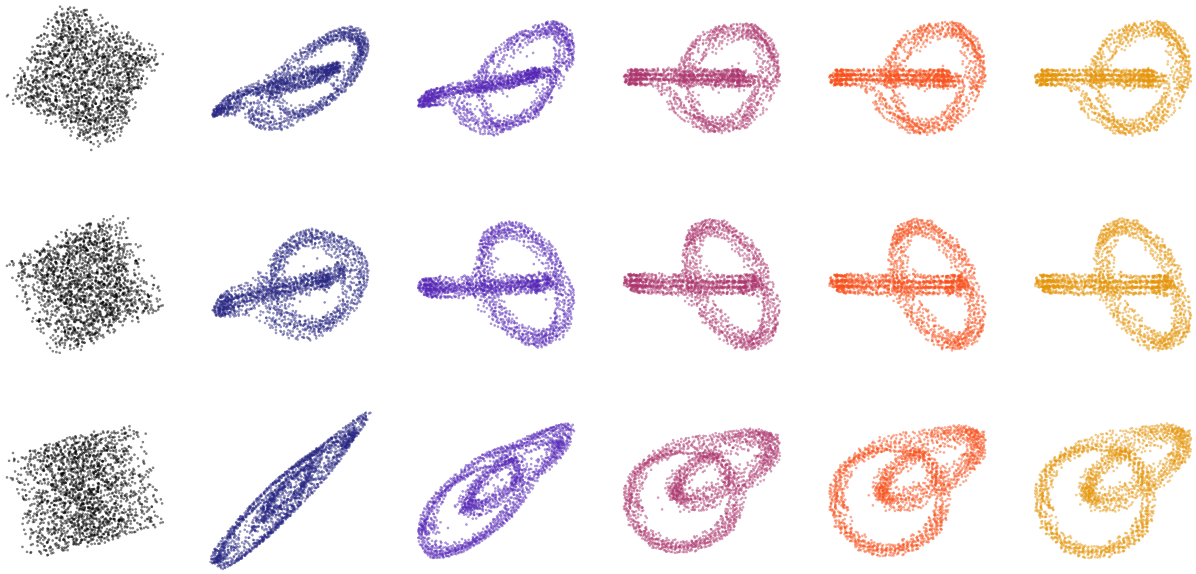


Figure C.I.11: The projections of the barycentre learned by GD (Algorithm C.I.5) at different iterations. Each row corresponds to one of the $K = 3$ projections, and each column corresponds to evenly spaced iterations between the initialisation (uniform noise on the cube $[0, 1]^d$) for the positions, and the final iteration.

Moving on to the Stochastic Gradient Descent algorithm (Algorithm C.I.6), we observe in Fig. C.I.12 that the learned projections are close to the target measures, albeit with more artifacts are outlier points than for the GD solution.

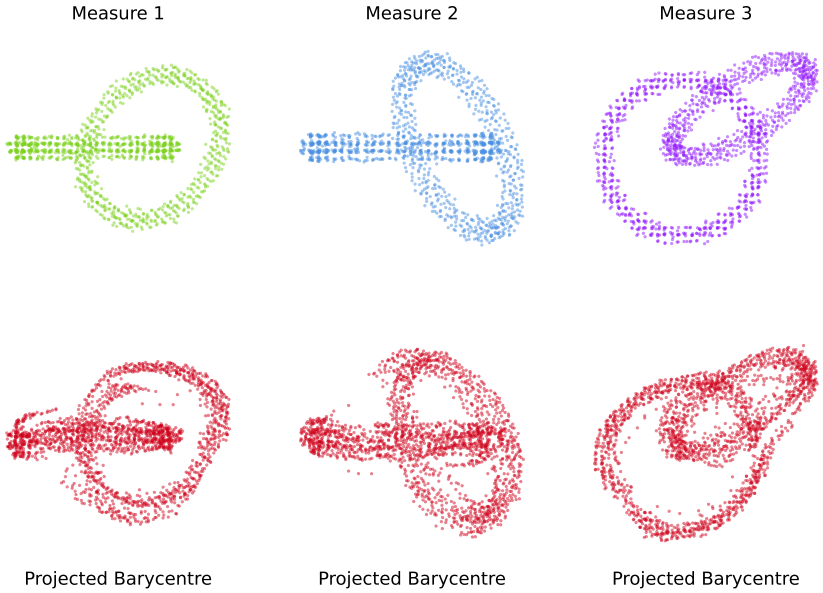


Figure C.I.12: The projections of the BGWB barycentre learned by SGD (Algorithm C.I.6).

In Fig. C.I.13, we visualise the learned barycentre in $\mathbb{R}^3$ as in Fig. C.I.10:

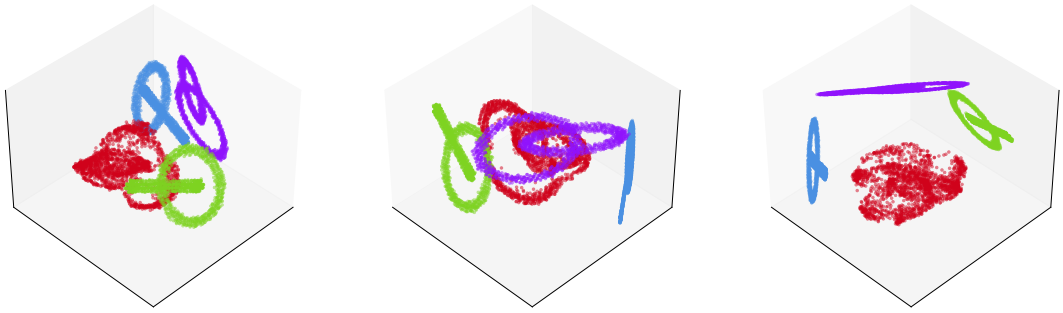


Figure C.I.13: The barycentre learned by SGD (Algorithm C.I.6) in $\mathbb{R}^3$, represented in red, and the embedded target measures (in blue, green and purple).

We present in Fig. C.I.14 the projections of the barycentre learned by SGD at different iterations. We observe, like in the GD case, that the iterations stabilise halfway through training, although this time the stopping criterion is met before the maximum number of iterations is reached.

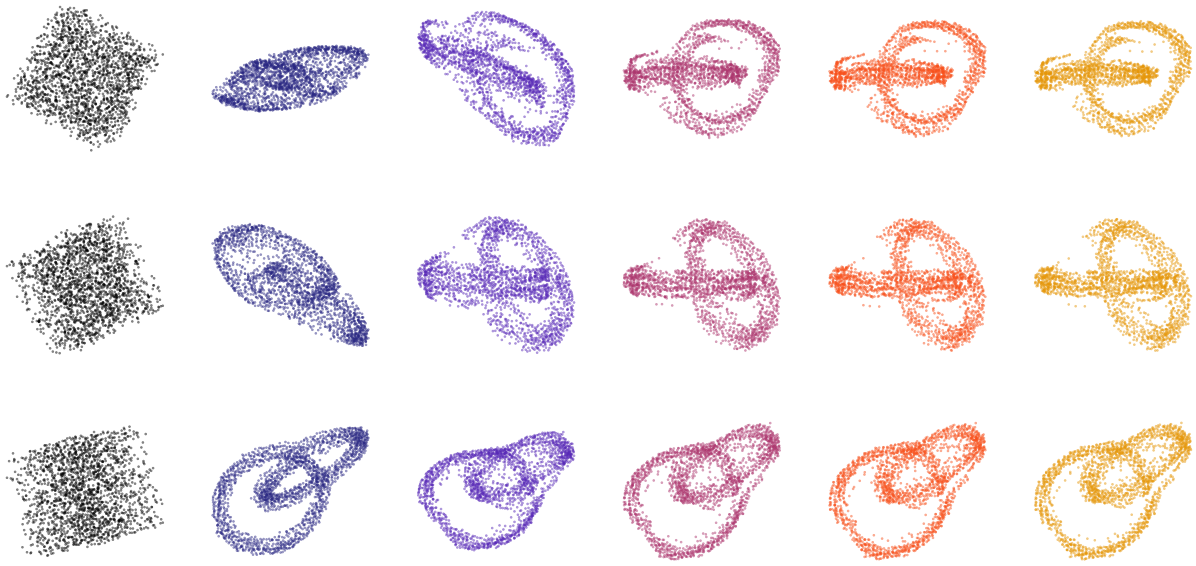


Figure C.I.14: The projections of the barycentre learned by SGD (Algorithm C.I.6) at different iterations. Each row corresponds to one of the $K = 3$ projections, and each column corresponds to evenly spaced iterations between the initialisation (uniform noise on the cube $[0, 1]^d$) for the positions, and the final iteration.

Regarding the BCD algorithm (Algorithm C.I.7), we observe in Fig. C.I.15 that the learned projections are reasonably close to the targets, but present numerous artifacts.

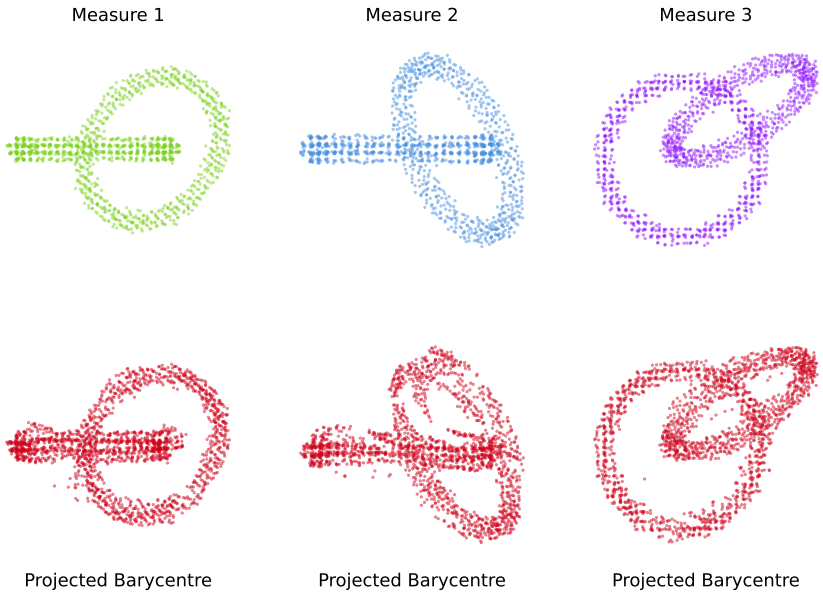


Figure C.I.15: The projections of the BGWB barycentre learned by BCD (Algorithm C.I.7).

In Fig. C.I.16, we visualise the learned barycentre in $\mathbb{R}^3$ as in Fig. C.I.10:

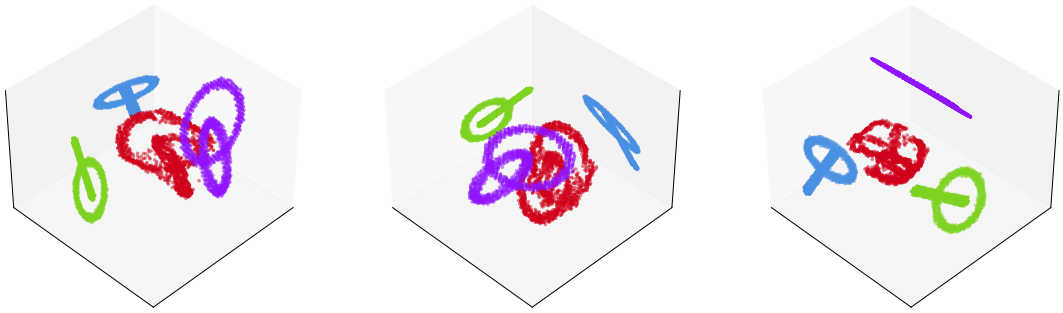


Figure C.I.16: The barycentre learned by BCD (Algorithm C.I.7) in $\mathbb{R}^3$, represented in red, and the embedded target measures (in blue, green and purple).

We present in Fig. C.I.17 the projections of the barycentre learned by BCD at different iterations. We observe that the iterations reach acceptable positions very quickly (in practice, in about 5 iterations), but fail to reach the visual quality of the GD and SGD solutions.

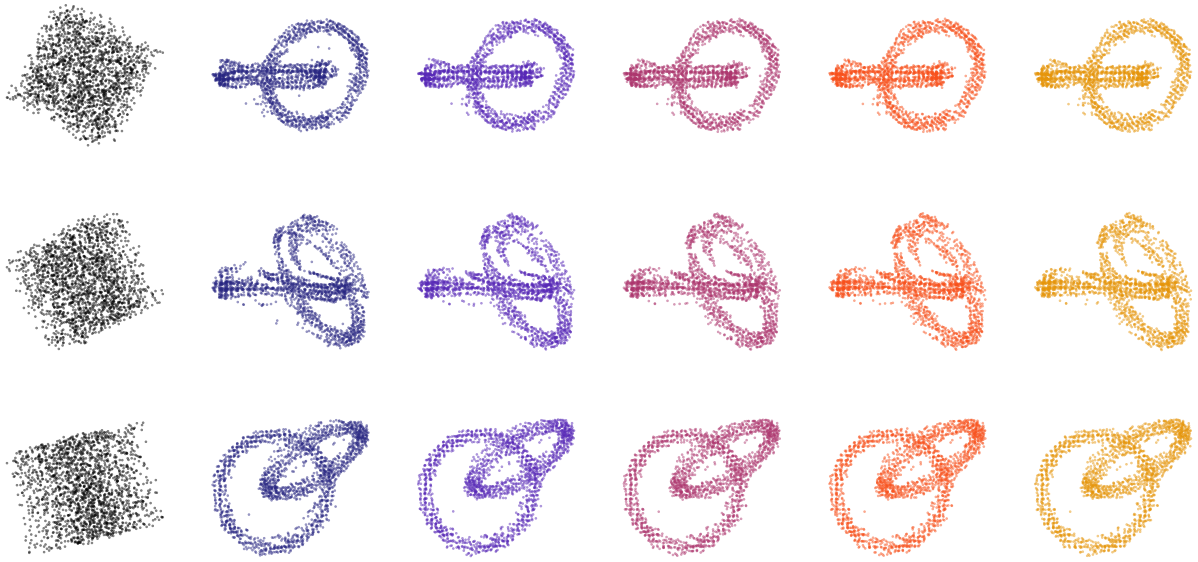


Figure C.I.17: The projections of the barycentre learned by BCD (Algorithm C.I.7) at different iterations. Each row corresponds to one of the $K = 3$ projections, and each column corresponds to evenly spaced iterations between the initialisation (uniform noise on the cube $[0, 1]^d$) for the positions, and the final iteration.

We now compare the respective losses and energies of the three algorithms, as well as the value of the energy $\mathcal{E}_{\mathrm{BGWB}}(X, P)$ at each iteration. Due to convergence instability, we had to lower the learning rate of the SGD algorithm by a factor of $1/K$, which is crucial to bear in mind when comparing convergence speeds. The main takeaway is that the BCD algorithm converges much faster to an acceptable solution, while GD and SGD convergence slower to a solution that can reach better lower loss values and better visual results.

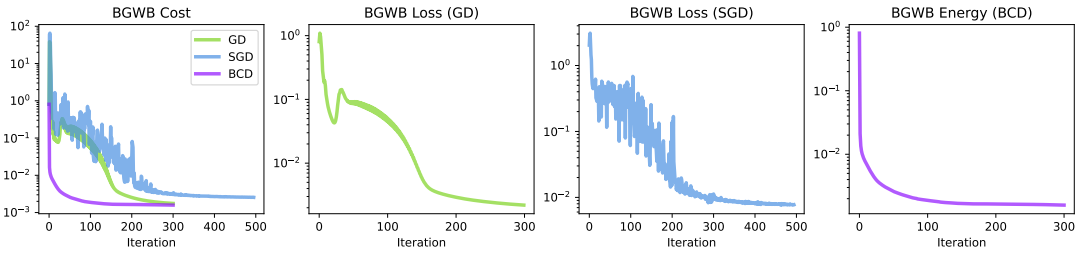


Figure C.I.18: BGWB cost and respective losses of GD, SGD and BCD algorithms on a particular reconstruction BGWB problem.

C.II

Computing Optimal Transport Barycentres

C.II.1 Introduction . . . 284
C.II.1.1 Related Works and Motivation 284
C.II.1.2 Contributions and Outline 285
C.II.2 Lifting Ground Barycentres to Measures . . . 286
C.II.3 A Fixed-Point Algorithm . . . 287
C.II.3.1 Algorithm Definition 287
C.II.3.2 Convergence of Fixed-Point Iterations 289
C.II.3.3 Expression of the Iterates when the Plans are Maps 295
C.II.3.4 Extension to the Entropic Case 297
C.II.3.5 The Particular Case of Conditionally Independent Couplings 299
C.II.4 Focus on the Discrete Case . . . 300
C.II.4.1 Discrete Expression and Algorithms 300
C.II.4.2 Correspondence of Gradient Descent with Fixed-Point Iterations 303
C.II.4.3 Discrete Uniqueness Discussion 304
C.II.4.4 Application to Gaussian Mixture Model Barycentres 305
C.II.5 Numerical Illustrations . . . 306
C.II.5.1 Toy Example for Barycentre Computation 306
C.II.5.2 Illustration with Norm Powers 307
C.II.5.3 Study of the Support Size of Iterates of G 309
C.II.5.4 Comparison with the Multi-Marginal Formulation 310
C.II.5.5 Generalised Wasserstein Barycentre Computation 313
C.II.5.6 Non-linear Generalised Wasserstein Barycentre Computation 313
C.II.5.7 Gaussian Mixture Model Barycentres 314
C.II.5.8 Colour Transfer on a Barycentre of Colour Distributions 315

Abstract

Wasserstein barycentres represent average distributions between multiple probability measures for the Wasserstein distance. The numerical computation of Wasserstein barycentres is notoriously challenging. A common approach is to use Sinkhorn iterations, where an entropic regularisation term is introduced to make the problem more manageable. Another approach involves using fixed-point methods, akin to those employed for computing Fréchet means on manifolds. The convergence of such methods for 2-Wasserstein barycentres, specifically with a quadratic cost function and absolutely continuous measures, was studied by Alvarez-Esteban et al. in [Álv+16]. In this chapter, we delve into the main ideas behind this fixed-point method and explore how it can be generalised to accommodate more diverse transport costs and generic probability measures, thereby extending its applicability to a broader range of problems. We show convergence results for this approach and illustrate its numerical behaviour on several barycentre problems. This chapter is based on the paper:

[TDG24] Eloi Tanguy, Julie Delon and Nathaël Gozlan.
"Computing Barycentres of Measures for Generic Transport Costs".
arxiv preprint 2501.04016 (Dec. 2024).

C.II.1 Introduction

C.II.1.1 Related Works and Motivation

Wasserstein barycentres represent a powerful concept in Optimal Transport theory, enabling the computation of average distributions between multiple probability measures. These barycentres preserve the geometric structure of the underlying distributions, making them particularly suited for machine learning tasks. They have proven useful in numerous applications, including image processing [Rab+12], computer graphics [Sol+15; BPC16], statistics [BCP19], domain adaptation [MM21], generative modelling [Kor+22], fairness in machine learning [Gor+19] or model selection in Bayesian learning [Bac+22]. Wasserstein barycentres are also at the core of clustering methods such as K-means, to define centroids in spaces of probability measures [Ho+17; Mi+18].

The classical notion of barycentre refers to the weighted average of a set of points (x_k) with positive weights (λ_k) summing to 1, in a metric space (E, d). Formally, a barycentre $\bar{x}$ is a point that minimises the weighted sum of (typically squared) distances:

$$\bar{x} \in \operatorname*{argmin}_{x \in E} \sum_{k=1}^{K} \lambda_k d^2(x, x_k).$$

This concept can be extended to the space of probability measures, where d can be replaced for instance by a transportation cost $\mathcal{T}_c$. We remind that for two probability measures μ and ν on metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the optimal transport cost between μ and ν for the ground cost c is defined as

$$\mathcal{T}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \mathrm{d}\pi,$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. Considering K different cost functions c_k, the barycentre problem can be written in this setting as

$$\bar{\mu} \in \operatorname*{argmin}_{\mu} \sum_{k=1}^{K} \lambda_k \mathcal{T}_{c_k}(\mu, \nu_k). \tag{C.II.1}$$

When $(\mathcal{X}, d_\mathcal{X}) = (\mathcal{Y}, d_\mathcal{Y})$ is a Polish space and $c = d_\mathcal{X}^p$ with $p \geq 1$, $\mathrm{W}_p(\mu, \nu) := (\mathcal{T}_{d^p}(\mu, \nu))^{\frac{1}{p}}$ defines a distance between probability measures (with finite moment of order p), called p-Wasserstein distance. In this case, the barycentre $\bar{\mu}$ defined above is called a Wasserstein barycentre. Generalisation to a barycentre of a probability measure on $\mathcal{P}(\mathcal{X})$ and the consistency of their discrete approximations is also studied by several authors [AC17].

The theoretical analysis of Wasserstein barycentres begins with the foundational work by Carlier and Ekeland [CE10], who studied the existence, uniqueness and dual formulations for barycentre problems with generic continuous cost functions. Subsequent work by [AC11] re-established the existence and dual formulations of such barycentres for the quadratic Wasserstein distance W_2 on Euclidean spaces, and showed uniqueness under the hypothesis that one of the original measures is absolutely continuous. More recent studies have broadened these results: [CCE24] extended the theoretical analysis to Wasserstein medians (W_1), studying their stability properties, and investigated dual and multi-marginal formulations. [BFR25] further extended the framework to W_p distances for $p > 1$, proving existence and uniqueness of barycentres for absolutely continuous measures on $\mathbb{R}^d$. A follow-up study by [BFR24] analysed the general case for strictly convex and $\mathcal{C}^2$ cost functions with non-degenerate Hessian.

From a computational perspective, calculating Wasserstein barycentres is known to be a highly challenging problem, classified as NP-hard. According to [AB22], although polynomial-time algorithms exist for computing Wasserstein barycentres with a fixed number of points, their

computational complexity scales exponentially with respect to the dimension of the space, or with respect to the number of marginals. This makes direct computation infeasible for high-dimensional problems or large sets of distributions, which are common in practical applications.

To tackle these computational challenges, several approximate methods have been developed for Wasserstein barycentres. The first chapter to propose an algorithmic solution for computing these barycentres was by [Rab+12], which computed Sliced Wasserstein barycentres through a gradient descent approach. This method leveraged the sliced Wasserstein distance to achieve an efficient approximation, significantly simplifying computations.

A natural approach to develop easily computable approximations of such barycentres is to replace transport costs $\mathcal{T}_c$ by regularised versions

$$\mathcal{T}_{c,\varepsilon}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu),$$

as proposed in [CD14]. When the support of the distributions and barycentre is fixed (a grid for instance), the problem can be rewritten as a KL projection problem and the so-called entropic barycentre can be computed efficiently with a modified version of Sinkhorn's algorithm [Ben+15; PC19b].

In order to deal with distributions without imposed support a second approach also described in [CD14] relies on a fixed-point algorithm inspired by the computation of Fréchet means on manifolds. Each step of this fixed point approach consists in replacing the current barycentre μ by its image measure by the map $\sum_{k=1}^K \lambda_k T_k$, where the T_k are optimal maps between μ and ν_k (assuming these maps exist). The authors of [Álv+16] were the first to establish a rigorous proof of convergence for this fixed-point approach in the case of absolutely continuous measures ν_k: more precisely, they proved convergence of a subsequence to a fixed point and showed that if the fixed point is unique, it is indeed a barycentre. Their study focuses specifically on the case of W_2 barycentres, with applications demonstrated mainly on Gaussian measures. Although their proof is only provided for absolutely continuous measures, this fixed point approach is frequently used for discrete measures and probably the baseline free-support method provided in numerical optimal transport libraries [Fla+21]. Building on the same ideas as [Álv+16], the author of [Lin23] extends the investigation of the fixed point algorithm for discrete measures on $\mathbb{R}^d$, limited to just one single iteration, and deriving a worst-case error bound in the W_2 and W_1 settings. The iterative solver of [Álv+16] has also been extended in high dimensional settings by [Kor+22], which use a neural solver for computing the optimal maps T_k.

In closely related directions, several other approaches have been proposed to compute Wasserstein barycentres over Riemannian manifolds [KP17], or Gromov-Wasserstein barycentres [BB25; BBS23] and the approach we develop in this chapter share similarities with [BB25].

C.II.1.2 Contributions and Outline

In this chapter, we develop a fixed-point approach to compute barycentres between probability measures for generic transport costs, i.e. solutions of the optimisation problem Eq. (C.II.1). Our only hypotheses are that we work on compact spaces, and that the ground costs c_k are continuous and such that $\mathrm{argmin}_x \sum_{k=1}^K \lambda_k c_k(x, x_k)$ is uniquely defined. In particular, we do not assume existence of optimal transport maps between μ and the ν_k, and we do not assume anything on the probability measures μ and ν_k. We propose an iterative fixed-point algorithm generalising [Álv+16] in this generic case. We show that the sequences generated by this algorithm have converging sub-sequences, that limits must be fixed-points of a certain mapping G, and that a barycentre for Eq. (C.II.1) is also a fixed point of G. We show that these results still hold for entropic regularised transport costs.

Numerically, we show that our approach specifically allows to extend the recent definition of generalised Wasserstein barycentres presented in [DGS21], notably by considering non-linear functions between the ambient space and the subspaces of measures ν_k. It also enables efficient computation of barycentres for the mixture Wasserstein metric [DD20], which until now were calculated using their multi-marginal equivalent formulation.

The chapter is organised as follows. In Section C.II.2, we introduce a novel notion of Optimal Transport barycentres in a certain space between measures ν_k on potentially different spaces for generic costs c_k. In Section C.II.3, we propose a fixed-point algorithm which generalises [Álv+16] and converges to solutions (in a certain sense). We re-write the problem in a discrete setting in Section C.II.4 and illustrate our method in Section C.II.5 on several numerical examples, providing a publicly available Python toolkit.

C.II.2 Lifting Ground Barycentres to Measures

We work with probability measures ν_k on compact metric spaces $(\mathcal{Y}_k, d_{\mathcal{Y}_k})_{k\in[\![1,K]\!]}$, of which we will seek a "barycentre" μ in a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$. To compare a measure $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$ and $\mu \in \mathcal{P}(\mathcal{X})$ we consider continuous cost functions $c_k : \mathcal{X} \times \mathcal{Y}_k \longrightarrow \mathbb{R}_+$. A barycentre will be a minimiser of the sum of the transport costs with respect to the measure ν_k, leading to the following energy for a measure $\mu \in \mathcal{P}(\mathcal{X})$:

$$V(\mu) := \sum_{k=1}^{K} \mathcal{T}_{c_k}(\mu, \nu_k), \tag{C.II.2}$$

hence our minimisation problem reads

$$\operatorname*{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} V(\mu). \tag{C.II.3}$$

Note that to introduce barycentre weights λ_k, it suffices to replace c_k with $\lambda_k c_k$, which allows us to include weights in the costs and alleviate notation. We summarise our standing assumptions on the spaces and costs in Assumption C.II.1:

Assumption C.II.1. The metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}_k, d_{\mathcal{Y}_k})$ are compact, and the costs $c_k : \mathcal{X} \times \mathcal{Y}_k \longrightarrow \mathbb{R}_+$ are continuous.

Existence of solutions for Problem Eq. (C.II.3) was established by [CE10, Proposition 2] under Assumption C.II.1.

Remark C.II.1. Uniqueness was proven in [CE10, Proposition 4] if, essentially, for at least one k, the problem $\mathcal{T}_{c_k}(\mu, \nu_k)$ has a Monge solution, for which they assume that each ν_k is absolutely continuous on $\mathcal{Y}_k = \overline{\Omega}$ with Ω an open and bounded subset of $\mathbb{R}^d$ with with $\nu_k(\partial\Omega) = 0$. They also assume that the costs $c_k(\cdot, y)$ are Lipschitz with a uniform constant L and that c_k verifies the Twist condition: $c_k(\cdot, y)$ is differentiable, with $\partial_x c_k(x, \cdot)$ injective.

The definition of a barycentre between measures ν_k can be seen as a lifting of a notion of barycentre within $\mathcal{X}$ of points $(y_1, \cdots, y_K) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$. To give mathematical meaning to this intuition and to our method, we will make the following assumption throughout the chapter:

Assumption C.II.2. For all $(y_1, \cdots, y_K) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$, the set $\operatorname*{argmin}_{x\in\mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k)$ has a unique element.

The uniqueness of the optimisation problem in Assumption C.II.2 allows us to introduce the ground barycentre function B:

$$B : \left\{ \begin{array}{ccc} \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K & \longrightarrow & \mathcal{X} \\ (y_1, \cdots, y_K) & \longmapsto & \operatorname*{argmin}_{x\in\mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k). \end{array} \right. \tag{C.II.4}$$

For convenience, we introduce $\mathcal{Y} := \Pi_k \mathcal{Y}_k$, equipped with the product distance, with the notation $Y := (y_1, \cdots, y_K)$ for an element of $\mathcal{Y}$, as well as the total cost function:

$$C := \left\{ \begin{array}{ccc} \mathcal{X} \times \mathcal{Y} & \longrightarrow & \mathbb{R}_+ \\ (x, y_1, \cdots, y_K) & \longmapsto & \sum_{k=1}^K c_k(x, y_k) \end{array} \right. . \tag{C.II.5}$$

Equipped with these convenient notations, we can write the multi-marginal formulation of our barycentre problem:

$$\underset{\pi \in \Pi(\nu_1, \cdots, \nu_K)}{\operatorname{argmin}} \int_{\mathcal{Y}} C(B(Y), Y) \mathrm{d}\pi(Y). \tag{C.II.6}$$

The barycentre problem defined in Eq. (C.II.3) is related to the multi-marginal formulation through the following equation, due to [CE10, Proposition 3.3]:

$$\underset{\mu \in \mathcal{P}(\mathcal{X})}{\operatorname{argmin}} V(\mu) = B\# \underset{\pi \in \Pi(\nu_1, \cdots, \nu_K)}{\operatorname{argmin}} \int_{\mathcal{Y}} C(B(Y), Y) \mathrm{d}\pi(Y). \tag{C.II.7}$$

The following technical result uses the continuity of the c_k and Assumption C.II.2 to show that B is continuous.

Lemma C.II.1. The function $B : \mathcal{Y} \longrightarrow \mathcal{X}$ defined in Eq. (C.II.4) is continuous.

Proof. The proof uses standard compactness arguments, showing that for $Y_n \xrightarrow[n \longrightarrow +\infty]{} Y \in \mathcal{Y}$, $(B(Y_n))$ can only have $B(Y)$ as a subsequential limit. □

Another important technical result is the regularity of transport costs, which we will use repeatedly. We gather well-known results in Lemma C.II.2.

Lemma C.II.2. Consider E, F compact metric spaces and let $c : E \times F \longrightarrow \mathbb{R}_+$ a measurable cost function. The optimal transport cost $\mathcal{T}_c$ has the following regularity for the weak convergence of measures depending on c:

1. If c is lower-semi-continuous, then $\mathcal{T}_c$ is lower-semi-continuous.
2. If c is continuous, then $\mathcal{T}_c$ is continuous.
3. If $E = F$ and c is l.s.c. with $c(x, y) = 0 \Longrightarrow x = y$, then $\mathcal{T}_c(\mu, \nu) = 0 \Longrightarrow \mu = \nu$.

Proof. Regarding item 1), by [San15, Theorem 1.42], Kantorovich duality holds for c l.s.c. and thus $\mathcal{T}_c$ can be written as a supremum of l.s.c. functions, hence is l.s.c.. For item 2), the result is verbatim [San15, Theorem 1.51]. For item 3), if $\mathcal{T}_c(\mu, \nu) = 0$ then there exists $\pi \in \Pi(\mu, \nu)$ such that $\int_{E^2} c(x, y) \mathrm{d}\pi(x, y) = 0$ (existence follows from lower semi-continuity, as in [San15, Theorem 1.5]). Thus for π-almost-every (x, y), $c(x, y) = 0$, which by assumption gives $x = y$, hence (using the same technique as in [San15, Proposition 5.1]) for any test function $\phi \in \mathcal{C}^0(E, \mathbb{R})$:

$$\int_E \phi(x) \mathrm{d}\mu(x) = \int_{E^2} \phi(x) \mathrm{d}\pi(x, y) = \int_{E^2} \phi(y) \mathrm{d}\pi(x, y) = \int_E \phi(y) \mathrm{d}\nu(y),$$

which shows that $\mu = \nu$. □

C.II.3 A Fixed-Point Algorithm

C.II.3.1 Algorithm Definition

In this section, we define a sequence $(\mu_t) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$ that will approach a barycentre of fixed measures $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$. We propose a modified version of the iterated scheme from [Álv+16]

to solve Eq. (C.II.3). To define an iteration mapping, for $\mu \in \mathcal{P}(\mathcal{X})$, we consider the set of multi-marginal couplings

$$\Gamma(\mu) := \left\{ \gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K) \ : \ \forall k \in [\![1, K]\!], \ \gamma_{0,k} \in \Pi^*_{c_k}(\mu, \nu_k) \right\}, \tag{C.II.8}$$

where, for all k, $\gamma_{0,k}$ denotes the $\mathcal{X} \times \mathcal{Y}_k$ marginal of γ and $\Pi^*_{c_k}(\mu, \nu_k)$ denotes the set of all optimal couplings for the transport problem between μ and ν_k associated to the cost function c_k. The existence of such multi-couplings is a consequence of the well-known "gluing lemma" (see [San15, Lemma 5.5]). Gluing K plans $\pi^*_k \in \Pi(\mu, \nu_k)$ with Lemma C.II.3 provides an explicit element of $\Gamma(\mu)$.

Lemma C.II.3. For $k \in [\![1, K]\!]$, let $\pi_k \in \Pi(\mu, \nu_k)$. Write the disintegration of π_k with respect to its first marginal μ as $\pi_k(\mathrm{d}x, \mathrm{d}y_k) = \mu(\mathrm{d}x)\pi^x_k(\mathrm{d}y_k)$. Then the measure γ defined as:

$$\gamma(\mathrm{d}x, \mathrm{d}y_1, \cdots, \mathrm{d}y_K) := \mu(\mathrm{d}x)\pi^x_1(\mathrm{d}y_1) \cdots \pi^x_K(\mathrm{d}y_K) \tag{C.II.9}$$

is such that $\forall k \in [\![1, K]\!], \ \gamma_{0,k} \in \Pi(\mu, \nu_k)$.

Proof. Taking the notations of the statement of the Lemma, take $k \in [\![1, K]\!]$ and a test function $\phi \in \mathcal{C}^0_b(\mathcal{X} \times \mathcal{Y}_k)$:

$$\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}_1\times\cdots\times\mathcal{Y}_K} \phi(x, y_k)\mathrm{d}\gamma(x, y_1, \cdots, y_K) &= \int_{\mathcal{X}\times\mathcal{Y}_1\times\cdots\times\mathcal{Y}_K} \phi(x, y_k)\mathrm{d}\mu(x)\mathrm{d}\pi^x_1(y_1)\cdots\mathrm{d}\pi^x_K(y_K) \\
&= \int_{\mathcal{X}\times\mathcal{Y}_k} \phi(x, y_k) \left(\prod_{l\neq k} \int_{\mathcal{Y}_l} \mathrm{d}\pi^x_l(y_l) \right) \mathrm{d}\mu(x)\mathrm{d}\pi^x_k(y_k) \\
&= \int_{\mathcal{X}\times\mathcal{Y}_1\times\cdots\times\mathcal{Y}_K} \phi(x, y_k)\mathrm{d}\pi_k(x, y_k).
\end{aligned}$$

□

By abuse of notation, we will denote $B\#\gamma := B\#\gamma_{1,\cdots,K}$, where $\gamma_{1,\cdots,K} \in \mathcal{P}(\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k)$ is the marginal of γ with respect to $(y_1, \cdots, y_K)$. In terms of random variables, if $(X, Y_1, \cdots, Y_K) \sim \gamma$, then $B\#\gamma = \mathrm{Law}[B(Y_1, \cdots, Y_K)]$. Denoting $B\#\Gamma(\mu) := \{B\#\gamma, \ \gamma \in \Gamma(\mu)\}$, we define the multi-valued mapping G which maps $\mu \in \mathcal{P}(\mathcal{X})$ to the set of next iterates $G(\mu) \subset \mathcal{P}(\mathcal{X})$:

$$G := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \rightrightarrows & \mathcal{P}(\mathcal{X}) \\ \mu & \mapsto & B\#\Gamma(\mu) \end{array} \right. . \tag{C.II.10}$$

Note that this construction is similar to that of [Álv+16, Remark 3.4]. Moreover, the candidate barycentre $\overline{\mu} = B\#\gamma_{1,\cdots,K}$ is closely related to the multi-marginal formulation of the barycentre problem (see Eq. (C.II.7)). Indeed, set $\pi := \gamma_{1,\cdots,K} \in \Pi(\mu_1, \cdots, \mu_K)$, notice that π is a candidate for the multi-marginal problem of a particular structure induced by the reference measure μ. In the case where the plans $\gamma_{0,k}$ are induced by maps T_k, then this structure is the coupling $(T_1, \cdots, T_K)\#\mu$. In terms of random variables, if $X \sim \mu$, then the chosen coupling is $(T_1(X), \cdots, T_K(X))$.

Taking inspiration from the W_2^2 case, we can see informally the iterate $\overline{\mu} \in G(\mu)$ as a local linearisation of $\mathcal{P}(\mathcal{X})$. To illustrate this intuition, we consider the case $\mathcal{X} = \mathcal{Y}_1 = \cdots = \mathcal{Y}_K$ and assume that for each k, the set of optimal plans $\Pi^*_{c_k}(\mu, \nu_k)$ is reduced to (I, T_k), or in other words, that the Monge problem has a unique solution. Informally, one may see the set of maps $T : \mathcal{X} \longrightarrow \mathcal{X}$ sending μ to a measure $T\#\mu \in \mathcal{P}(\mathcal{X})$ as the tangent space to $\mathcal{P}(\mathcal{X})$ at μ. As a result, the problem of finding a barycentre $\overline{\mu}$ can be seen from the viewpoint of the reference measure μ in the tangent space $T_\mu\mathcal{P}(\mathcal{X})$ as the problem of finding $S \in T_\mu\mathcal{P}(\mathcal{X})$ such that $S\#\mu$ would minimise the cost V. Our approach takes a barycentre of the optimal maps T_k by choosing the candidate $S := B \circ (T_1, \cdots, T_K)$. In the case of the squared-Euclidean cost on the common space $\mathbb{R}^d$, this amounts to $S := \sum_k \lambda_k T_k$, which is exactly the Linearised Optimal Transport barycentre approximation for the reference measure μ, as introduced in [MDC20, Section 4.3]. We illustrate this viewpoint schematically in Fig. C.II.1.

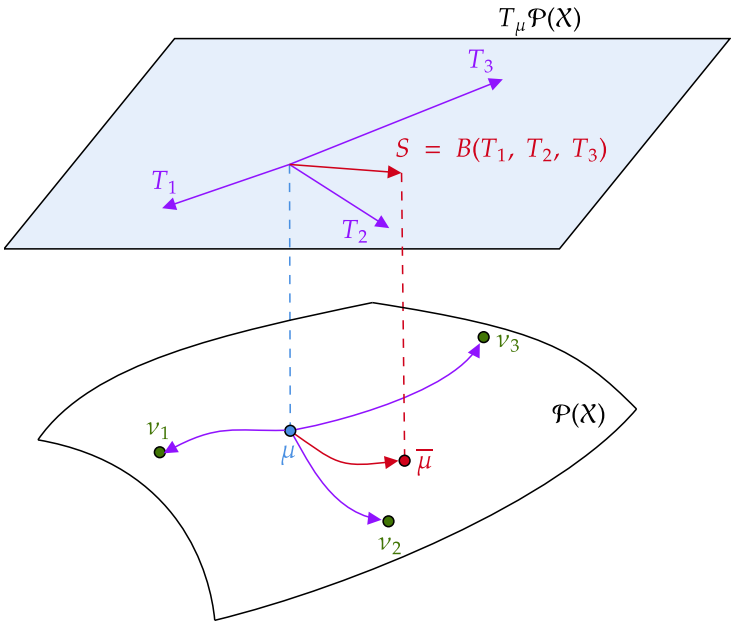


Figure C.II.1: Illustration of the informal linearisation interpretation for the barycentre candidate $\overline{\mu} = B \circ (T_1, \cdots, T_K)\#\mu$.

Starting from a measure $\mu_0 \in \mathcal{P}(\mathcal{X})$, our algorithm consists of choosing iterates through the multi-function G:

$$\forall t \in \mathbb{N},\ \mu_{t+1} \in G(\mu_t).$$

We dedicate the next section to a theoretical study of the convergence of this fixed-point iteration.

C.II.3.2 Convergence of Fixed-Point Iterations

We can formulate a regularity result of the multi-valued map G: namely, we will show that G is *upper hemi-continuous*. For the sake of simplicity, we will take the following definition[1]:

Definition C.II.1. A multi-valued function $\varphi : E \rightrightarrows F$ from a compact metric space E to parts of a compact metric space F is said to be *upper hemi-continuous* (u.h.c.) if for any sequence $(x_n, y_n) \in (E \times F)^{\mathbb{N}}$ such that $y_n \in \varphi(x_n)$ and $x_n \xrightarrow[n\longrightarrow+\infty]{} x \in E$, there exists an extraction such that $y_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{} y \in F$ with $y \in \varphi(x)$.

For more technical reasons, we also need to introduce the notion of *lower hemi-continuity*[2]

Definition C.II.2. A multi-valued function $\varphi : E \rightrightarrows F$ from a compact metric space space E to parts of a compact metric space space F is said to be *lower hemi-continuous* (l.h.c.) if for any sequence $(x_n) \in E^{\mathbb{N}}$ such that $x_n \xrightarrow[n\longrightarrow+\infty]{} x \in E$, then for any $y \in F$ such that $y \in \varphi(x)$, there exists an extraction α and a sequence $(y_n) \in F^{\mathbb{N}}$ such that $y_n \in \varphi(x_{\alpha(n)})$ and $y_n \xrightarrow[n\longrightarrow+\infty]{} y$.

To illustrate the technical differences between these two notions, we consider two specific multi-valued functions in Fig. C.II.2.

[1] We refer to [AB94, Chapter 17] for a more general definition and introduction to these concepts on Polish spaces. We choose a stronger sequential definition from [AB94, Theorem 17.20], which in their vocabulary corresponds to u.h.c multi-functions with compact values.

[2] whose formulation is is equivalent to [AB94, Definition 17.2], by their [AB94, Theorem 17.21].

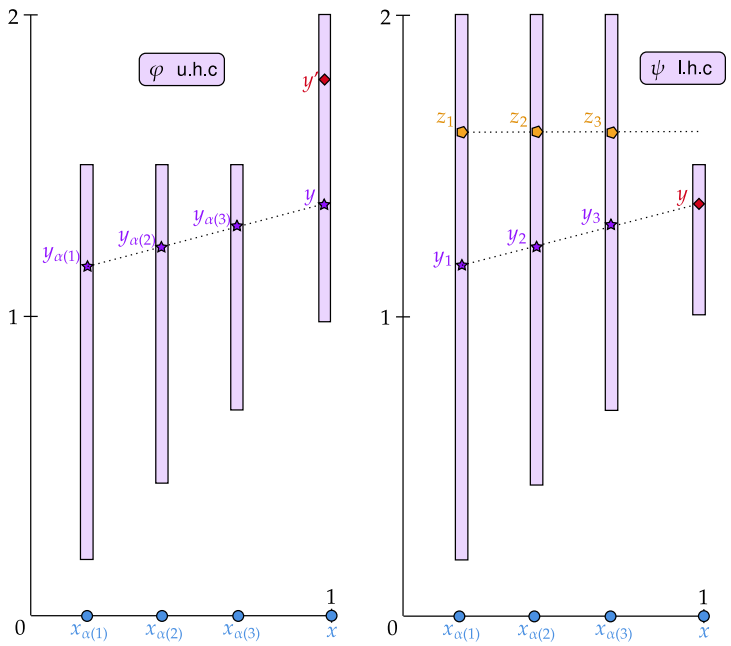


Figure C.II.2: *Left:* the multi-function $\varphi : [0,1] \rightrightarrows [0,2]$ defined by $\forall x \in [0,1)$, $\varphi(x) = [x, 3/2]$ and $\varphi(1) = [1,2]$ is u.h.c.. Indeed, taking any sequence (x_n, y_n) such that $y_n \in \varphi(x_n)$ and $x_n \xrightarrow[n\longrightarrow+\infty]{} x$, there exists an extraction α such that $y_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{} y \in \varphi(x)$. However, φ is not l.h.c. at 1 since the target $y' := 7/4 \in \varphi(1)$ can never be a limit of a sequence (x_n, y_n) with $x_n \xrightarrow[n\longrightarrow+\infty]{} 1$ and $y_n \in \varphi(x_n)$.
Right: $\psi : [0,1] \rightrightarrows [0,2]$ defined by $\forall x \in [0,1)$, $\psi(x) = [x, 2]$ and $\psi(1) = [1, 3/2]$ is l.h.c.. Take $x_n \xrightarrow[n\longrightarrow+\infty]{} x$ and a target $y \in \psi(x)$. Then there exists an extraction α and a sequence (y_n) such that $y_n \in \psi(x_n)$ and $y_n \xrightarrow[n\longrightarrow+\infty]{} y$. However, ψ is not u.h.c: take $x_n \xrightarrow[n\longrightarrow+\infty]{} 1$ and the sequence $z_n := 5/3$. We have $\forall n \in \mathbb{N}$, $z_n \in \psi(x_n)$, however any subsequence of (z_n) converges to $5/3 \notin \psi(1)$.

Finally, an hemi-continuous multi-map is one that is both u.h.c. and l.h.c.:

Definition C.II.3. A multi-valued function $\varphi : E \rightrightarrows F$ from a compact metric space space E to parts of a compact metric space space F is said to be *hemi-continuous* if it is both u.h.c. (Definition C.II.1) and l.h.c. (Definition C.II.2).

We begin with technical lemmas on the hemi-continuity properties of sets of couplings.

Lemma C.II.4. Consider E, F compact metric spaces and $\nu \in \mathcal{P}(F)$. The multi-function

$$\Pi_\nu := \left\{ \begin{array}{ccc} \mathcal{P}(E) & \rightrightarrows & \mathcal{P}(E \times F) \\ \mu & \mapsto & \Pi(\mu, \nu) \end{array} \right. \tag{C.II.11}$$

is hemi-continuous.

Proof. **u.h.c..** We apply Definition C.II.1: introduce $\mu_n \xrightarrow[n\longrightarrow+\infty]{w} \mu \in \mathcal{P}(E)$ and $\pi_n \in \Pi(\mu_n, \nu)$. Since $\mathcal{P}(E \times F)$ is compact, we can introduce α an extraction such that $\pi_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \pi \in \mathcal{P}(E \times F)$. By continuity of marginalisation, we deduce $\pi \in \Pi(\mu, \nu)$, which shows that Π_ν is u.h.c. by definition.

l.h.c.. We consider W_1, the 1-Wasserstein distance on $\mathcal{P}(E)$ (i.e. $\mathcal{T}_{d_E}$), and use the same notation for the 1-Wasserstein distance on $\mathcal{P}(E^2)$, with the distance $d_{E^2}((x,y),(x',y')) := \max(d_E(x,x'), d_E(y,y'))$, both of which metrise the weak convergence by [Vil09, Corollary 6.13]. We apply Definition C.II.2: take $\mu_n \xrightarrow[n\longrightarrow+\infty]{w} \mu \in \mathcal{P}(E)$, and let $\pi \in \Pi(\mu, \nu)$. Consider (X, Y)

two coupled random variables of law π, and for $n \in \mathbb{N}$, take X_n a random variable such that (X, X_n) is an optimal coupling for $\mathrm{W}_1(\mu, \mu_n)$, and let $\pi_n := \mathrm{Law}(X_n, Y)$. We have

$$\mathrm{W}_1(\pi, \pi_n) \leq \mathbb{E}\left[d_{E^2}\left((X, Y), (X_n, Y)\right)\right] = \mathbb{E}\left[\max(d_E(X, X_n), d_E(Y, Y))\right] = \mathrm{W}_1(\mu, \mu_n),$$

then by metrisation, we get $\mathrm{W}_1(\mu, \mu_n) \xrightarrow[n \longrightarrow +\infty]{} 0$, then $\pi_n \xrightarrow[n \longrightarrow +\infty]{w} \pi$, concluding the proof that Π_ν is l.h.c.. □

We can apply Berge's maximisation theorem to show that the set of *optimal* transport plans is upper hemi-continuous for a continuous cost function:

Lemma C.II.5. Consider E, F compact metric spaces, a continuous cost $c : E \times F \longrightarrow \mathbb{R}_+$ and $\nu \in \mathcal{P}(F)$. The multi-function

$$[\Pi_c^*]_\nu := \left\{ \begin{array}{ccc} \mathcal{P}(E) & \rightrightarrows & \mathcal{P}(E \times F) \\ \mu & \mapsto & \Pi_c^*(\mu, \nu) \end{array} \right. \tag{C.II.12}$$

is upper hemi-continuous.

Proof. By compactness, the map $\pi \longmapsto \int_{E \times F} c \mathrm{d}\pi$ is continuous, and by Lemma C.II.4, the multi-map $\mu \rightrightarrows \Pi(\mu, \nu)$ is hemi-continuous (with compact values), hence by Berge's maximisation theorem from [AB94, Theorem 17.31], the map

$$[\Pi_c^*]_\nu : \mu \longmapsto \Pi_c^*(\mu, \nu) = \underset{\pi \in \Pi(\mu, \nu)}{\mathrm{argmin}} \int_{E \times F} c \mathrm{d}\pi$$

is upper hemi-continuous. □

Remark C.II.2. The multifunction $[\Pi_c^*]_\nu$ is not **lower** hemi-continuous. Indeed, take the following points of $\mathbb{R}^2$:

$$\begin{aligned} &\forall n \in \mathbb{N},\ x_n := (-1, 2^{-n}),\ y_n := (1, -2^{-n}), \\ &x_\infty := (-1, 0),\ y_\infty := (1, 0),\ w := (0, 1),\ z := (0, -1), \end{aligned}$$

and the following discrete measures (see Fig. C.II.3):

$$\forall n \in \mathbb{N},\ \mu_n := \frac{1}{2}(\delta_{x_n} + \delta_{y_n}),\ \mu_\infty := \frac{1}{2}(\delta_{x_\infty} + \delta_{y_\infty}),\ \nu := \frac{1}{2}(\delta_w + \delta_z).$$

We have $\mu_n \xrightarrow[n \longrightarrow +\infty]{w} \mu_\infty$, and a unique OT plan for the cost $c(\cdot, \cdot) := \|\cdot - \cdot\|_2^2$ between μ_n and ν, which sends x_n to w and y_n to z:

$$\forall n \in \mathbb{N},\ \Pi_c^*(\mu_n, \nu) = \{\pi_n\},\ \pi_n := \frac{1}{2}(\delta_{x_n, w} + \delta_{y_n, z}),$$

with $\pi_n \xrightarrow[n \longrightarrow +\infty]{w} \pi_\infty := \frac{1}{2}(\delta_{x_\infty, w} + \delta_{y_\infty, z})$. However, the set of optimal plans between the limit μ_∞ and ν has more than one element, since $\|x_\infty - w\|_2^2 = \|x_\infty - z\|_2^2$ and $\|y_\infty - w\|_2^2 = \|y_\infty - z\|_2^2$:

$$\Pi_c^*(\mu, \nu) = \left\{(1 - t)\pi_\infty + t\pi',\ t \in [0, 1]\right\},\ \pi' := \frac{1}{2}(\delta_{x_\infty, z} + \delta_{y, w}).$$

We conclude that there does not exist an extraction α and a sequence (π'_n) such that $\forall n \in \mathbb{N},\ \pi'_n \in \Pi_c^*(\mu_{\alpha(n)}, \nu)$ and $\pi'_n \xrightarrow[n \longrightarrow +\infty]{w} \pi'$.

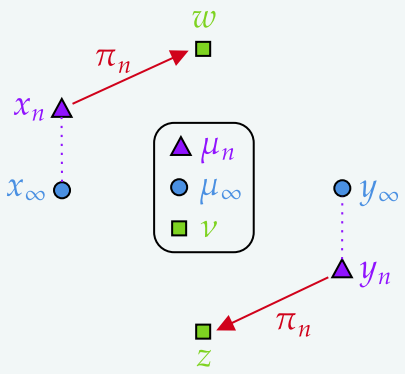


Figure C.II.3: Counter-example from Remark C.II.2 showing that $\Pi_c^*(\cdot, \nu)$ is not lower hemi-continuous in general.

A direct corollary of Lemma C.II.5 is the upper hemi-continuity of Γ and G. For notational convenience, we introduce $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$.

Proposition C.II.1. The multi-map

$$\Gamma := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \rightrightarrows & \mathcal{P}(\mathcal{Z}) \\ \mu & \mapsto & \Gamma(\mu) \end{array} \right.$$

where $\Gamma(\mu)$ is defined in Eq. (C.II.8) and G defined in Eq. (C.II.10) are upper hemi-continuous (and compact-valued).

Proof. Let $\mu \in \mathcal{P}(\mathcal{X})$. To show that $G(\mu)$ and $\Gamma(\mu)$ are compact, it suffices to show that $\Gamma(\mu)$ is closed, since $\mathcal{P}(\mathcal{Z})$ is compact, and $G(\mu) = B\#\Gamma(\mu)$ with B continuous by Lemma C.II.1. Take $(\gamma_n) \in \Gamma(\mu)^{\mathbb{N}}$ such that $\gamma_n \xrightarrow[n\longrightarrow+\infty]{} \gamma \in \mathcal{P}(\mathcal{Z})$. We show that $\gamma \in \Gamma(\mu)$. For $k \in [\![1, K]\!]$ and $n \in \mathbb{N}$, we have $\gamma_n \in \Gamma(\mu)$, hence $[\gamma_n]_{0,k} \in \Pi_{c_k}^*(\mu, \nu_k)$. By continuity of marginalisation, we deduce that $\gamma \in \Pi(\mu, \nu_1, \cdots, \nu_K)$. By continuity of $\pi \longmapsto \int_{\mathcal{X}\times\mathcal{Y}_k} c_k \mathrm{d}\pi$ (which holds by compactness), we deduce that $\gamma_{0,k} \in \Pi_{c_k}^*(\mu, \nu_k)$, hence $\gamma \in \Gamma(\mu)$.

For the u.h.c. of Γ, take a sequence $(\mu_n) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$ such that $\mu_n \xrightarrow[n\longrightarrow+\infty]{w} \mu \in \mathcal{P}(\mathcal{X})$, and take a sequence $(\gamma_n) \in \mathcal{P}(\mathcal{Z})^{\mathbb{N}}$, with $\gamma_n \in \Gamma(\mu_n)$. Since $\gamma_n \in \mathcal{P}(\mathcal{Z})$ which is compact, take α an extraction such that $\gamma_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \gamma \in \mathcal{P}(\mathcal{Z})$. We will show that $\gamma \in \Gamma(\mu)$.

Start with $k := 1$. For $n \in \mathbb{N}$, we have $\gamma_{\alpha(n)} \in \Gamma(\mu_{\alpha(n)})$, hence $\pi^{(1)}_{\alpha(n)} := [\gamma_{\alpha(n)}]_{0,1} \in \Pi_{c_1}^*(\mu_{\alpha(n)}, \nu_1)$. By Lemma C.II.5, the map $\mu \longmapsto \Pi_{c_1}^*(\mu, \nu_1)$ is u.h.c., hence by definition, since $\mu_{\alpha(n)} \xrightarrow[n\longrightarrow+\infty]{w} \mu \in \mathcal{P}(\mathcal{X})$ and $\pi^{(1)}_{\alpha(n)} \in \Pi_{c_1}^*(\mu_{\alpha(n)}, \nu_1)$, there exists an extraction α_1 such that $\pi^{(1)}_{\alpha\circ\alpha_1(n)} \xrightarrow[n\longrightarrow+\infty]{w} \pi^{(1)} \in \Pi_{c_1}^*(\mu, \nu_1)$.

Continuing this method for $k \in [\![2, K]\!]$ with successive sub-extractions α_k, setting $\beta := \alpha \circ \alpha_1 \circ \cdots \circ \alpha_K$, we have for any $k \in [\![1, K]\!]$, $[\gamma_{\beta(n)}]_{0,k} = \pi^{(k)}_{\beta(n)} \xrightarrow[n\longrightarrow+\infty]{w} \pi^{(k)} \in \Pi_{c_k}^*(\mu, \nu_k)$. The continuity of marginalisation implies $\gamma_{0,k} = \pi^{(k)}$, and in turn shows that $\gamma \in \Gamma(\mu)$, concluding that Γ is u.h.c.

For G, the fact that $G(\mu) = B\#\Gamma(\mu)$ and the continuity of B prove that G is u.h.c. using the u.h.c. of Γ by [AB94, Theorem 17.23]. □

In order to study the energy of iterates of G, we first require a technical result on the error of sub-optimal ground barycentres for B. We introduce a radius constant $R := \max_{(x,Y)\in\mathcal{X}\times\mathcal{Y}} d_{\mathcal{X}}(x, B(Y))$, which is finite since $\mathcal{X}$ and $\mathcal{Y}$ are compact, and B is continuous. We need to make a trivial assumption to ensure that $R > 0$:

Assumption C.II.3. There exists $x \in \mathcal{X}$ and $Y \in \mathcal{Y}$ such that $x \neq B(Y)$.

Lemma C.II.6 is a generalisation of the following elementary Euclidean property in $\mathbb{R}^d$ for the cost $\|\cdot - \cdot\|_2^2$, for which $B(y_1, \cdots, y_K) = \sum_{k=1}^K \lambda_k y_k$ verifies the following identity:

$$\forall x \in \mathbb{R}^d,\ \forall (y_1, \cdots, y_K) \in (\mathbb{R}^d)^K,\ \overline{x} := \sum_{k=1}^K \lambda_k y_k :\ \sum_{k=1}^K \lambda_k \|x - y_k\|_2^2 = \sum_{k=1}^K \lambda_k \|\overline{x} - y_k\|_2^2 + \|x - \overline{x}\|_2^2.$$

Lemma C.II.6. There exists a function $\delta = \eta \circ d_{\mathcal{X}}$, with $\eta : [0, R] \longrightarrow \mathbb{R}_+$ lower-semi-continuous, non-decreasing and verifying $\eta(s) = 0 \iff s = 0$, such that

$$\forall (x, Y) \in \mathcal{X} \times \mathcal{Y},\ C(x, Y) \geq C(B(Y), Y) + \delta(x, B(Y)). \tag{C.II.13}$$

Proof. — *Step 1*: Definition of η. First, for $(x, Y) \in \mathcal{X} \times \mathcal{Y}$, let $\Delta(x, Y) := C(x, Y) - C(B(Y), Y)$. By definition of B, $\Delta(x, Y) \geq 0$, and $\Delta(x, Y) = 0 \iff x = B(Y)$. By assumption, B and C are continuous, which implies that Δ is also continuous.

We now introduce $S := \max_{(x,Y) \in \mathcal{X} \times \mathcal{Y}} \Delta(x, Y)$. Assumption C.II.3 ensures $S > 0$. Define now the function η:

$$\eta := \left\{ \begin{array}{ccc} [0, R] & \longrightarrow & [0, S] \\ u & \longmapsto & \min\limits_{(x,Y) \in \mathcal{X} \times \mathcal{Y}} \{\Delta(x, Y)\ :\ d_{\mathcal{X}}(x, B(Y)) \geq u\} \end{array} \right. . \tag{C.II.14}$$

We show that for $u \in [0, R]$, the infimum is attained. First, let $f := (x, Y) \longmapsto d_{\mathcal{X}}(x, B(Y))$, we remark that

$$\forall (x, Y) \in \mathcal{X} \times \mathcal{Y},\ d_{\mathcal{X}}(x, B(Y)) \geq u \iff (x, Y) \in f^{-1}([u, R]).$$

By continuity of f and compactness of $\mathcal{X} \times \mathcal{Y}$, $\mathcal{K}_u := f^{-1}([u, R])$ is a compact subset of $\mathcal{X} \times \mathcal{Y}$. $\mathcal{K}_u$ is not empty since there exists $(x_R, Y_R) \in \mathcal{X} \times \mathcal{Y}$ such that $d_{\mathcal{X}}(x_R, B(Y_R)) = R$ (by continuity, compactness and definition of R).

— *Step 2*: Proof of Eq. (C.II.13). Let $(x, Y) \in \mathcal{X} \times \mathcal{Y}$, and $u := d_{\mathcal{X}}(x, B(Y))$. By definition, $(x, Y) \in \mathcal{K}_u$, hence $\eta(u) \leq \Delta(x, Y)$, which is equivalent to Eq. (C.II.13).

— *Step 3*: Lower semi-continuity of η. Let $u_n \xrightarrow[n \longrightarrow +\infty]{} u \in [0, R]$, and for $n \in \mathbb{N}$ introduce $(x_n, Y_n) \in \mathcal{K}_{u_n}$ such that $\eta(u_n) = \Delta(x_n, Y_n)$. Since $(\eta(u_n)) \in [0, S]^{\mathbb{N}}$, consider an extraction α such that $\eta(u_{\alpha(n)}) \xrightarrow[n \longrightarrow +\infty]{} a_\alpha \in [0, S]$. By compactness of $\mathcal{X} \times \mathcal{Y}$, we can extract from $(x_{\alpha(n)}, Y_{\alpha(n)})_n$ a subsequence such that $(x_{\alpha\circ\beta(n)}, Y_{\alpha\circ\beta(n)}) \xrightarrow[n \longrightarrow +\infty]{} (x_{\alpha,\beta}, Y_{\alpha,\beta}) \in \mathcal{X} \times \mathcal{Y}$. By construction of the sequence $(x_n, Y_n)_n$, we have

$$\forall n \in \mathbb{N},\ d_{\mathcal{X}}(x_{\alpha\circ\beta(n)}, B(Y_{\alpha\circ\beta(n)})) \geq u_{\alpha\circ\beta(n)}, \tag{C.II.15}$$

since $(x_{\alpha\circ\beta(n)}, Y_{\alpha\circ\beta(n)}) \in \mathcal{K}_{u_{\alpha\circ\beta(n)}}$. Taking the limit in Eq. (C.II.15) yields $d_{\mathcal{X}}(x_{\alpha,\beta}, B(Y_{\alpha,\beta})) \geq u$, by continuity of B, Lemma C.II.1. This shows that $(x_{\alpha,\beta}, Y_{\alpha,\beta}) \in \mathcal{K}_u$, hence $\eta(u) \leq \Delta(x_{\alpha,\beta}, Y_{\alpha,\beta})$. However, by continuity of Δ, and since $\Delta(x_{\alpha(n)}, Y_{\alpha(n)}) \xrightarrow[n \longrightarrow +\infty]{} a_\alpha$, it follows that $\Delta(x_{\alpha,\beta}, Y_{\alpha,\beta}) = a_\alpha$. Since the subsequential limit a_α was chosen arbitrarily, we conclude that $\eta(u) \leq \liminf\limits_{n \longrightarrow +\infty} \eta(u_n)$, hence η is lower semi-continuous.

— *Step 4*: η is non-decreasing. Let $0 \leq u \leq v \leq R$, we have $\mathcal{K}_v \subset \mathcal{K}_u$, hence

$$\eta(u) = \min_{(x,Y) \in \mathcal{K}_u} \Delta(x, Y) \leq \min_{(x,Y) \in \mathcal{K}_v} \Delta(x, Y) = \eta(v).$$

— *Step 5*: Separation property. Let $u \in [0, R]$ such that $\eta(u) = 0$. This implies that there exists $(x, Y) \in \mathcal{X} \times \mathcal{Y}$ such that $\Delta(x, Y) = 0$ and $d_{\mathcal{X}}(x, B(Y)) \geq u$. Now by Step 1 this implies $x = B(Y)$, thus $d_{\mathcal{X}}(x, B(Y)) = 0$ and finally $u = 0$. □

Given the inequality in Eq. (C.II.13), we can now find an informative inequality between $V(\overline{\mu})$ and $V(\mu)$ for any $\overline{\mu} \in G(\mu)$. Applying Proposition C.II.2 to the W_2 case for absolutely continuous measures yields [Álv+16, Proposition 4.3], wherein the cost $\mathcal{T}_\delta$ is simply W_2^2. This decrease was also studied by [Lin23, Proposition 4.4] in the discrete setting W_p^p.

Proposition C.II.2. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\overline{\mu} \in G(\mu)$. Then $V(\mu) \geq V(\overline{\mu}) + \mathcal{T}_\delta(\mu, \overline{\mu})$. If μ^* is a barycentre, then $G(\mu^*) = \{\mu^*\}$.

Proof. Let $\overline{\mu} = B\#\gamma \in G(\mu)$ with $\gamma \in \Gamma(\mu)$, by definition of $\mathcal{T}_{c_k}$ and by optimality of the bi-marginals $\gamma_{0,k}$ of γ:

$$\sum_{k=1}^{K} \mathcal{T}_{c_k}(\mu, \nu_k) = \int_{\mathcal{X}\times\mathcal{Y}} C(x,Y)\mathrm{d}\gamma(x,Y) \tag{C.II.16}$$

$$\geq \int_{\mathcal{X}\times\mathcal{Y}} (C(B(Y),Y) + \delta(x,B(Y)))\mathrm{d}\gamma(x,Y) \tag{C.II.17}$$

$$\geq \sum_{k=1}^{K} \mathcal{T}_{c_k}(B\#\gamma, \nu_k) + \mathcal{T}_\delta(\mu, B\#\gamma) \tag{C.II.18}$$

$$= V(\overline{\mu}) + \mathcal{T}_\delta(\mu, \overline{\mu}). \tag{C.II.19}$$

The inequality in Eq. (C.II.17) comes from Lemma C.II.6, and the inequality in Eq. (C.II.18) comes from the definition of $\Gamma(\mu)$ (Eq. (C.II.8)), which allows us to write for $k \in [\![1, K]\!]$:

$$\int_{\mathcal{X}\times\mathcal{Y}} c_k(B(Y), y_k)\mathrm{d}\gamma(x,Y) = \int_{\mathcal{X}\times\mathcal{Y}_k} c_k \mathrm{d}\pi_k,$$

where we introduce the coupling $\pi_k := (B, P_k)\#[\gamma_{1,\cdots,K}]$, with $P_k(y_1, \cdots, y_K) = y_k$. The first marginal of π is $B\#[\gamma_{1,\cdots,K}]$ (which we write $B\#\gamma$ for legibility), and the second marginal is ν_k. Similarly,

$$\int_{\mathcal{X}\times\mathcal{Y}} \delta(x, B(Y))\mathrm{d}\gamma(x,Y) = \int_{\mathcal{X}\times\mathcal{X}} \delta \mathrm{d}[(I,B)\#\gamma] \geq \mathcal{T}_\delta(\mu, B\#\gamma).$$

If μ^* is a barycentre, then by definition for any $\overline{\mu} \in G(\mu)$, we have $V(\overline{\mu}) \geq V(\mu^*)$, thus Eqs. (C.II.17) and (C.II.18) are equalities, and $\mathcal{T}_\delta(\mu^*, \overline{\mu}) = 0$. By Lemmas C.II.2 and C.II.6, the cost δ guarantees the separation property of the transport cost $\mathcal{T}_\delta$, hence $\mu^* = \overline{\mu}$. □

The inequality in Proposition C.II.2 shows that the amount of decrease in the energy between two iterations is lower-bounded by a transport discrepancy $\mathcal{T}_\delta$ (we remind that in the squared-Euclidean case, $\mathcal{T}_\delta = \mathrm{W}_2^2$). We can now show convergence of iterates of G, in the sense that any weakly converging subsequence converges towards a fixed point of G.

Theorem C.II.1. For any $\mu_0 \in \mathcal{P}(\mathcal{X})$, let (μ_t) verifying $\mu_{t+1} \in G(\mu_t)$. Then (μ_t) has converging subsequences, and any weakly converging subsequence necessarily converges towards a $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mu \in G(\mu)$.

Proof. Fix a sequence (μ_t) such that $\mu_{t+1} \in G(\mu_t)$ and write $\mu_{t+1} = B\#[\gamma_t]_{1,\cdots,K}$ with $\gamma_t \in \Gamma(\mu_t)$. Since $\mathcal{X}$ is compact, the space $\mathcal{P}(\mathcal{X})$ is also compact, and so the sequence (μ_t) is tight. Consider an extraction α such that $\mu_{\alpha(t)} \xrightarrow[t\longrightarrow+\infty]{w} \mu \in \mathcal{P}(\mathcal{X})$. By u.h.c. of Γ (Proposition C.II.1), there exists an extraction β such that $\gamma_{\alpha\circ\beta(t)} \xrightarrow[t\longrightarrow+\infty]{w} \gamma \in \Gamma(\mu)$.

By Proposition C.II.2, the sequence $(V(\mu_t))$ is non-increasing and non-negative, hence it is convergent, imposing $\lim_{t\longrightarrow+\infty} \left[V(\mu_{\alpha\circ\beta(t)}) - V(\mu_{\alpha\circ\beta(t)+1})\right] = 0$. Using the lower-bound in Proposition C.II.2 we obtain:

$$\forall t \in \mathbb{N},\ 0 \leq \mathcal{T}_\delta(\mu_{\alpha\circ\beta(t)}, \mu_{\alpha\circ\beta(t)+1}) \leq V(\mu_{\alpha\circ\beta(t)}) - V(\mu_{\alpha\circ\beta(t)+1}),$$

and take the limit inferior:

$$0 \leq \liminf_{t\longrightarrow+\infty} \mathcal{T}_\delta(\mu_{\alpha\circ\beta(t)}, \mu_{\alpha\circ\beta(t)+1}) \leq 0. \tag{C.II.20}$$

We remind that $(\mu_{\alpha\circ\beta(t)+1})_t$ is a sequence in $\mathcal{P}(\mathcal{X})$ which is compact, and take $\rho \in \mathcal{P}(\mathcal{X})$ a subsequential limit of $(\mu_{\alpha\circ\beta(t)+1})_t$. By lower-semi-continuity of $\mathcal{T}_\delta$ (which holds by applying

Lemma C.II.2 item 1) with Lemma C.II.6), Eq. (C.II.20) provides $\mathcal{T}_\delta(\mu,\rho)=0$. By Lemma C.II.2 item 3), we obtain that $\rho=\mu$, thus any subsequential limit of $(\mu_{\alpha\circ\beta(t)+1})_t$ is μ, which proves that it converges weakly to μ.

Writing abusively $B\#\gamma$ for $B\#\gamma_{1,\cdots,K}$, we conclude:

$$\begin{array}{ccc} \mu_{\alpha\circ\beta(t)+1} & \xrightarrow[t\longrightarrow+\infty]{w} & \mu \\ \| & & \| \\ B\#\gamma_{\alpha\circ\beta(t)} & \xrightarrow[t\longrightarrow+\infty]{w} & B\#\gamma \end{array}$$

hence we have found $\gamma\in\Gamma(\mu)$ such that $\mu=B\#\gamma$, proving $\mu\in G(\mu)$. □

Fixed-points of G may not be unique and may not be barycentres, as shown the the following example. Take the following measures:

$$\mu:=\frac{1}{2}\left(\delta_{(0,1)}+\delta_{(0,-1)}\right),\ \nu_1:=\frac{1}{2}\left(\delta_{(-2,1)}+\delta_{(2,-1)}\right),\ \nu_2:=\frac{1}{2}\left(\delta_{(-2,-1)}+\delta_{(2,1)}\right).$$

Between μ and ν_k, the unique OT plan for the squared-Euclidean cost is given by a permutation, with

$$\pi_1^*=\frac{1}{2}\left(\delta_{(0,1)\otimes(-2,1)}+\delta_{(0,-1)\otimes(2,-1)}\right),$$

and likewise for π_2^*. The next iterate of G and H are both equal to μ itself, which is distinct from the unique barycenter $\mu^*=\frac{1}{2}(\delta_{(-2,0)}+\delta_{(2,0)})$. We show this example in Fig. C.II.4.

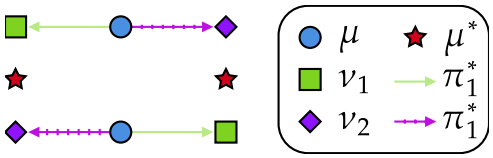


Figure C.II.4: Example showing non-barycentre measure μ which is a fixed-point of G and H.

C.II.3.3 Expression of the Iterates when the Plans are Maps

In some cases, the plans introduced in $\Gamma(\mu)$ (Eq. (C.II.8)) are induced by maps, which is to say that they are each supported on a set of the form $(x,T_k(x))$. This is the case in the specific setting chosen by [Álv+16], which is to say that all measures are absolutely continuous on $\mathbb{R}^d$ and the costs are all $c(x,y)=\|x-y\|_2^2$. By Brenier's Theorem (as stated in [San15, Theorem 1.22], for example), this implies that optimal transport couplings are supported on the graph of a map. This property holds under the weaker condition that the costs verify the Twist condition (see [Vil09, Theorem 10.28] for example). In this case, each set optimal transport plans $\Pi_{c_k}^*(\mu,\nu_k)$ is composed of one element $(I,T_k)\#\mu$, and as a result, the expression of $G(\mu)$ becomes substantially simpler, namely $G(\mu)=\{B\circ(T_1,\cdots,T_K)\#\mu\}$. In the linearisation interpretation (Fig. C.II.1), this expression can be understood as taking the ground barycentre of the maps T_k using the ground map B.

Drawing inspiration from this observation, we can define an alternative iteration consisting in choosing a map T_k as the barycentric projection of the coupling $\gamma_{0,k}\in\Pi_{c_k}^*(\mu,\nu_k)$ for $\gamma\in\Gamma(\mu)$: see Definition C.II.4 and Fig. C.II.5.

Definition C.II.4. The **barycentric projection** of a coupling $\pi\in\Pi(\mu,\nu)$ for $\mu\in\mathcal{P}(E)$ and $\nu\in\mathcal{P}(F)$ is the map $\overline{\pi}:E\longrightarrow F$, which is defined for μ-almost-every $x\in E$ as:

$$\overline{\pi}(x)=\int_F y\pi_x(\mathrm{d}y),$$

where we wrote the disintegration $\pi(\mathrm{d}x,\mathrm{d}y)=\mu(\mathrm{d}x)\pi_x(\mathrm{d}y)$. In terms of random variables,

one may write this expression as:

$$\overline{\pi}(x) = \mathop{\mathbb{E}}_{(X,Y)\sim\pi}[Y \mid X = x].$$

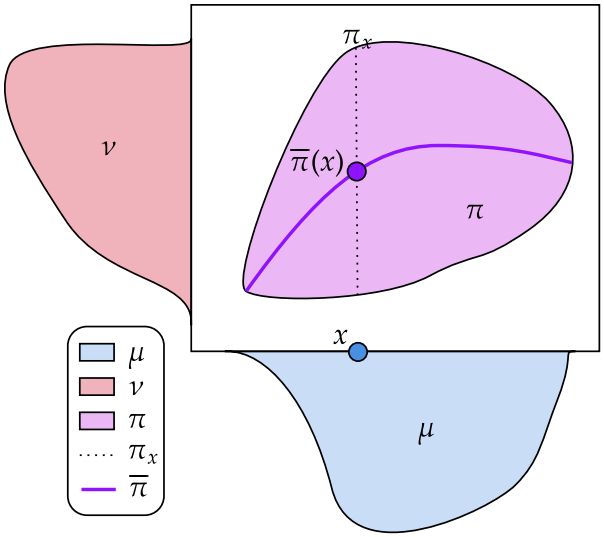


Figure C.II.5: Illustration of a barycentric projection. The disintegration of the coupling π with respect to its first marginal μ at x is the measure π_x concentrated on the dotted line. The barycentric projection of π evaluated at x is the mean of the measure π_x.

Note that for this expression to be well-defined, the target space F must be a *convex space*, i.e. a space where one may define convex combinations of points (or, more precisely, expectations of probability measures). In the case $\mathcal{X} = \mathcal{Y}_1 = \cdots = \mathcal{Y}_K$, a meaningful choice of convex combination is the ground barycentre B. We can apply this barycentric projection idea to define an alternate multi-mapping $H : \mathcal{P}(\mathcal{X}) \rightrightarrows \mathcal{P}(\mathcal{X})$:

$$\forall \mu \in \mathcal{P}(\mathcal{X}),\ H(\mu) := \{B \circ (\overline{\gamma_{0,1}}, \cdots, \overline{\gamma_{0,K}})\#\mu,\ \gamma \in \Gamma(\mu)\}\,. \tag{C.II.21}$$

In general, for $\pi \in \Pi(\mu,\nu)$, $\overline{\pi}\#\mu \neq \nu$, hence one does not necessarily have $\forall \widetilde{\mu} \in H(\mu),\ V(\widetilde{\mu}) \leq V(\mu)$. However, if each $\Pi^*_{c_k}(\mu,\nu_k)$ are composed of plans supported by maps, then $H(\mu) = G(\mu)$. In the case of discrete measures and for the squared Euclidean cost, the iterations of H correspond to the approach proposed in [CD14, Algorithm 2].

A fixed-point of H and not of G. In this paragraph, we present a counter-example which we found with Nicolas Juillet. Consider the initial measure $\mu := \frac{1}{2}\left(\delta_{(0,1)} + \delta_{(0,-1)}\right)$ and the targets:

$$\nu_1 := \frac{1}{4}\left(\delta_{(-2,1)} + \delta_{(-3,1)} + \delta_{(-2,-1)} + \delta_{(-3,-1)}\right),\ \nu_2 := \frac{1}{4}\left(\delta_{(2,1)} + \delta_{(3,1)} + \delta_{(2,-1)} + \delta_{(3,-1)}\right).$$

For the cost $c(x,y) := \|x-y\|_2^2$, there is a unique OT plan between μ and each respective ν_k: denoting $\pi_k^* := \Pi^*(\mu,\nu_k)$, we have:

$$\pi_1^* = \frac{1}{4}\left(\delta_{(0,1)\otimes(-2,1)} + \delta_{(0,1)\otimes(-3,1)} + \delta_{(0,-1)\otimes(-2,-1)} + \delta_{(0,-1)\otimes(-3,-1)}\right),$$

and likewise for π_2^*. It follows that the next iteration of H is point-valued with $H(\mu) = \{\mu_H\}$, $\mu_H = \mu$. Regarding G, for the conditionally independent multi-coupling $\gamma \in \Gamma(\mu)$, we have:

$$\mu_G := B\#\gamma = \frac{1}{4}\left(\delta_{(0,1)} + \delta_{(0,-1)}\right) + \frac{1}{8}\left(\delta_{(-1,1)} + \delta_{(1,1)} + \delta_{(-1,-1)} + \delta_{(1,-1)}\right).$$

We summarise these results in Fig. C.II.6.

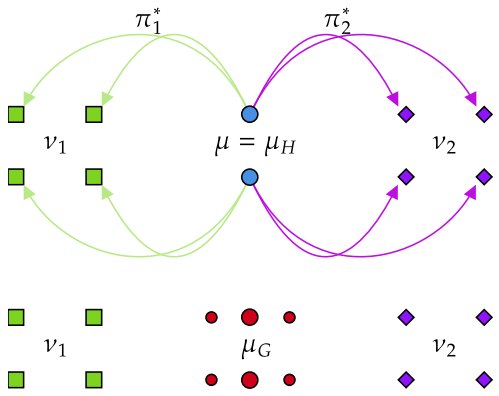


Figure C.II.6: Example showing non-barycentre measure μ which is a fixed-point of H but not G.

The measure μ is different from the unique barycentre

$$\mu^* = \frac{1}{4}\left(\delta_{(-1,1)} + \delta_{(1,1)} + \delta_{(-1,-1)} + \delta_{(1,-1)}\right),$$

however one has the relationship $\mu_G = \frac{1}{2}\mu + \frac{1}{2}\mu^*$. If one were to choose a different multi-coupling $\gamma \in \Gamma(\mu)$, one would obtain a different iterate, for instance consider the coupling $\gamma' \in \Gamma(\mu)$ which is such that for $(X, Y_1, Y_2) \sim \gamma'$, the couplings (X, Y_1) and (X, Y_2) are opposed, more precisely, we assume the following almost-sure conditional equalities for $\varepsilon \in \pm 1$:

$$\begin{aligned} Y_1 | \, (X = (0,\varepsilon), Y_2 = (2,\varepsilon)) &= (-3,\varepsilon), \quad Y_1 | \, (X = (0,\varepsilon), Y_2 = (3,\varepsilon)) = (-2,\varepsilon), \\ Y_2 | \, (X = (0,\varepsilon), Y_1 = (-2,\varepsilon)) &= (3,\varepsilon), \quad Y_2 | \, (X = (0,\varepsilon), Y_1 = (-3,\varepsilon)) = (2,\varepsilon), \end{aligned}$$

then the next iteration is $\mu'_G := B\#\gamma' = \mu^*$. If one takes yet another coupling γ'' which is this time symmetrical, in the sense that $(X, Y_1, Y_2) \sim \gamma''$ and $\varepsilon \in \pm 1$:

$$\begin{aligned} Y_1 | \, (X = (0,\varepsilon), Y_2 = (2,\varepsilon)) &= (-2,\varepsilon), \quad Y_1 | \, (X = (0,\varepsilon), Y_2 = (3,\varepsilon)) = (-3,\varepsilon), \\ Y_2 | \, (X = (0,\varepsilon), Y_1 = (-2,\varepsilon)) &= (2,\varepsilon), \quad Y_2 | \, (X = (0,\varepsilon), Y_1 = (-3,\varepsilon)) = (3,\varepsilon), \end{aligned}$$

then the next iterate is $\mu''_G := B\#\gamma'' = \mu$.

C.II.3.4 Extension to the Entropic Case

In this section, we explain how our results from Section C.II.3.2 extend to Entropic-Regularised Optimal transport, wherein we introduce for a regularisation $\varepsilon > 0$:

$$\mathcal{T}_{c,\varepsilon}(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{E\times F} c \mathrm{d}\pi + \varepsilon \, \mathrm{KL}(\pi | \mu \otimes \nu), \quad V_\varepsilon := \sum_{k=1}^{K} \mathcal{T}_{c_k,\varepsilon}(\mu,\nu_k). \tag{C.II.22}$$

Strict convexity of the KL divergence yields existence and uniqueness of entropic optimal transport plans (denoted $\Pi^*_{c,\varepsilon}(\mu,\nu)$), and by [GNB22, Theorem 1.4], the (single-valued) map $\mu \longmapsto \Pi^*_{c,\varepsilon}(\mu,\nu)$ is continuous for the weak convergence, provided that the cost c is continuous. Akin to the OT case, we define the map $\Gamma_\varepsilon : \mathcal{P}(\mathcal{X}) \longrightarrow \mathcal{P}(\mathcal{Z})$ by:

$$\Gamma_\varepsilon(\mu) := \mu(\mathrm{d}x)\pi^x_{1,\varepsilon}(\mathrm{d}y_1)\cdots\pi^x_{K,\varepsilon}(\mathrm{d}y_K), \ \forall k \in [\![1,K]\!], \ \pi_{k,\varepsilon} := \Pi^*_{c_k,\varepsilon}(\mu,\nu_k), \tag{C.II.23}$$

and the iteration functional $G_\varepsilon(\mu) := B\#\Gamma_\varepsilon(\mu)$. Using Lemma C.II.6 and some technical manipulations of the KL divergence, we adapt Proposition C.II.2 to this entropic case.

Proposition C.II.3. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\overline{\mu} := G_\varepsilon(\mu)$. Then $V_\varepsilon(\mu) \geq V_\varepsilon(\overline{\mu}) + \mathcal{T}_\delta(\mu, \overline{\mu})$. If μ^* is a barycentre, then $G_\varepsilon(\mu^*) = \mu^*$.

Proof. We begin as in Proposition C.II.2, with $\gamma := \Gamma_\varepsilon(\mu)$:

$$\sum_{k=1}^K \mathcal{T}_{c_k,\varepsilon}(\mu, \nu_k) = \int_{\mathcal{X}\times\mathcal{Y}} C(x,Y)\mathrm{d}\gamma(x,Y) + \varepsilon \sum_{k=1}^K \mathrm{KL}(\gamma_{0,k}|\mu\otimes\nu_k) \tag{C.II.24}$$

$$\geq \int_{\mathcal{X}\times\mathcal{Y}} (C(B(Y),Y) + \delta(x,B(Y)))\mathrm{d}\gamma(x,Y) + \varepsilon \sum_{k=1}^K \mathrm{KL}(\gamma_{0,k}|\mu\otimes\nu_k). \tag{C.II.25}$$

For convenience, write $\gamma_\otimes := \mu\otimes\nu_1\otimes\cdots\otimes\nu_K$. Using the notation from Eq. (C.II.23), notice that $\dfrac{\mathrm{d}\gamma}{\mathrm{d}\gamma_\otimes} = \prod_{k=1}^K \dfrac{\mathrm{d}\pi_{k,\varepsilon}}{\mathrm{d}(\mu\otimes\nu_k)}$, which implies that $\sum_k \mathrm{KL}(\gamma_{0,k}|\mu\otimes\nu_k) = \mathrm{KL}(\gamma|\gamma_\otimes)$. Putting this with Eq. (C.II.25) yields

$$V_\varepsilon(\mu) \geq \sum_{k=1}^K \int_{\mathcal{X}\times\mathcal{Y}_k} c_k \mathrm{d}(B,P_k)\#\gamma + \varepsilon\,\mathrm{KL}(\gamma|\gamma_\otimes) + \int_{\mathcal{X}^2} \delta\mathrm{d}(I,B)\#\gamma. \tag{C.II.26}$$

Now let $f : \mathcal{X}\times\mathcal{Y}_1\times\cdots\times\mathcal{Y}_K \longrightarrow \mathcal{X}\times\mathcal{Y}_1\times\cdots\times\mathcal{Y}_K$ the continuous function defined by $f(x,y_1,\cdots,y_K) := (B(y_1,\cdots,y_K),y_1,\cdots,y_K)$. We apply the data processing inequality (use [PW14, Theorem 3.5], or apply [AGS05, Lemma 9.4.5]): $\mathrm{KL}(\gamma|\gamma_\otimes) \geq \mathrm{KL}(f\#\gamma|f\#\gamma_\otimes)$. Now we use the disintegration formula and the change-of-reference formula for KL. Notice that the first marginals of $f\#\gamma$ and $f\#\gamma_\otimes$ are both equal to $\overline{\mu}$, and that $(f\#\gamma)_{1,\cdots,K} \in \Pi(\nu_1,\cdots,\nu_K)$ and $(f\#\gamma_\otimes)_{1,\cdots,K} = \nu_1\otimes\cdots\otimes\nu_K$.

$$\begin{aligned}
\mathrm{KL}(f\#\gamma|f\#\gamma_\otimes) &= \mathrm{KL}((f\#\gamma)_0|(f\#\gamma_\otimes)_0) + \int_{\mathcal{X}} \mathrm{KL}((f\#\gamma)^x|\nu_1\otimes\cdots\otimes\nu_K)\mathrm{d}\overline{\mu}(x)\\
&= 0 + \int_{\mathcal{X}} \mathrm{KL}((f\#\gamma)^x|(f\#\gamma)^x_1\otimes\cdots\otimes(f\#\gamma)^x_K)\mathrm{d}\overline{\mu}(x)\\
&+ \mathrm{KL}((f\#\gamma)^x_1\otimes\cdots\otimes(f\#\gamma)^x_K|\nu_1\otimes\cdots\otimes\nu_K)\mathrm{d}\overline{\mu}(x)\\
&\geq \sum_{k=1}^K \int_{\mathcal{X}} \mathrm{KL}((f\#\gamma)^x_k|\nu_k)\mathrm{d}\overline{\mu}(x) = \sum_{k=1}^K \mathrm{KL}((f\#\gamma)_{0,k})|\mu\otimes\nu_k).
\end{aligned}$$

Now we notice that $(f\#\gamma)_{0,k} = (B,P_k)\#\gamma \in \Pi(\overline{\mu},\nu_k)$, which with Eq. (C.II.26) provides:

$$\begin{aligned}
V_\varepsilon(\mu) &\geq \sum_{k=1}^K \left(\int_{\mathcal{X}\times\mathcal{Y}_k} c_k\mathrm{d}(B,P_k)\#\gamma + \varepsilon\,\mathrm{KL}((B,P_k)\#\gamma|\mu\otimes\nu_k)\right) + \int_{\mathcal{X}^2}\delta\mathrm{d}(I,B)\#\gamma\\
&\geq \sum_{k=1}^K \mathcal{T}_{c,\varepsilon}(\overline{\mu},\nu_k) + \mathcal{T}_\delta(\mu,\overline{\mu}) = V_\varepsilon(\overline{\mu}) + \mathcal{T}_\delta(\mu,\overline{\mu}).
\end{aligned}$$

The rest of the proof follows as in Proposition C.II.2. □

From Proposition C.II.3, we deduce an adaptation of Theorem C.II.1 to the entropic case.

Theorem C.II.2. For any $\mu_0 \in \mathcal{P}(\mathcal{X})$, let (μ_t) verifying $\mu_{t+1} = G_\varepsilon(\mu_t)$. Then (μ_t) has converging subsequences, and any weakly converging subsequence necessarily converges towards a $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mu = G_\varepsilon(\mu)$.

Proof. The proof can be adapted from Theorem C.II.1 without difficulty, in particular given the fact that each $\mu \longmapsto \Pi^*_{c_k,\varepsilon}(\mu,\nu_k)$ is continuous with respect to the weak convergence of measures, which ensures that Γ_ε is also continuous. □

C.II.3.5 The Particular Case of Conditionally Independent Couplings

In Eq. (C.II.8), we chose all possible multi-couplings with optimal bi-marginals. It is possible to restrict the set of couplings to the smaller set of multi-couplings with conditionally independent marginals, i.e. multi-couplings $\gamma \in \Pi(\mu, \nu_1, \cdots, \nu_K)$ such that there exists $\pi_k \in \Pi^*_{c_k}(\mu, \nu_k)$ for $k \in [\![1, K]\!]$ such that $\gamma_{0,k} = \pi_k$ and specifically:

$$\gamma(\mathrm{d}x, \mathrm{d}y_1, \cdots, \mathrm{d}y_K) := \mu(\mathrm{d}x)\ \pi_1^x(\mathrm{d}y_1) \cdots \pi_K^x(\mathrm{d}y_K),$$

as in Eq. (C.II.9). In terms of random variables, this corresponds to the choice of $(X, Y_1, \cdots, Y_K)$ such that $(X, Y_k) \sim \pi_k$ and conditionally to X, the variables $Y_1, \cdots, Y_K$ are independent. We denote by $\Gamma_\otimes(\mu)$ the set of such couplings, and consider the associated multi-map $G_\otimes := B\#\Gamma_\otimes$ as in Eq. (C.II.10). It is clear that $\forall \mu \in \mathcal{P}(\mathcal{X}),\ G_\otimes(\mu) \subset G(\mu)$. In particular, this implies subsequential convergence converges of iterates $\mu_{t+1} = G_\otimes(\mu_t)$ to a fixed-point of G. In Proposition C.II.4, we show that the convergence is to a fixed-point of $G_\otimes$ in the discrete case (measures with finite support). First, we emphasise that with a discrete initialisation measure and discrete measures (ν_k), the support of the sequence (μ_t) is finite and always contained in:

$$\{B(y_1, \cdots, y_K),\ \forall k \in [\![1, K]\!],\ y_k \in \mathrm{supp}(\nu_k)\},$$

which ensures that iterates remain discrete.

Proposition C.II.4. Take $\mu_0 \in \mathcal{P}(\mathcal{X})$ a discrete measure and $\nu_1, \cdots \nu_k \in \mathcal{P}(\mathcal{Y}_1) \times \cdots \times \mathcal{P}(\mathcal{Y}_K)$ discrete measures. Then any sub-sequential limit $\mu \in \mathcal{P}(\mathcal{X})$ of the sequence (μ_t) defined by $\mu_{t+1} \in G_\otimes(\mu_t)$ verifies $\mu \in G_\otimes(\mu)$.

Proof. We follow a technique used in the proof of in [Goz+17, Theorem 9.6], specifically [Goz+17, page 65]. As commented before the statement of the result, the sequence (μ_t) remains discrete. Write for $t \in \mathbb{N},\ \mu_{t+1} = B\#\gamma_t$ with $\gamma_t \in \Gamma_\otimes(\mu_t)$, and take an extraction α such that $\mu_{\alpha(t)} \xrightarrow[t \longrightarrow +\infty]{w} \mu$. As done in Theorem C.II.1, the u.h.c. property of Γ allows us to extract a subsequence β such that $\gamma_{\alpha\circ\beta(t)} \xrightarrow[t \longrightarrow +\infty]{w} \gamma \in \Gamma(\mu)$, since we have the (point-wise) inclusion $\Gamma_\otimes \subset \Gamma$. As shown in the proof of Theorem C.II.1, the sequence $(\mu_{\alpha\circ\beta(t)})_t$ weakly converges to μ, hence we now want to show that $\gamma \in \Gamma_\otimes(\mu)$, which would allow to conclude $\mu \in G_\otimes(\mu)$.

For $t \in \mathbb{N}$ and $k \in [\![1, K]\!]$, introduce $\pi^{(k)}_{\alpha\circ\beta(t)} := [\gamma_{\alpha\circ\beta(t)}]_{0,k}$, and its disintegration with respect to $\mu_{\alpha\circ\beta(t)}$ as

$$\pi^{(k)}_{\alpha\circ\beta(t)}(\mathrm{d}x, \mathrm{d}y_k) = \mu_{\alpha\circ\beta(t)}(\mathrm{d}x)\ [\pi^{(k)}_{\alpha\circ\beta(t)}]^x(\mathrm{d}y_k).$$

As argued above the statement of the proposition, the sequence $(\mu_t)_t$ remains discrete with a support contained in $B(\prod_k \mathrm{supp}(\nu_k))$ and thus (γ_t) also remains discrete, and its first marginal μ_t has a finite support of size at most $n := \prod_k \#\,\mathrm{supp}(\nu_k)$ on fixed points $(x_1, \cdots, x_n)$. For simplicity, we will see the measures (μ_t) as supported on $\mathcal{X}_n := \{x_1, \cdots, x_n\}$ with possibly zero mass at some of these points, and in such cases, we define $[\pi^{(k)}_{\alpha\circ\beta(t)}]^x$ as the null measure $\mathcal{M}_0$. Since $\gamma_{\alpha\circ\beta(t)} \in \Gamma_\otimes(\mu_{\alpha\circ\beta(t)})$, by definition we can write its disintegration with respect to $\mu_{\alpha\circ\beta(t)}$ as:

$$\gamma_{\alpha\circ\beta(t)}(\mathrm{d}x, \mathrm{d}y_1, \cdots, \mathrm{d}y_K) = \mu_{\alpha\circ\beta(t)}(\mathrm{d}x)\ [\pi^{(1)}_{\alpha\circ\beta(t)}]^x(\mathrm{d}y_1) \cdots [\pi^{(K)}_{\alpha\circ\beta(t)}]^x(\mathrm{d}y_K).$$

For $i \in [\![1, n]\!]$ and $k \in [\![1, K]\!]$, there exists an extraction $\chi_{i,k}$ such that the sequence $\left([\pi^{(k)}_{\alpha\circ\beta\circ\chi_{i,k}(t)}]^{x_i}\right)_{t\in\mathbb{N}}$ converges weakly to a $[\pi^{(k)}]^{x_i} \in \mathcal{P}(\mathcal{X}) \times \{\mathcal{M}_0\}$. We choose the extractions as successive sub-extractions, such that $\chi_{1,2}$ is a sub-extraction of $\chi_{1,1}$, until $\chi_{n,K}$ which is a sub-extraction of all previous extractions. We then define $\chi := \chi_{n,K}$. The extraction χ is such that for $i \in [\![1, n]\!]$ and $k \in [\![1, K]\!]$, the sequence $\left([\pi^{(k)}_{\alpha\circ\beta\circ\chi(t)}]^{x_i}\right)_{t\in\mathbb{N}}$ converges weakly to $[\pi^{(k)}]^{x_i}$. By verifying against test functions, we deduce the following disintegration holds for γ:

$$\gamma(\mathrm{d}x, \mathrm{d}y_1, \cdots, \mathrm{d}y_K) = \mu(\mathrm{d}x)\ [\pi^{(1)}]^x(\mathrm{d}y_1) \cdots [\pi^{(K)}]^x(\mathrm{d}y_K),$$

which shows that $\gamma \in \Gamma_\otimes(\mu)$, and thus $\mu \in G_\otimes(\mu)$. □

Remark C.II.3. The proof of Proposition C.II.4 can also be written for discrete measures with at-most-countable supports through a diagonal extraction argument, we kept to finite supports for legibility.

C.II.4 Focus on the Discrete Case

In this section, we will formulate the fixed-point algorithm in the discrete case, and discuss some algorithmic aspects.

C.II.4.1 Discrete Expression and Algorithms

Consider discrete measures $\nu_k := \sum_{i=1}^{n_k} b_{k,i}\delta_{y_{k,i}} \in \mathcal{P}(\mathbb{R}^{d_k})$ where $\forall k \in [\![1, K]\!]$, $\forall i \in [\![1, n_k]\!]$, $y_{k,i} \in \mathbb{R}^{d_k}$. We stack the support of ν_k into $Y_k \in \mathbb{R}^{n_k \times d_k}$ such that $[Y_k]_{i,\cdot} = y_{k,i}$, and similarly introduce $b_k := (b_{k,i})_{i=1}^{n_k} \in \triangle_{n_k}$.

First, our objective is to re-write the iteration Eq. (C.II.10) in this discrete setting, with an initial measure $\mu = \sum_{i=1}^{n} a_i\delta_{x_i} \in \mathcal{P}(\mathbb{R}^d)$. For each k, we choose $\pi_k \in \mathbb{R}_+^{n\times n_k}$ an optimal transport plan, which is to say a solution of the Kantorovich linear program:

$$\operatorname*{argmin}_{\Pi(a,b_k)} \sum_{i=1}^{n}\sum_{j=1}^{n_k} c_k(x_i, y_{k,j})\pi_{i,j},$$

where $\Pi(a, b_k) := \left\{\pi \in \mathbb{R}_+^{n\times n_k} \;:\; \pi\mathbf{1} = a,\ \pi^\top\mathbf{1} = b_k\right\}$. Seeing multi-couplings $\gamma \in \Gamma(\mu)$ as a tensors $\gamma \in \mathbb{R}^{n\times n_1\times\cdots,n_k}$, the discrete expression of G reads:

$$G(\mu) = \left\{ \sum_{j_1,\cdots,j_K} \left(\sum_{i=1}^{n} \gamma_{1,j_1,\cdots,j_K}\right) \delta\left(B(y_{1,j_1},\cdots,y_{K,j_K})\right),\ \gamma \in \Gamma(\mu) \right\}. \tag{C.II.27}$$

A visualisation of Eq. (C.II.27) using the multi-coupling from Eq. (C.II.9) is provided in Fig. C.II.7.

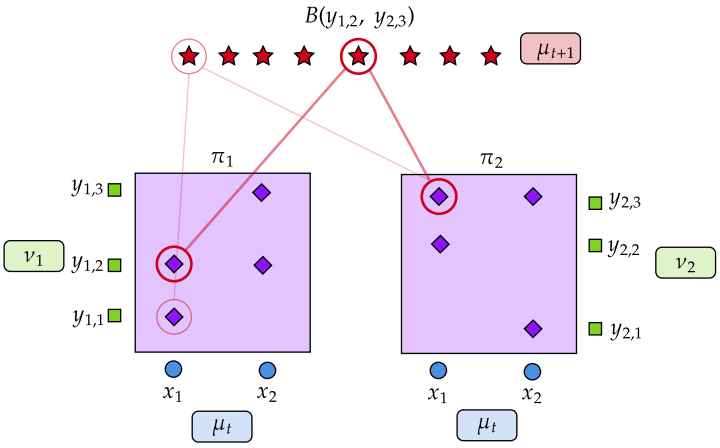


Figure C.II.7: Visual explanation of the discrete fixed-point iteration G. For each point x_i in the support of μ_t, we look at all the points $(y_{1,j_1},\cdots,y_{K,j_K})$ which are assigned from x_i by the multi-coupling γ, then the ground barycentre $B(y_{1,j_1},\cdots,y_{K,j_K})$ is taken on all these tuples, with weights given by the multi-coupling.

As in [BB25], we can use a generalisation of the North-West Corner (NWC) method to compute $\gamma \in \Gamma(\mu)$ with prescribed bi-marginals $\pi_k \in \Pi(a, b_k)$. In Algorithm C.II.1, we present the NWC strategy. The idea is to fill the entries of γ greedily using entries $\pi^{(k)}_{i,j_k}$ for increasing $i, j_1, \cdots, j_K$ (see [PC19b, Section 3.4.2] for a presentation of the method in the standard setting).

Algorithm C.II.1: North-West Corner Gluing.

Data: For $k \in [\![1, K]\!]$, transport plan $\pi_k \in \Pi(a, b_k)$, with $a \in \triangle_n$ and $b_k \in \triangle_{n_k}$.
Result: Gluing $\mathrm{NWC}(\pi_1, \cdots, \pi_K) = \gamma \in \Pi(a, b_1, \cdots, b_K)$ such that each $\gamma_{0,k} = \pi_k$.

Initialisation: $\gamma = 0_{n \times n_1 \times \cdots \times n_K}$ and for $k \in [\![1, K]\!]$, $P_k = \pi_k$.
for $i \in [\![1, n]\!]$ **do**
  Set $(j_1, \cdots, j_K) = (1, \cdots, 1)$ and $u = a_i$;
  while $u > 0$ **do**
    Compute $v = \min(P^{(1)}_{i,j_1}, \cdots, P^{(K)}_{i,j_K})$;
    Assign $\gamma_{i,j_1,\cdots,j_K} = v$ and decrease $u \leftarrow u - v$;
    for $k \in [\![1, K]\!]$ **do**
      Decrease $P^{(k)}_{i,j_k} \leftarrow P^{(k)}_{i,j_k} - v$;
      if $P^{(k)}_{i,j_k} = 0$ **then**
        Increment $j_k \leftarrow j_k + 1$;
      end
    end
  end
end

Noticing that Eq. (C.II.27) only requires the $n_1 \times \cdots \times n_K$-tensor $\rho := \gamma_{1,\cdots,K}$, it is possible to only store the indices $(j_1, \cdots, j_K)$ such that $\rho_{j_1,\cdots,j_K} > 0$, as well as the corresponding weights $\rho_{j_1,\cdots,j_K}$. This avoids the prohibitive memory cost of storing the full tensor γ, and takes advantage of the sparsity of the multi-coupling γ: if each π_k is an extremal point of $\Pi(a, b_k)$, we conjecture that $\mathrm{NWC}(\pi_1, \cdots, \pi_K)$ is an extremal point of $\Pi(a, b_1, \cdots, b_K)$, and thus $\#\operatorname{supp}\gamma \leq n + \sum_k n_k - K$ (adapting techniques from [ABM16, Theorem 2]).

Thanks to Eq. (C.II.27) we formalise the fixed-point iterations in the discrete case in Algorithm C.II.2.

Algorithm C.II.2: Discrete iteration of G.

Data: barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k:
$Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and cost function $c_k : \mathbb{R}^d \times \mathbb{R}^{d_k} \longrightarrow \mathbb{R}_+$.
Number of iterations T, initial size $n \geq 1$ and stopping criterion $\alpha \geq 0$.
Result: Barycentre $\mu_T = \sum_{i=1}^{N_t} a_i^{(T)} \delta_{x_i^{(T)}}$.

Initialisation: Choose $\mu_0 = \sum_{i=1}^{n} a_i^{(0)} \delta_{x_i^{(0)}}$ with $a^{(0)} \in \triangle_n$ and $X^{(0)} \in \mathbb{R}^{n \times d}$.
for $t \in [\![0, T-1]\!]$ **do**
  for $k \in [\![1, K]\!]$ **do**
    Solve the OT problem: $\pi^{(k)} \in \underset{\pi \in \Pi(a^{(t)}, b_k)}{\operatorname{argmin}} \sum_{i,j} \pi_{i,j} c_k(x_i^{(t)}, y_{k,j})$;
  end
  Compute $\gamma = \mathrm{NWC}(\pi^{(1)}, \cdots, \pi^{(K)})$;
  Compute $\rho = \gamma_{1,\cdots,K} = [\sum_i \gamma_{j_1,\cdots,j_K}]_{j_1,\cdots,j_k}$ and write
  $\operatorname{supp}\rho = \left((j_1^{(i)}, \cdots, j_K^{(i)})\right)_{i=1}^{N_t}$;
  for $i \in [\![1, N_t]\!]$ **do**
    Compute $x_i^{(t+1)} = B(y_{1,j_1^{(i)}}, \cdots, y_{K,j_K^{(i)}})$ and $a_i^{(t+1)} = \rho_{j_1,\cdots,j_K}$;
  end
  if $\mathrm{W}_2^2(\mu_{t+1}, \mu_t) < \frac{\alpha}{N_t} \|X^{(t)}\|_2^2$ **then**
    Declare convergence and terminate.
  end
end
return $a^{(T)}, X^{(T)}$

Given our considerations on the support of NWC gluing, we expect (without formal proof) the upper bound $\#\operatorname{supp}\mu_T \leq n + T(\sum_k n_k) - TK$. This is the same conclusion as [BB25], which they also state without proof about their Gromov-Wasserstein fixed-point iteration [BB25, Algorithm 5.2]. From a memory perspective, the algorithm does not require the storage of each $\gamma \in \mathbb{R}^{N_t \times n_1 \times \cdots \times n_K}$, as remarked for the NWC algorithm.

In some specific cases, the expression in Eq. (C.II.27) becomes simpler. If the weights a and b_k are all uniform and $n = n_1 = \cdots = n_K$, then the Birkhoff-von-Neumann Theorem allows the choice of each transport plan π_k as permutation assignments $[\pi_k]_{i,j} = \frac{1}{n}\mathbb{1}(\sigma_k(i) = j)$. In this case, the expression of $G(\mu)$ becomes:

$$G(\mu) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(B(y_{1,\sigma_1(i)}, \cdots, y_{K,\sigma_K(i)})\right). \tag{C.II.28}$$

If one takes the barycentric projections of the OT plans $\pi^{(k)}$ in Eq. (C.II.27), one obtains a discrete expression of H (from Eq. (C.II.21)) written in Eq. (C.II.29) and visualised in Fig. C.II.8.

$$H(\mu) = \left\{ \sum_{i=1}^{n} a_i \delta \left[B\left((1/a_i)\sum_{j=1}^{n_1} \pi^{(1)}_{i,j} y_{1,j}, \cdots, (1/a_i)\sum_{j=1}^{n_K} \pi^{(K)}_{i,j} y_{K,j} \right) \right],\ \pi^{(k)} \in \Pi^*_{c_k}(\mu, \nu_k) \right\}. \tag{C.II.29}$$

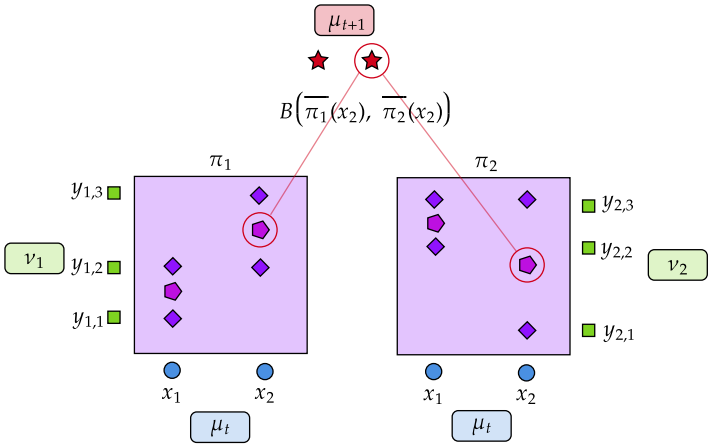


Figure C.II.8: Visual explanation of the discrete fixed-point iteration H. For each point x_i in the support of μ_t and $k \in [\![1, K]\!]$, we look at all the points (y_{k,j_k}) which are assigned from x_i by the OT plan $\pi^{(k)}$ between μ and ν_k, then take their barycenter $\overline{\pi}_k(x_i)$ (pentagons on the figure). The point x_i in the support of μ_t is then sent to the point $B(\overline{\pi}_1(x_i), \cdots, \overline{\pi}_K(x_i))$ in μ_{t+1}.

Contrary to G, for H the number of points in the support of μ_t remains the same, and the weights a remain fixed. In this setting, the optimisation is done solely on the positions, which can be seen as a Lagrangian formulation. Note that in the squared-Euclidean case, Eq. (C.II.29) is the formula proposed in [CD14, Equation 8] and currently implemented in the Python OT library [Fla+21]. A technical difference is that [CD14] also proposes an optimisation over the barycentre weights (by sub-gradient descent), while the fixed-point approach by [Álv+16] and ours do not. Furthermore, [CD14] suggests a computational simplification by using barycentric projections of *entropic* plans (as in Section C.II.3.4), for which, as for H, there are no theoretical guarantees (to our knowledge).

The practical advantage of the map-supported expressions in Eqs. (C.II.28) and (C.II.29) over Eq. (C.II.27) is computational: since the support size of μ_t cannot increase, the cost of computing the OT plans at Line 4 is smaller. We shall see in Section C.II.4.3 that in some cases, Kantorovich solutions are almost-surely permutations for random supports. While convenient, this expression only holds when all the measures have the same amount of points, in contrast to the barycentric expression Eq. (C.II.29).

Algorithm C.II.3: Discrete iteration of H.

Data: barycentre coefficients $(\lambda_k) \in \triangle_K$, for $k \in [\![1, K]\!]$, support of ν_k: $Y_k \in \mathbb{R}^{n_k \times d_k}$, weights of ν_k: $b_k \in \triangle_{n_k}$ and cost function $c_k : \mathbb{R}^d \times \mathbb{R}^{d_k} \longrightarrow \mathbb{R}_+$. Number of iterations T, barycentre size $n \geq 1$, weights $a \in \triangle_n$ and stopping criterion $\alpha \geq 0$.

Result: Barycentre $\mu_T = \sum_{i=1}^n a_i \delta_{x_i^{(T)}}$.

Initialisation: Choose $\mu_0 = \sum_{i=1}^n a_i \delta_{x_i^{(0)}}$ with $X^{(0)} \in \mathbb{R}^{n \times d}$.
for $t \in [\![0, T-1]\!]$ **do**
  for $k \in [\![1, K]\!]$ **do**
    Solve the OT problem: $\pi^{(k)} \in \underset{\pi \in \Pi(a, b_k)}{\operatorname{argmin}} \sum_{i,j} \pi_{i,j} c_k(x_i^{(t)}, y_{k,j})$;
  end
  for $i \in [\![1, n]\!]$ **do**
    Compute $x_i^{(t+1)} = B\left((1/a_i)\sum_{j=1}^{n_1} \pi_{i,j}^{(1)} y_{1,j}, \cdots, (1/a_i)\sum_{j=1}^{n_K} \pi_{i,j}^{(K)} y_{K,j}\right)$;
  end
  if $\mathrm{W}_2^2(\mu_{t+1}, \mu_t) < \frac{\alpha}{n}\|X^{(t)}\|_2^2$ **then**
    Declare convergence and terminate.
  end
end
return $X^{(T)}$

We present the iteration of H as a cost-effective alternative to G, which is in some sense a simplification of the Block-Coordinate Descent (BCD) method, wherein the update with respect to the support $X \in \mathbb{R}^{n \times d}$ with transport plans (π_k) fixed is done by computing:

$$X^* \in \underset{X \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} \sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^{n_k} \pi_{i,j}^{(k)} c_k(x_i, y_{k,j}). \tag{C.II.30}$$

In practice, apart from the case of the squared Euclidean cost, the optimisation in Eq. (C.II.30) is not tractable, and one must resort to Gradient Descent (GD) methods. BCD methods with GD for the update of X can be seen as a variant of the full GD method which minimises $X \longmapsto V(\frac{1}{n}\sum_i \delta_{x_i})$, and we leave their study for future work.

C.II.4.2 Correspondence of Gradient Descent with Fixed-Point Iterations

The fixed-point method of [Álv+16] applied to Bures-Wasserstein barycentres also corresponds to a gradient descent algorithm with a specific step size, as remarked by [Alt+21]. This also holds for discrete measures. Indeed, writing $X = \{x_1, \cdots, x_n\}$ and assuming $\mu_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, an alternative to fixed-point iterations would be to apply a gradient descent directly on the non convex functional $F : X \longmapsto \sum_{k=1}^K \lambda_k \mathcal{T}_{c_k}(\mu_X, \nu_k)$. For differentiable costs c_k, assuming that $\nu_k = \frac{1}{n}\sum_{i=1}^n \delta_{y_{k,i}}$, one step of such a gradient descent writes

$$\forall i \in [\![1, n]\!],\ x_i^{(t+1)} = x_i^{(t)} - \alpha \sum_{k=1}^K \lambda_k \nabla_x c_k(x_i^{(t)}, y_{k,\sigma_k^{(t)}(i)}), \tag{C.II.31}$$

where we choose an element of $\Pi_{c_k}^*(\mu_{X^{(t)}}, \nu_k)$ induced by a permutation $\sigma_k^{(t)}$ between $\{x_1^{(t)}, \cdots, x_n^{(t)}\}$ and $\{y_{k,1}, \cdots, y_{k,n}\}$. The whole optimisation algorithm consists in alternating such gradient steps on X with updates of the optimal assignments $\{\sigma_k^{(t)}\}$, depending on the new point positions. In the fixed-point approach, this gradient step on each $x_i^{(t)}$ is replaced by the computation of $B(y_{1,\sigma_1^{(t)}(i)}, \cdots, y_{K,\sigma_K^{(t)}(i)})$, which corresponds to a full descent on X for a given configuration of assignments before updating the said assignments (in other words, alternate minimisation). For generic costs c_k, one may also use a gradient descent strategy to compute

barycentres $B(y_{1,\sigma_1^{(t)}(i)}, \cdots, y_{K,\sigma_K^{(t)}(i)})$, that is gradient descents on the K functionals $x \longmapsto \sum_{k=1}^K c_k(x, y_{k,\sigma_k^{(t)}(i)})$, and such descents write exactly as Eq. (C.II.31). In this case, the only difference between both approaches is that the fixed point algorithm applies the whole descent on X before updating assignments, while gradient descent on F alternates steps of gradient descent on X with updates of the assignments.

When $c_k = \|\cdot - \cdot\|_2^2$, both approaches are equivalent if the gradient step is chosen as $\alpha = \frac{1}{2}$. Indeed, a gradient iteration on F writes

$$\forall i \in [\![1, n]\!],\ x_i^{(t+1)} = (1-2\alpha)x_i^{(t)} + 2\alpha \sum_{k=1}^K \lambda_k y_{k,\sigma_k^{(t)}(i)} = \sum_{k=1}^K \lambda_k y_{k,\sigma_k^{(t)}(i)}.$$

It follows that for $\alpha = \frac{1}{2}$, one step of gradient descent computes directly the barycentre for the current configuration of assignments $\{\sigma_k^{(t)}\}$, which is precisely one iteration of the fixed-point algorithm. For different cost functions, similar optimal steps may be formulated, but the step may depend on i and $x_i^{(t)}$.

Choosing the best strategy between the fixed point approach and the gradient descent surely depends on the set of costs. When B is easily computable (more efficiently than by gradient descent), the fixed point algorithm moves the points faster than gradient descent. However, it is not obvious what should be the better option for complex costs c_k in practice. More generally, one could wonder if updating assignments more often (which is the case for the gradient descent on F) might not help avoiding local minima of the whole functional which is non convex in X. We did not observe this behaviour in practice in our experiments and therefore recommand the fixed point approach as the default choice.

C.II.4.3 Discrete Uniqueness Discussion

In this section, we investigate conditions to have uniqueness in the discrete Kantorovich problem between measures $\mu = \sum_{i=1}^{n_x} a_i \delta_{x_i} \in \mathcal{P}(\mathbb{R}^{d_x})$ and $\nu = \sum_{j=1}^{n_y} b_j \delta_{y_j} \in \mathcal{P}(\mathbb{R}^{d_y})$:

$$\min_{\pi \in \Pi(a,b)} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \pi_{i,j} c(x_i, y_j). \tag{C.II.32}$$

For convenience, we introduce $X := (x_1, \cdots, x_{n_x}) \in \mathbb{R}^{n_x \times d_x}$ and $Y := (y_1, \cdots, y_{n_y}) \in \mathbb{R}^{n_y \times d_y}$. The following result shows that if the cost matrix $M := (X, Y) \longmapsto (c(x_i, y_j))_{i,j} \in \mathbb{R}^{n_x \times n_y}$ is not orthogonal to a face of the transportation polytope, then the discrete Kantorovich problem has a unique solution. For convenience, we write $\pi \cdot M := \sum_{i,j} \pi_{i,j} M_{i,j}$.

Proposition C.II.5. Let $a \in \triangle_{n_x}$ and $b \in \triangle_{n_y}$ be fixed weights and $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \longrightarrow \mathbb{R}_+$ a cost function. Consider the cost matrix function

$$M := \left\{ \begin{array}{ccc} \mathbb{R}^{n_x \times d_x} \times \mathbb{R}^{n_y \times d_y} & \longrightarrow & \mathbb{R}^{n_x \times n_y} \\ (X, Y) & \longmapsto & (c(x_i, y_j))_{i,j} \end{array} \right. ,$$

and let $(X, Y) \in \mathbb{R}^{n_x \times d_x} \times \mathbb{R}^{n_y \times d_y}$. Denote by $\operatorname{Extr} \Pi(a, b)$ the (finite) set of extremal points of the transportation polytope $\Pi(a, b)$.

$$\min_{\pi \in \Pi(a,b)} \pi \cdot M(X, Y) \text{ has a unique solution } \iff M(X, Y) \notin \bigcup_{\pi_1 \neq \pi_2 \in \operatorname{Extr} \Pi(a,b)} (\pi_1 - \pi_2)^\perp. \tag{C.II.33}$$

Proof. Since $\Pi(a, b)$ is convex and compact in $\mathbb{R}^{n_x \times n_y}$, by the Krein-Milman theorem, it is the convex hull of the set of its extreme points, denoted $\operatorname{Extr} \Pi(a, b)$. With the definition

$$\Pi(a, b) = \left\{ \pi \in \mathbb{R}^{n_x \times n_y} : \pi \geq 0,\ \pi \mathbf{1} = a,\ \pi^\top \mathbf{1} = b \right\},$$

we see that $\Pi(a,b)$ is a polytope, and thus $\operatorname{Extr}\Pi(a,b)$ is finite. Since the Kantorovich problem is a linear problem, the set of optimal solutions is exactly the set of convex combinations of optimal extremal points. As a result, we have non-uniqueness in Eq. (C.II.32) if and only if there exists $\pi_1 \neq \pi_2 \in \operatorname{Extr}\Pi(a,b) : \pi_1 \cdot M(X,Y) = \pi_2 \cdot M(X,Y)$. We conclude that uniqueness holds if and only if $\forall \pi_1 \neq \pi_2 \in \operatorname{Extr}\Pi(a,b) : M(X,Y) \notin (\pi_1 - \pi_2)^{\perp}$. □

A consequence of Proposition C.II.5 is that if $M\#\mathscr{L}^{n_x \times d_x + n_y \times d_y}$ does not give mass to hyperplanes of $\mathbb{R}^{n_x \times n_y}$, then the Kantorovich problem has a unique solution for $\mathscr{L}^{n_x \times d_x + n_y \times d_y}$-almost-every (X,Y). Furthermore, if the measures have the same amount of points ($n_x = n_y$) and the weights are uniform, then the extreme points of $\Pi(a,b)$ are permutations, which provides a theoretical justification for the convenient expression in Eq. (C.II.28).

C.II.4.4 Application to Gaussian Mixture Model Barycentres

In this section, we explain how our fixed-point algorithm can be applied to compute barycentres between Gaussian Mixture Models (GMMs), providing a new numerical method for the GMM barycentre notion introduced in [DD20, Section 5]. The notation $S_d^{++}(\mathbb{R})$ will refer to the cone of positive definite symmetric $d \times d$ matrices.

We consider the case where the measures are Gaussian Mixture Models, seen as discrete measures over the space of Gaussian measures on $\mathbb{R}^d$: $\mathcal{X} := \mathcal{N} := \left\{\mathcal{N}(m,S) : m \in \mathbb{R}^d,\ S \in S_d^{++}(\mathbb{R})\right\}$, equipped with the 2-Wasserstein distance, which has a specific expression called the *Bures-Wasserstein distance*:

$$\mathrm{W}_2^2(\mathcal{N}(m_1,S_1),\mathcal{N}(m_2,S_2)) = \|m_1 - m_2\|_2^2 + \underbrace{\operatorname{Tr}\left(S_1 + S_2 - 2(S_1^{1/2} S_2 S_1^{1/2})^{1/2}\right)}_{d_{\mathrm{BW}}^2(S_1,S_2) :=}. \tag{C.II.34}$$

Alternatively, one could see the same problem differently, setting $\mathcal{X} := \mathbb{R}^d \times S_d^{++}(\mathbb{R})$ equipped with the distance defined in Eq. (C.II.34). To remind the definition of barycentres between Gaussian mixture models from [DD20], we will consider measures that lie on the same space of Gaussian measures: $\mathcal{X} = \mathcal{Y}_1 = \cdots = \mathcal{Y}_K = \mathcal{N}$. Next, we choose cost functions c_k on $\mathcal{N}$ as the squared Bures-Wasserstein distance W_2^2 scaled by λ_k. Given mixture models $\mu, \nu \in \mathcal{P}(\mathcal{N})$ of the form

$$\mu = \sum_{i=1}^{n} a_i \delta_{\mathcal{N}(m_i,S_i)},\ \nu = \sum_{j=1}^{m} b_j \delta_{\mathcal{N}(m'_j,S'_j)},$$

the Optimal Transport cost $\mathcal{T}_{\mathrm{W}_2^2}(\mu,\nu)$ is the value of a discrete problem, which is precisely the Mixed Wasserstein Distance introduced in [DD20, Proposition 4]:

$$\mathcal{T}_{\mathrm{W}_2^2}(\mu,\nu) = \min_{\pi \in \Pi(a,b)} \sum_{i,j} \pi_{i,j} \mathrm{W}_2^2(\mathcal{N}(m_i,S_i),\mathcal{N}(m'_j,S'_j)). \tag{C.II.35}$$

Consider K GMM measures ν_k written as:

$$\nu_k = \sum_{j=1}^{n_k} b_{k,j} \delta_{\mathcal{N}(m_{k,j},S_{k,j})} \in \mathcal{P}(\mathcal{N}),$$

their GMM barycentre cost with weights (λ_k) for $\mu = \sum_{i=1}^{n} a_i \delta_{\mathcal{N}(m_i,S_i)} \in \mathcal{P}(\mathcal{N})$ reads:

$$V(\mu) = \sum_{k=1}^{K} \lambda_k \min_{\pi_k \in \Pi(a,b_k)} \sum_{i,j} \pi_{i,j} \left(\|m_i - m_{k,j}\|_2^2 + d_{\mathrm{BW}}^2(S_i,S_{k,j})\right). \tag{C.II.36}$$

We now turn to the expression of the ground barycentre function $B : \mathcal{N}^K \longrightarrow \mathcal{N}$. This corresponds to a 2-Wasserstein barycentre problem in the Gaussian case, which was first studied by [AC11, Theorem 6.1] (showing existence and uniqueness):

$$B(\mathcal{N}(m_1,S_1),\cdots,\mathcal{N}(m_K,S_K)) = \mathcal{N}(\overline{m},\overline{S}),\ \overline{m} := \sum_{k=1}^{K} \lambda_k m_k,\ \overline{S} := \operatorname*{argmin}_{S \in S_d^{++}(\mathbb{R})} \sum_{k=1}^{K} \lambda_k d_{\mathrm{BW}}^2(S,S_k).$$

A fixed-point formulation of this problem is presented in [Álv+16] as a particular case of their study of the fixed-point algorithm for the ground cost $\|\cdot - \cdot\|_2^2$ and absolutely continuous measures. This problem is presented again in [BJL19], were they prove additional convergence guarantees. We recall from [Álv+16; BJL19] the fixed-point algorithm to compute the barycentre of K Gaussians $(\mathcal{N}(m_k, S_k))$ and weights $(\lambda_1, \cdots, \lambda_K)$, which consists in iterating the function $G_{\mathcal{N}} : S_d^{++}(\mathbb{R}) \longrightarrow S_d^{++}(\mathbb{R})$:

$$G_{\mathcal{N}}(S) = S^{-1/2}\left(\sum_{k=1}^{K} \lambda_k (S^{1/2} S_k S^{1/2})^{1/2}\right)^2 S^{-1/2}. \tag{C.II.37}$$

Now that we have defined the ground barycentre map B, we can apply our fixed-point algorithm to compute a barycentre. Given a reference GMM with n components $\mu = \sum_{i=1}^n a_i \delta_{\mathcal{N}(m_i,S_i)}$, for $k \in [\![1, K]\!]$, solve the discrete Kantorovich problem between μ and ν_k (Eq. (C.II.35)) and choose $\pi_k \in \Pi^*_{\mathrm{W}_2^2}(\mu, \nu_k)$. The GMM of $G(\mu)$ associated to the choice of plans $\pi_k \in \Pi(a, b_k)$ in the iteration scheme is the GMM $\overline{\mu}$ defined by:

$$\overline{\mu} = \sum_{j_1,\cdots,j_K} \sum_{i=1}^{n} \frac{1}{a_i^{K-1}} \pi^{(1)}_{i,j_1} \times \cdots \times \pi^{(K)}_{i,j_K} \delta[B(\mathcal{N}(m_{1,j_1}, S_{1,j_1}), \cdots, \mathcal{N}(m_{K,j_K}, S_{K,j_K}))].$$

As we argued in Section C.II.4.1, it is computationally wise to consider a variant of the fixed-point iterations which use the barycentric projections of the couplings π_k (see Eq. (C.II.21)). To use this in the case of the space $\mathcal{N}$, we need to choose a notion of convex combination in $\mathcal{N}$ to be able to compute the images of the barycentric projections. The most meaningful choice is a Wasserstein Gaussian barycentre, which corresponds to using the ground barycentre map B (this time with weights given by the disintegration of the coupling in question).

Remark C.II.4. The metric space $(\mathcal{N}, \mathrm{W}_2)$ is not compact, however we consider discrete measures (GMMs). We will show how one can restrict $\mathcal{N}$ to a compact subset containing all barycentres. Combining [DD20, Corollary 3] and [Álv+16, Theorem 4.2, Equations 20 and 21], shows that the barycentre is within a certain compact subset of $\mathcal{P}(\mathcal{N})$ of measures supported on Gaussians with covariances whose eigenvalues are in a segment $[r, R]$, where $0 < r < R$ are explicit constants depending on the covariances of the components of $\nu_1, \cdots, \nu_K$. As for the means, they can be constrained to the convex hull of the means of the components of the mixtures ν_k.

C.II.5 Numerical Illustrations

In this section, we provide numerical experiments to illustrate the fixed-point method (specifically its barycentric variant presented in Algorithm C.II.3) on various toy datasets. All code from this section is available in our companion Python toolkit. A numerical implementation of Algorithm C.II.2, which allows flexible support sizes, is also possible, but computationally much less appealing than Algorithm C.II.3.

C.II.5.1 Toy Example for Barycentre Computation

We begin with a simple example of barycentre computation in $\mathbb{R}^2$ of two discrete uniform measures with different support sizes and for the square-Euclidean cost $c_k(x, y) = \|x - y\|_2^2$. We observe convergence to the true barycentre in two iterations in Fig. C.II.9. The support size increases from 10 to 19 at the first iteration and remains at 19 at the final iteration.

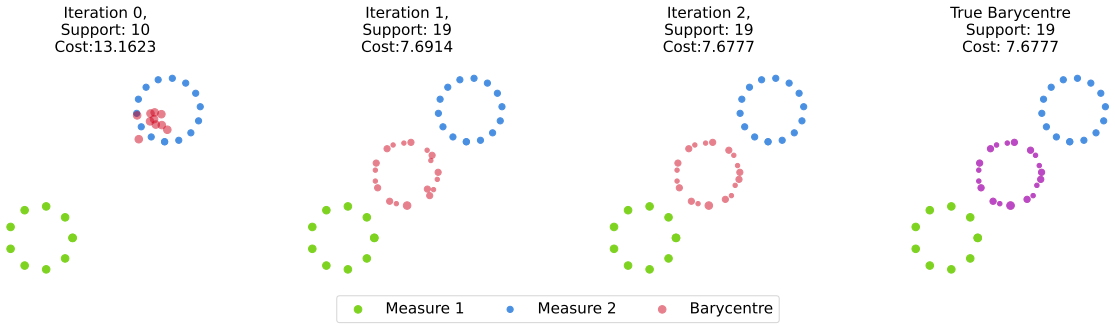


Figure C.II.9: Iterations of Algorithm C.II.2 for the square-Euclidean cost, and comparison with the true W_2^2 barycentre.

C.II.5.2 Illustration with Norm Powers

We consider discrete measures in $\mathbb{R}^2$ for costs $c_k(x, y) = \|x - y\|_p^q$, as illustrated in Fig. C.II.10.

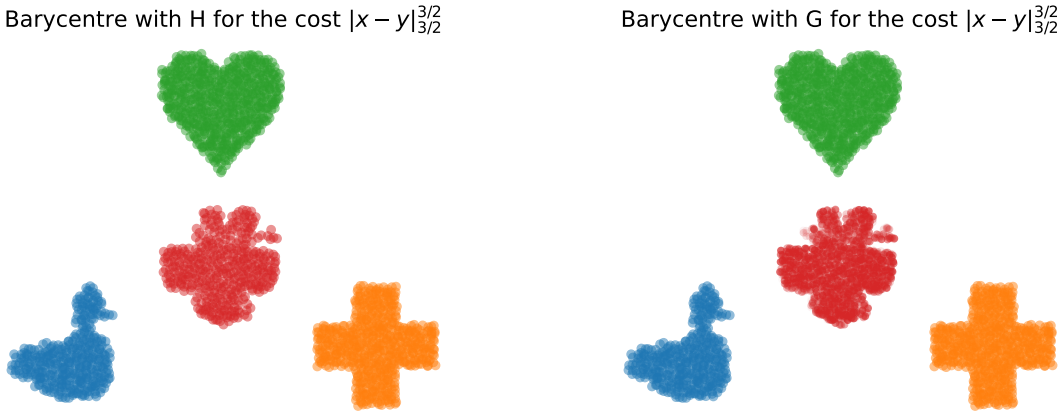
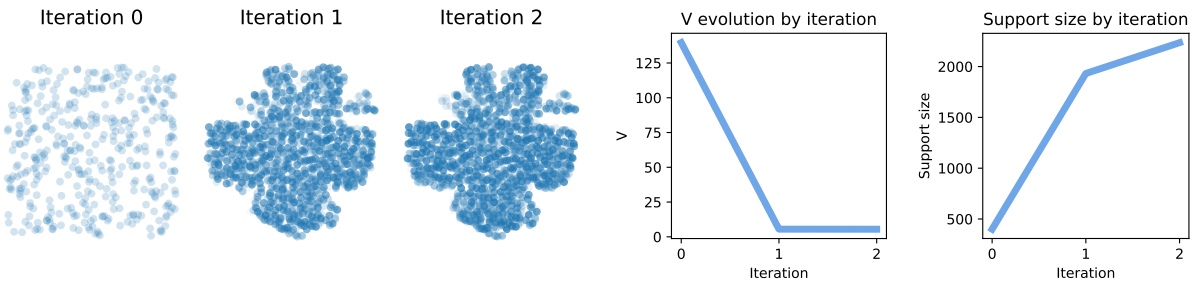


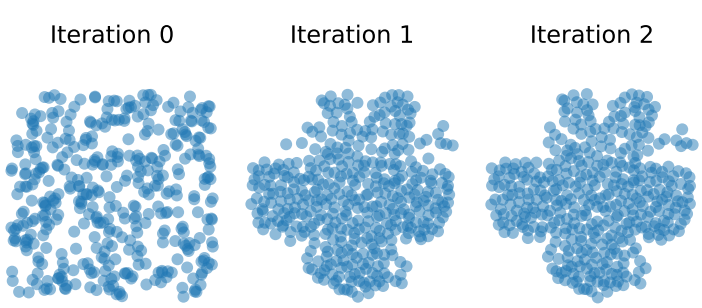
Figure C.II.10: Barycentres with initial support size $n = 400$ for $(p, q) = (\frac{3}{2}, \frac{3}{2})$ of three measures with sizes 561, 382, 629.

In Fig. C.II.11, we observe that for $(p, q) = (\frac{3}{2}, \frac{3}{2})$, the iterates of G (Algorithm C.II.2) have an energy that converges in one iteration, but the support size continues to grow at iteration 2. As for H (Algorithm C.II.3), we observe in Fig. C.II.12 convergence in one iteration. In Fig. C.II.13, we present barycentres for various pairs (p, q) using iterates of H.

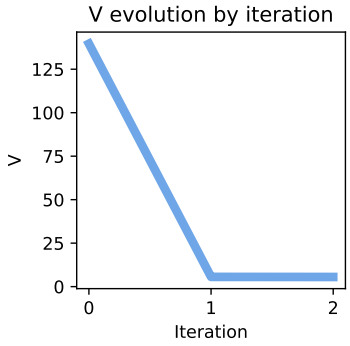


(a) Fixed-point iterations for $(p, q) = (\frac{3}{2}, \frac{3}{2})$. (b) Energy V. (c) Support size.

Figure C.II.11: Convergence of the iterations of G (Algorithm C.II.2).



(a) Fixed-point iterations for $(p, q) = (\frac{3}{2}, \frac{3}{2})$.



(b) Barycentre energy V of the iterations.

Figure C.II.12: Convergence of the iterations of H (Algorithm C.II.3).

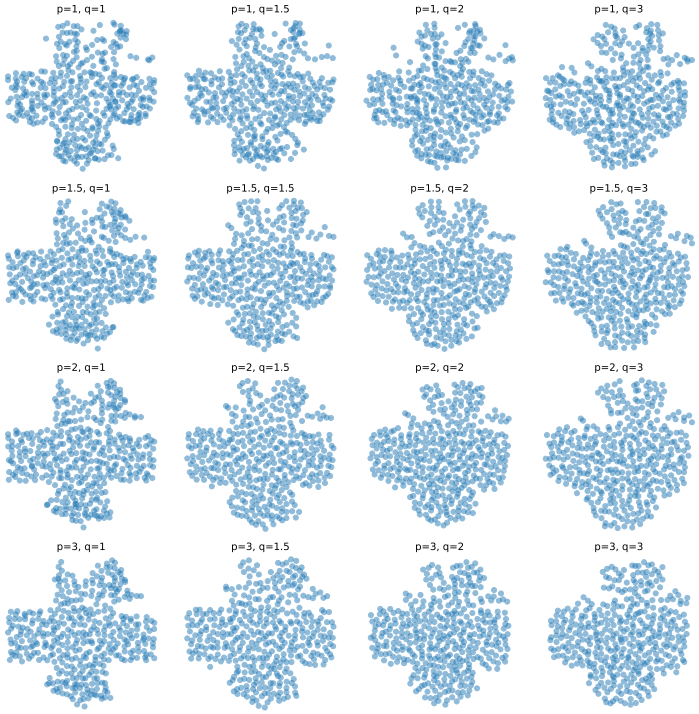


Figure C.II.13: Barycentres for the cost $\|x - y\|_p^q$ for different values of (p, q).

In the following, we consider a different setting where two of the three target measures are identical, and with a different third target. This will allow us to study the robustness properties of the associated barycentre, seeing the third different measure as an outlier. We represent the target measures and a barycentre in Fig. C.II.14, and compare different barycentres varying the parameters (p, q) of the cost $\| \cdot - \cdot \|_p^q$ in Fig. C.II.15. We observe that the barycentre obtained for $q = 1$ always takes the shape of the duck, as this power allows for greater robustness to outliers (here the heart-shaped cloud), regardless of the norm. The influence of the third point cloud becomes increasingly evident as p and q grow.

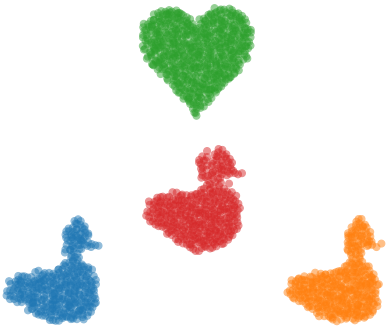


Figure C.II.14: Barycentre of three point clouds for the cost $\|\cdot - \cdot\|_{3/2}^{3/2}$.

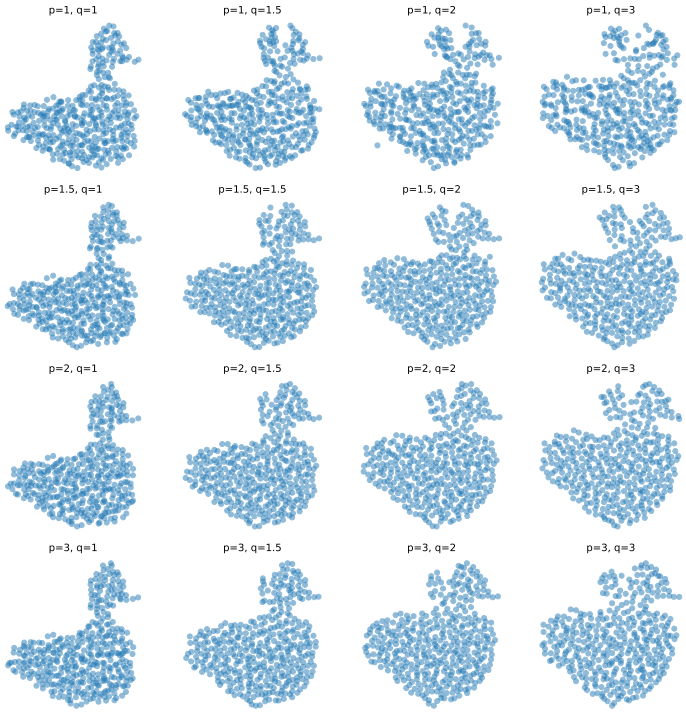
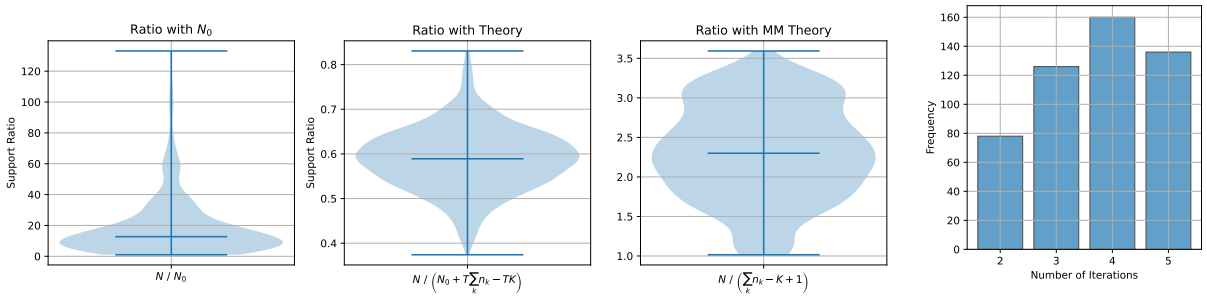


Figure C.II.15: Barycentres for the measures of figure C.II.14, for norms p raised to the power q.

C.II.5.3 Study of the Support Size of Iterates of G

In this section, we study the support size N of the final iteration of G (Algorithm C.II.2). As discussed in Section C.II.4.1, we expect (without formal proof) that the support size after T iterations is upper-bounded by $N_0 + T\sum_k n_k - TK$, where N_0 is the initial support size and n_k is the size of the k-th marginal. We verify this hypothesis on numerical experiments on numerous configurations varying $N_0, (n_k), d, (d_k)$ with measure points and weights generated randomly and for the square-Euclidean cost in Fig. C.II.16. We observe that the upper-bound is indeed respected, and that the algorithm attains convergence in a small number of iterations.



(a) Ratios of the final support size N with various support sizes.

(b) Iterations until convergence.

Figure C.II.16: Numerical study of the support size N of iterates of G. We ran 500 samples, each drawing a random number of measures $K \in [\![2, 10]\!]$, random dimensions $d, d_1, \cdots, d_K \in [\![1, 20]\!]$, random initial support sizes $N_0 \in [\![10, 100]\!]$ and random measure sizes $n_k \in [\![10, 100]\!]$.

Running the same experiment as in Fig. C.II.16 with $N_0 = n_1 = \cdots = n_K$ and uniform measure weights, we obtain, as expected in Eq. (C.II.28) that the support size N_t remains constant.

C.II.5.4 Comparison with the Multi-Marginal Formulation

Following Eq. (C.II.7), the discrete OT barycentre problem has a multi-marginal formulation, which can be written as follows, given measures $\nu_k = \sum_{j=1}^{m_k} b_{k,j}\delta_{y_{k,j}}$:

$$\underset{\pi \in \Pi(b_1, \cdots, b_K)}{\operatorname{argmin}} \sum_{j_1, \cdots, j_K} \pi_{j_1, \cdots, j_K} \sum_{k=1}^{K} c_k(B(y_{1,j_1}, \cdots, y_{K,j_K}), y_{k,j_k}). \tag{C.II.38}$$

Numerical solvers for Eq. (C.II.38), while slow, allow the computation of the exact solution of the barycentre problem. Comparing this solution to the output of our algorithm is technical, since the barycentric version of our algorithm imposes the size of the support of the barycentre in addition to imposing the weights, which introduces bias. We aim to illustrate that the speed of the barycentric algorithm, with a quantitative study of the error with respect to the multi-marginal "ground truth". Note that even in this square-euclidean experiment, there is no widespread multi-marginal solver, which is why we also contribute an implementation.

The experimental setup is the following: the K measures ν_k are all uniform measures with n points in $\mathbb{R}^d$ drawn independently from $\mathcal{N}(0, 1)$. For the fixed-point algorithm, the initial measure is also taken as a uniform measure over n points with $\mathcal{N}(0, 1)$ samples. We compare different numbers of iterations of the fixed-point algorithm and different choices of n, d, K. The plots show the ratios of the energy V and computation times for our algorithm divided by a Linear Programming multi-marginal solver, plotting 30% and 70% quantiles across 10 samples for each configuration. As expected in Eq. (C.II.28), since in this case the measures are uniform with a common support size, the iterates of H and G are identical in this setting.

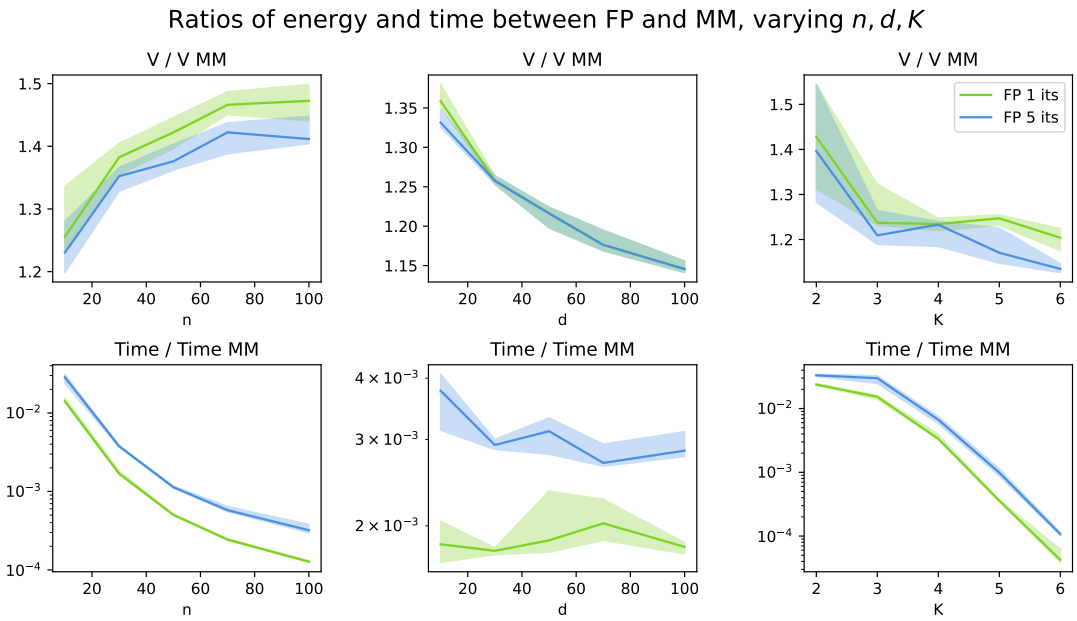


Figure C.II.17: Comparing the fixed-point solver with a linear programming multi-marginal solver. From left to right columns: varying n with $d = 10$ and $K = 3$; varying d with $n = 30$ and $K = 3$; varying K with $n = 10$ and $d = 10$. The comparison is made by dividing the energy value V (resp. computation time) of the fixed-point solution by the multi-marginal solution. The different curves correspond to $T = 1, 5, 10$ iterations (legend in the top-right).

From the results presented in Fig. C.II.17, it appears that the fixed-point algorithm converges in very few iterations, has an energy at most 50% worse than the exact multi-marginal solution, and is orders of magnitude faster, especially for larger measure sizes n and for greater numbers of marginals K. Note that for $n \geq 10$ and $K \geq 10$ for example, the multi-marginal problem is computationally intractable.

To compare with similar barycentre support sizes, in Fig. C.II.18 we experiment with fixed-point barycentres using H (Algorithm C.II.3) with $N_{\mathrm{FP}} = (n-1)K + 1$ points. The rationale behind this choice stems from the fact that discrete measures with $n_1, \cdots, n_K$ points have a barycentre with at most $\sum_k n_k - K + 1$ points ([ABM16, Theorem 2][3]).

[3] whose techniques are in fact not specific to the cost $\|\cdot - \cdot\|_2^2$

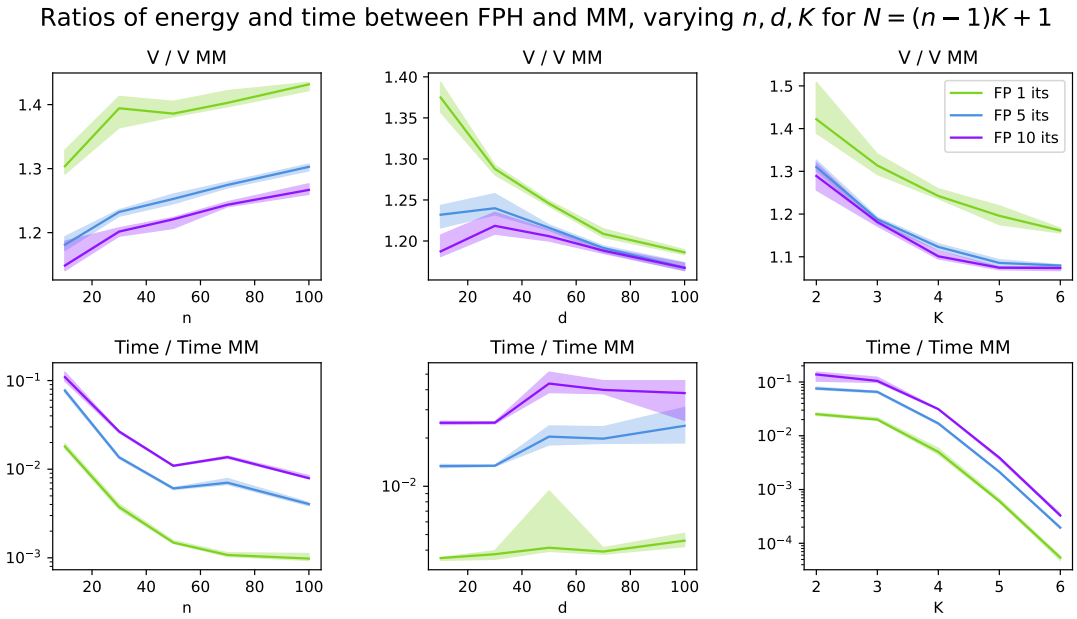


Figure C.II.18: Comparing the fixed-point solver from Algorithm C.II.3 for $N_{\text{FP}} = (n-1)K+1$ and the same setup as in Fig. C.II.17.

We now focus on the iterations of G (Algorithm C.II.2) in the case of uniform measures where the initialisation is taken with n points and the target measures have even spaced sizes $n_1 = \frac{n}{2} \cdots n_K = 2n$. This ensures that iterates of G differ from iterates of H, and we present the results in Fig. C.II.19.

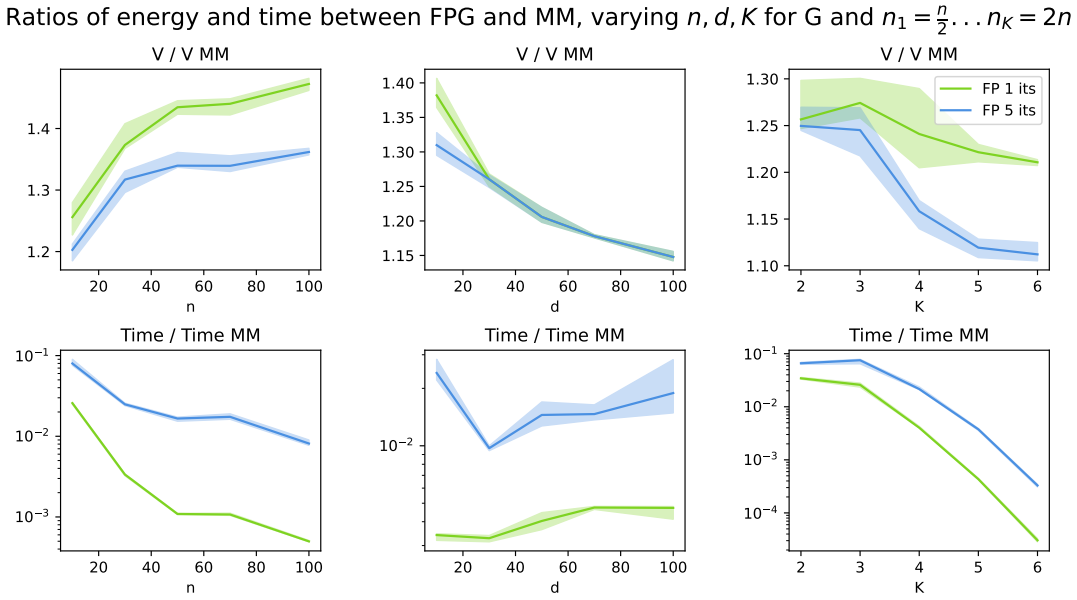
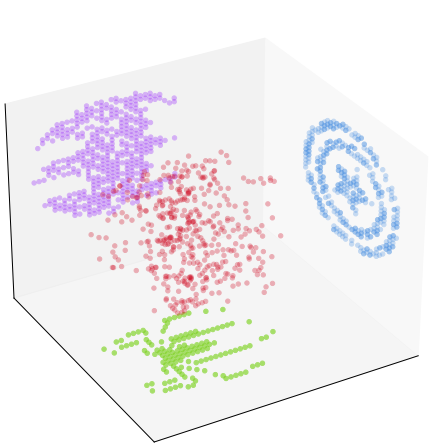


Figure C.II.19: Comparing the fixed-point solver from Algorithm C.II.2 with the MM solver, for an initialisation with n points and target measures with different sizes $n_1 = \frac{n}{2} \cdots n_K = 2n$.

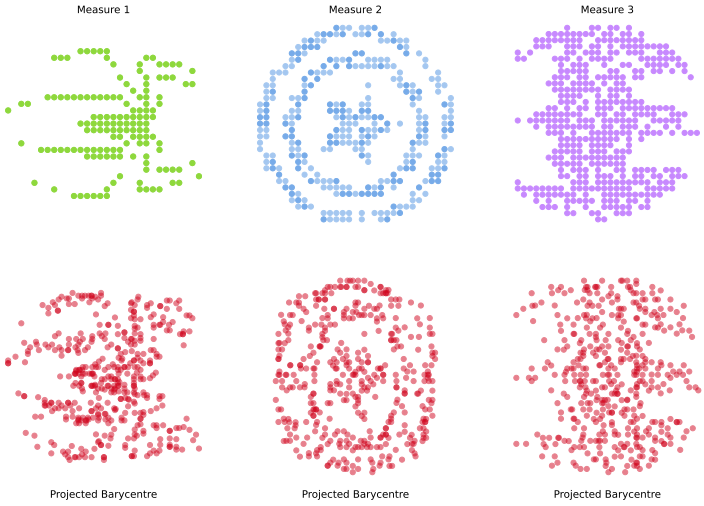
Figs. C.II.17 to C.II.19 suggests that the fixed-point methods proposed in Algorithms C.II.2 and C.II.3a re useful as a fast approximate solvers for the barycentre problem, and that settings with larger barycentre supports may require more iterations to converge. The main takeaway is that our methods remain competitive for large supports and number of target measures, yet its convergence speed and overall advantages are more pronounced for smaller supports.

C.II.5.5 Generalised Wasserstein Barycentre Computation

In Fig. C.II.20a, we illustrate the case where $c_k(x, y) = \|P_k x - y\|_2$, where $P_k : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$ is an orthogonal projection. The problem finds a 3D measure whose projections attempt to match the reference 2D measures, which we compare in Fig. C.II.20b. This is a modification of the exponent 2 from Generalised Wasserstein Barycentres [DGS21].



(a) Barycentre in $\mathbb{R}^3$ with immersed measures $\nu_k \in \mathcal{P}(\mathbb{R}^2)$.



(b) Projections of the barycentre into $\mathbb{R}^2$.

Figure C.II.20: Barycenter (using Algorithm C.II.3) with costs $c_k(x, y) = \|P_k x - y\|_2$, where P_k are orthogonal projections from $\mathbb{R}^3$ to the three axes-aligned planes of the orthonormal basis. We provide an animation in the companion code.

C.II.5.6 Non-linear Generalised Wasserstein Barycentre Computation

In this illustration, we look for a barycentre in $\mathbb{R}^2$ whose projections onto different circles match measures on these circles. We choose the costs $c_k(x, y) = \|P_k(x) - y\|_2^2$, where P_k is the projection onto the circle k. Since P_k is not linear, this is a direct generalisation of [DGS21].

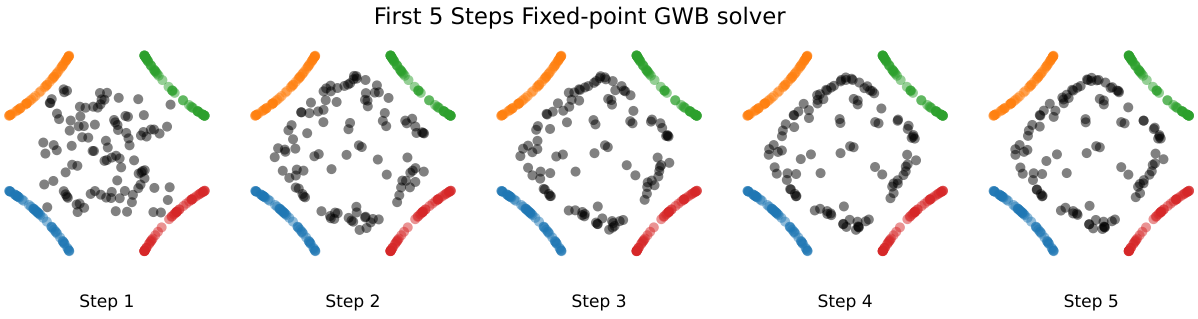


Figure C.II.21: First 5 iterations of the fixed-point algorithm (Algorithm C.II.3) for costs $c_k(x, y) = \|P_k(x) - y\|_2^2$, where P_k are projections onto four different circles on which the ν_k are supported (plotted in colour).

In this instance, convergence happens quickly, but a stationary point is only reached after about 5 iterations, as observed on the steps in Fig. C.II.21 and on the energy curve in Fig. C.II.22.

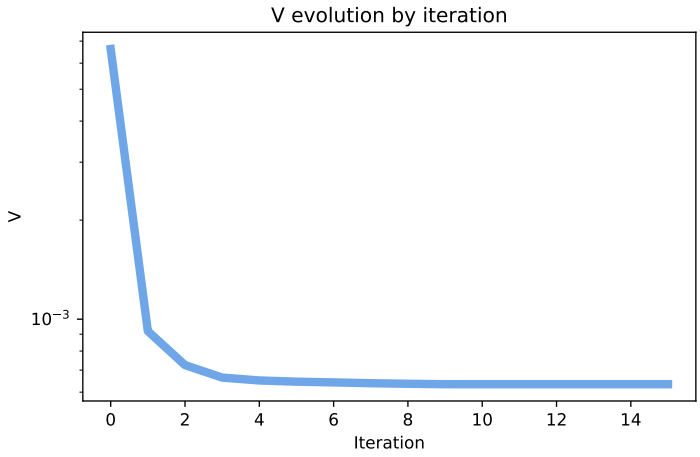


Figure C.II.22: Barycentre energy V of the fixed-point algorithm for H across iterations.

C.II.5.7 Gaussian Mixture Model Barycentres

We illustrate numerical solutions of the GMM Barycentre method introduced in Section C.II.4.4. In Fig. C.II.23, we compare the multi-marginal solution with the output of our algorithm (we use Algorithm C.II.3).

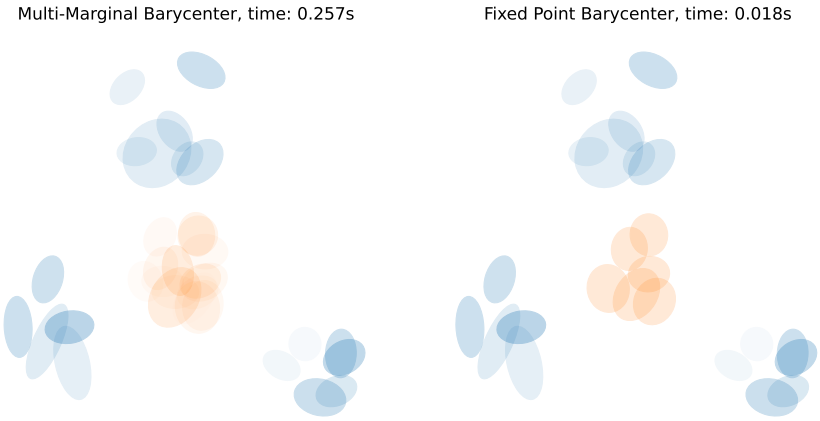


Figure C.II.23: Left: multi-marginal solution for the GMM barycentre problem. Right: fixed-point solution for $n = 6$ components.

Finally, in Fig. C.II.25 we illustrate barycentres between 4 GMMs shown in Fig. C.II.24 with different weights.

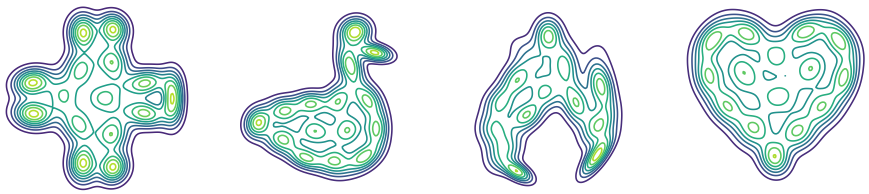


Figure C.II.24: Four GMMs of which we will compute barycentres in Fig. C.II.25.

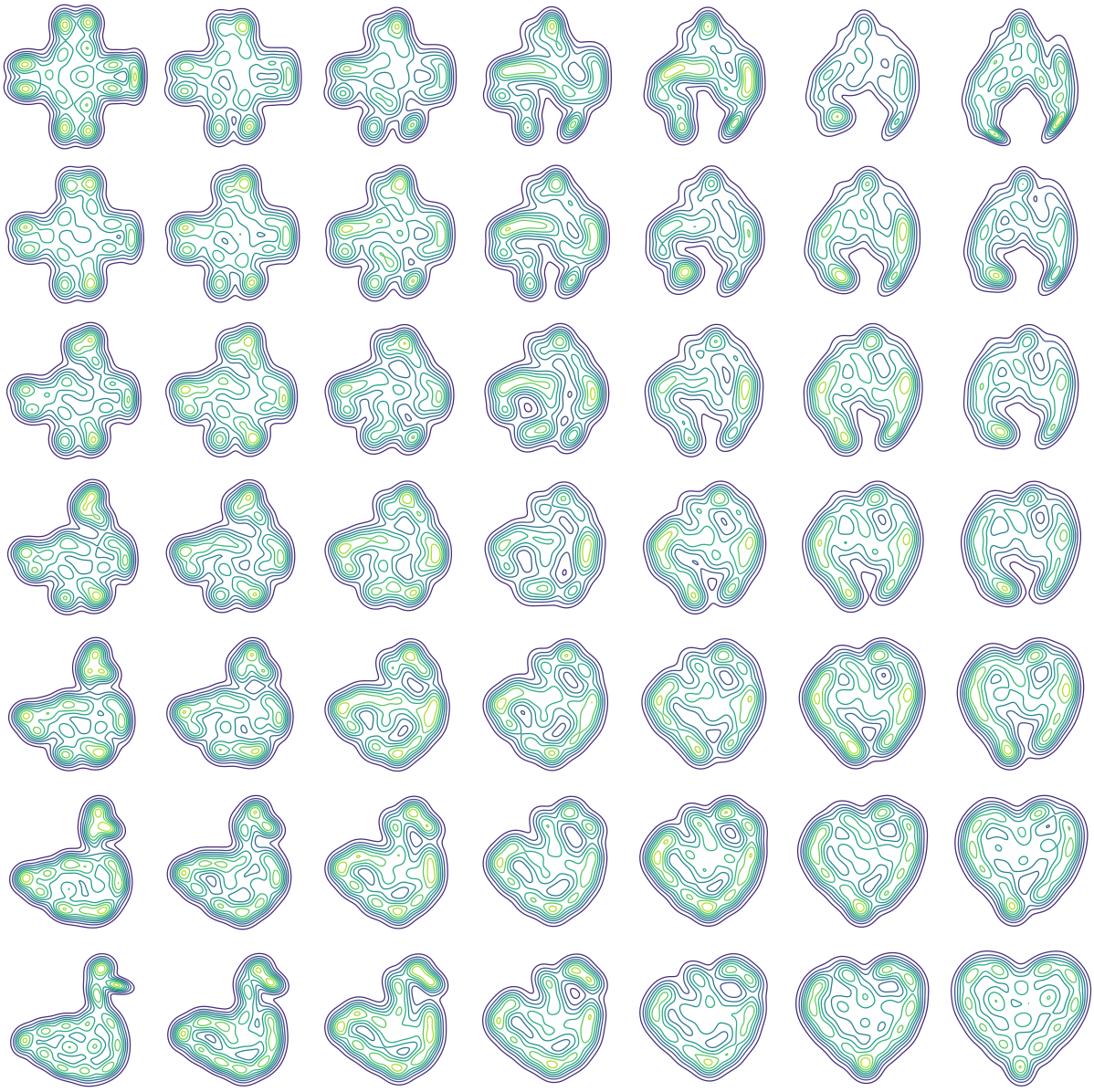


Figure C.II.25: GMM barycentres between the four corner GMMs computed with the fixed-point solver with $n = 15$ components. The GMMs are represented by the contours of their densities on $\mathbb{R}^2$.

C.II.5.8 Colour Transfer on a Barycentre of Colour Distributions

In this final experiment, we consider a colour transfer problem. The goal is to compute the barycentre of the colour distributions of several (here three) source images, some of which contain outlier colours, and then use this barycentre as a target measure to modify the colours of a new image (referred to as the *input* here). Figure C.II.26 shows the source images, the *input* image, and the same *input* image after transferring its colour distribution to that of the colour barycentre of the source images. The barycentre is computed either for a W_1 cost or for a W_2^2 cost. This transfer is evaluated on downsampled images, with the RGB matching of a colour c in the high-resolution image subsequently chosen as $c+\tau$, where τ is the colour translation obtained for the closest colour to c in the downsampled image (this amounts to viewing the matching as a piecewise constant translation field). Figure C.II.27 shows the colour distributions of the images in the RGB space. We observe that the W_1 cost enjoys greater robustness to the colour outliers compared to the usual W_2^2 cost.

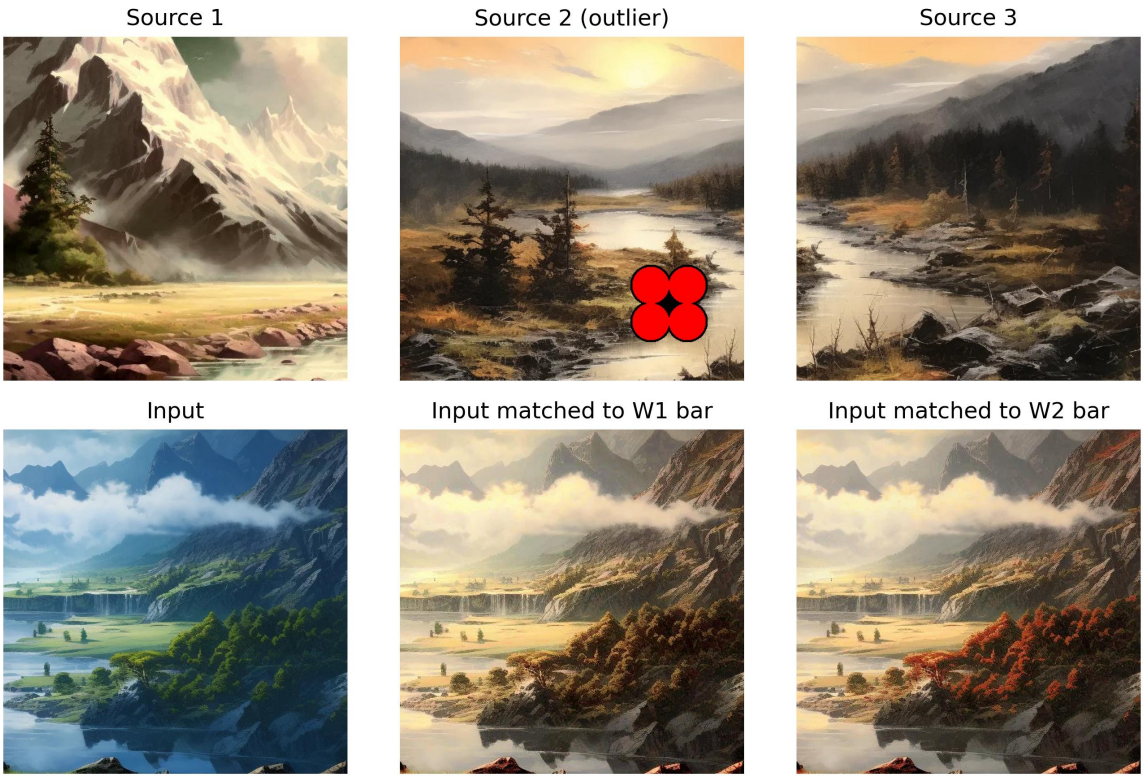


Figure C.II.26: Colour transfer applied to the input image towards the colour barycentre of the source images, for the costs W_1 and W_2^2. One of the source images contains unwanted colour artifacts, which we see as outliers.

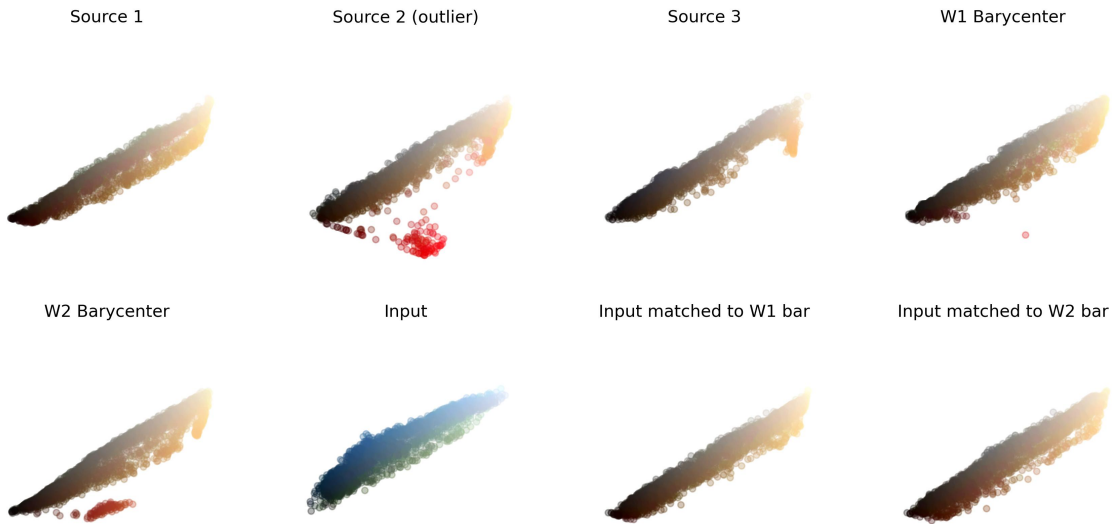


Figure C.II.27: Colour distributions of the different images from figure C.II.26 as well as the W_2^2 and W_1 barycentres of the source images.

Future Directions

There are numerous directions for future research. To begin with, in Theorems C.II.1 and C.II.2, we show subsequential convergence to fixed-points of G (resp. G_ε), which may not be barycentres. In cases where barycentres and fixed points may not be unique such as the discrete setting, it remains unclear if there exists fixed points that are not barycentres.

The barycentric fixed-point algorithm (iterating Eq. (C.II.21)) has no theoretical guarantees of convergence. Given its computational advantages and its current use in practice for the

squared Euclidean cost ([CD14], [Fla+21]), this is a timely question.

In Section C.II.3.3, we required a notion of barycentric projection for couplings $\pi \in \Pi^*_{c_k}(\mu, \nu_k)$. In $\mathbb{R}^d$, the underlying convex combinations are performed using the usual linear structure, however this does not generalise to arbitrary metric spaces. To consider these object more formally on generic (compact) metric spaces, it would be necessary to discuss in more detail the meaning of expectation in a space without a linear structure.

Throughout this work, we relied heavily on Assumption C.II.2, but in practice this can be difficult to verify for costs c_k: beyond the case $c_k = h(x - y)$ with h strictly convex, it is difficult to provide large classes of costs that yield this property on B (other examples include $c_k(x, y) = \|P_k x - y\|_2^2$ as in [DGS21] or W_2^2 for absolutely continuous measures). One could alternatively investigate a theoretical framework where B is a multi-function.

In the absolutely continuous case, the Twist condition can ensure uniqueness of the barycentre, as explained in Remark C.II.1. A natural question concerns almost-sure uniqueness in the discrete case, as was partially explored in Section C.II.4.3.

From a numerical standpoint, it has been observed that the fixed-point algorithm converges in very few iterations. A theoretical work extending the discrete Wasserstein case from [Lin23] would bridge a significant gap between theory and practical observation.

Acknowledgements

We would like to thank Christophe Gaillac for the initial discussions that motivated the introduction of barycentres with generic costs. This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

Part D

Some Contributions to Kernel Methods

Chapter D.I opens this part with a short introduction to Reproducing Kernel Hilbert Spaces, based on blackboard introductory presentations given by the author at the MAP5 laboratory.

Chapter D.II explores the idea of using cones in Reproducing Kernel Hilbert Spaces to represent gradients of convex functions, providing negative existence results for such methods. This chapter is based on joint work with Joan Glaunès.

Chapter D.III introduces an explicit construction of universal kernels on compact metric spaces, and a notion of approximate universality, with a focus on certain tractable kernels that are shown to be approximately universal. This chapter is based on the paper:

[Tan25] Eloi Tanguy.
"Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces".
arxiv preprint 2506.03661 (Jun. 2025).

D.I

A Gentle Introduction to RKHS

In this part, we will delve into theoretical considerations around Reproducing Kernel Hilbert Spaces (RKHS), which is a field of Mathematics that is relatively far removed from Optimal Transport on which the rest of this thesis is focused. To ease the reader into the topic, we will briefly introduce the field of RKHS theory. This introductory chapter is based on blackboard talks given by the author at the MAP5 laboratory in Paris, and is intended to be accessible to a wide audience.

D.I.1 Reproducing Kernel Hilbert Spaces

There are many different equivalent definitions of a Reproducing Kernel Hilbert Space (RKHS): in particular, it is possible to begin with a kernel and to construct the associated RKHS, or to begin with a Hilbert space with certain properties and to "discover" its kernel. We will focus on the latter viewpoint, and consider a Hilbert space $(H, \langle \cdot, \cdot \rangle_H)$ of functions $\mathcal{X} \longrightarrow \mathbb{R}$, where $\mathcal{X}$ is a set without a particular structure. First, we remind that in Definition D.I.1 the definition of a Hilbert space.

Definition D.I.1. A **Hilbert space** is a vector space H over $\mathbb{R}$ equipped with an inner product $\langle \cdot, \cdot \rangle_H$ that is **complete**, which is to say that every Cauchy sequence in H converges (for the topology induce by the norm $\|h\|_H := \sqrt{\langle h, h \rangle_H}$) to an element of H.
A **Cauchy sequence** is a sequence $(h_n)_{n \in \mathbb{N}}$ in H such that for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$, we have $\|h_n - h_m\|_H \leq \varepsilon$.

For the sake of simplicity, we consider spaces of real-valued functions and Hilbert spaces over $\mathbb{R}$, but the definitions can be extended to $\mathbb{C}^d$-valued functions and Hilbert spaces over $\mathbb{C}$. Given a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$, the **evaluation map** at a point $x \in \mathcal{X}$, defined by

$$\delta_x := \left\{ \begin{array}{ccc} H & \longrightarrow & \mathbb{R} \\ h & \longmapsto & h(x) \end{array} \right. \in \mathcal{L}(H, \mathbb{R}),$$

where $\mathcal{L}(H, \mathbb{R})$ is the space of linear maps from H to $\mathbb{R}$, is of particular interest in RKHS theory:

Definition D.I.2. A Hilbert space H of functions $\mathcal{X} \longrightarrow \mathbb{R}$ is said to be a **Reproducing Kernel Hilbert Space** (RKHS) if the evaluation map δ_x is continuous for every $x \in \mathcal{X}$. This means that for every $x \in \mathcal{X}$, there exists a constant $C_x \geq 0$ such that for all $h \in H$, we have:

$$|\delta_x(h)| = |h(x)| \leq C_x \|h\|_H.$$

We can write this condition $\delta_x \in H'$, where H' is the space of continuous linear maps from H to $\mathbb{R}$ (the topological dual of H).

In a RKHS, the norm is "strong" in the sense that convergence in H implies pointwise convergence of functions. In other words, if a sequence of functions $(h_n)_{n \in \mathbb{N}} \in H^{\mathbb{N}}$ is such that

$\|h_n - h\|_H \xrightarrow[n\longrightarrow+\infty]{} 0$ for some $h \in H$, then for every $x \in \mathcal{X}$, we have:

$$|h_n(x) - h(x)| = |\delta_x(h_n - h)| \leq C_x \|h_n - h\|_H \xrightarrow[n\longrightarrow+\infty]{} 0.$$

A simple example of a RKHS space is the space of "band-limited" functions, which we present in Example D.I.1.

Example D.I.1. We consider the space $H := \left\{ f \in \mathcal{C}^0(\mathbb{R}) \cap L^2(\mathbb{R}) : \operatorname{supp} \hat{f} \subset [-a, a] \right\}$ of continuous functions f on $\mathbb{R}$ verifying $\int_{\mathbb{R}} f^2 < +\infty$ and whose Fourier transform $\hat{f}$ is supported in the interval $[-a, a]$. We equip H with the L^2 inner product: $\langle f, g \rangle_H := \int_{\mathbb{R}} fg$. The space $(H, \langle \cdot, \cdot, \rangle)_H$ is an RKHS: we apply Definition D.I.2 and fix $x \in \mathbb{R}$ and $f \in H$, with the convention that

$$\hat{f} := \omega \longmapsto \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-i\omega x} \mathrm{d}x,$$

we compute using the Fourier inversion formula, the Cauchy-Schwarz inequality, and the Parseval identity:

$$\begin{aligned} |\delta_x(f)| &= \left| \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\omega) e^{i\omega x} \mathrm{d}\omega \right| \leq \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} |\hat{f}(\omega)| \mathrm{d}\omega \\ &\leq \frac{1}{\sqrt{2\pi}} \sqrt{\int_{-a}^{a} |\hat{f}(\omega)|^2 \mathrm{d}\omega} \sqrt{\int_{-a}^{a} 1^2 \mathrm{d}\omega} = \sqrt{\frac{a}{\pi}} \sqrt{\int_{\mathbb{R}} |\hat{f}(\omega)|^2 \mathrm{d}\omega} = \sqrt{\frac{a}{\pi}} \|f\|_H. \end{aligned}$$

To define the kernel of an RKHS, we will use a well-known result in Hilbert space theory:

Theorem D.I.1. Riesz Representation Theorem [RAG05, Theorem 13.31] Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space, for every continuous linear functional $\ell \in H'$, there exists a unique element $R\ell \in H$ such that for all $h \in H$, we have $\ell(h) = \langle h, R\ell \rangle_H$. The Riesz operator $R : H' \longrightarrow H$ is an isometric isomorphism[a].

[a] i.e. $\|R\ell\|_H = \|\ell\|_{H'} := \sup_{\|h\|_H \leq 1} |\ell(h)|$ and R is linear and bijective.

Using Theorem D.I.1, we can define the kernel of an RKHS using the evaluation maps, which by assumption are continuous:

Definition D.I.3. The **kernel** of an RKHS H is the map $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ defined by:

$$k := (x, y) \in \mathcal{X}^2 \longmapsto \langle R\delta_x, R\delta_y \rangle_H,$$

where $R\delta_x, R\delta_y \in H$ are the Riesz representations of δ_x, δ_y as in Theorem D.I.1.

For convenience, the element $R\delta_x \in H$ is often denoted by $K_x := R\delta_x$. The map $x \longmapsto K_x$ is called the *canonical feature map*, for reasons that we will not elaborate until Chapter D.III. Using the properties of the inner product $\langle \cdot, \cdot \rangle_H$ and of the Riesz representation, the following properties of the kernel k can be deduced:

Proposition D.I.1. The kernel $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ of an RKHS H satisfies the following properties for any $x, y \in \mathcal{X}$:

1) Symmetry: $k(x, y) = k(y, x)$.

2) Positivity:

$$\forall n \in \mathbb{N}^*,\ \forall x_1, \cdots, x_n \in \mathcal{X},\ \forall a \in \mathbb{R}^n,\ \sum_{i,j} a_i k(x_i, x_j) a_j \geq 0$$

3) Feature identity: $k(\cdot, x) = K_x$.

4) Reproducing property: $\forall h \in H,\ h(x) = \langle h, K_x \rangle_H$.

5) Self-reproducing property: $\langle k(\cdot, x), k(\cdot, y) \rangle_H$.

Proof. For 1), we have $k(x, y) = \langle K_x, K_y \rangle_H = \langle K_y, K_x \rangle_H = k(y, x)$. To show 2), we compute:

$$\sum_{i,j} a_i k(x_i, x_j) a_j = \sum_{i=1}^n a_i \left\langle K_{x_i}, \sum_{j=1}^n a_j K_{x_j} \right\rangle_H = \left\langle \sum_{i=1}^n a_i K_{x_i}, \sum_{j=1}^n a_j K_{x_j} \right\rangle_H = \left\| \sum_{i=1}^n a_i K_{x_i} \right\|_H^2 \geq 0.$$

For 3), since $K_y = R\delta_y$ it holds that $K_x(y) = \delta_y(K_x) = \langle K_x, K_y \rangle_H = k(y, x)$, concluding that $K_x = k(\cdot, x)$. Regarding 4), we use $K_x = R\delta_x$ and obtain $h(x) = \delta_x(h) = \langle h, K_x \rangle_H$. For 5), we apply 4) to $h := K_y$. □

A natural question is whether a suitable function $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ can be a reproducing kernel of an RKHS H. The answer is that the properties 1) and 2) of Proposition D.I.1 are sufficient. Before stating the result, we introduce a notation for such functions:

Definition D.I.4. A function $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ is said to be **symmetric-positive** if it verifies 1) and 2) of Proposition D.I.1. In that case, we write $k \in S^2(\mathcal{X})$.

Theorem D.I.2 (Moore-Aronszajn Theorem [Aro50])**.** Let $k \in S^2(\mathcal{X})$ be a symmetric-positive function. Then there exists a unique (up to isometry) RKHS $(H, \langle \cdot, \cdot \rangle_H)$ with kernel k.

The idea of the proof of Theorem D.I.2 is to begin with the space

$$H_0 := \left\{ \sum_{i=1}^n a_i k(\cdot, x_i),\ n \in \mathbb{N},\ a \in \mathbb{R}^n,\ (x_1, \cdots, x_n) \in \mathcal{X}^n \right\},$$

equipped with the inner product:

$$\left\langle \sum_{i=1}^n a_i k(\cdot, x_i), \sum_{j=1}^m b_j k(\cdot, y_j) \right\rangle_{H_0} = \sum_{i,j} a_i k(x_i, y_j) b_j.$$

The space $(H_0, \langle \cdot, \cdot \rangle_{H_0})$ is only a pre-Hilbert space, and the technicality of the proof resides in studying its Hilbertian completion (limits of Cauchy sequences in H_0). A self-sufficient proof of Theorem D.I.2 can be found in Jean-Philippe Vert's course notes. Some commonly used kernels are:

- The Gaussian (or RBF) kernel: $k(x, y) := \exp(-\|x - y\|_2^2 / s^2)$
- The polynomial kernel: $k(x, y) := (c + \langle x, y \rangle_2)^d$
- The Laplace kernel: $k(x, y) := \exp(-\|x - y\|_2 / s)$
- Radial kernels for some finite positive measure μ on $\mathbb{R}_+$: $k(x, y) := \int_0^{+\infty} e^{-t\|x-y\|_2^2} \mathrm{d}\mu(t)$
- Taylor kernels for coefficients $(a_n) \in \mathbb{R}_+^{\mathbb{N}}$ with sufficient decay: $k(x, y) := \sum_{n=0}^{+\infty} a_n \langle x, y \rangle^n$

D.I.2 Kernel Interpolation

In this section, we explain the concept of kernel interpolation, which allows for the approximation of a function $f : \mathcal{X} \longrightarrow \mathbb{R}$ whose values are known at a finite set of points $\{x_1, \dots, x_n\} \subset \mathcal{X}$. In other words, given values (y_i) and points (x_i), we want to find the simplest possible function of an RKHS H verifying $h(x_i) = y_i$ at each i. Again, we focus on the case of a real-valued functions for simplicity. Mathematically, the goal is to find a solution of the following exact interpolation problem:

$$\underset{\substack{h \in H \\ \forall i \in [\![1,n]\!],\, h(x_i)=y_i}}{\operatorname{argmin}} \quad \|h\|_H^2. \tag{D.I.1}$$

Minimising the norm $\|h\|_H$ can be understood as a way of finding the simplest possible function in H verifying the interpolation conditions. To motivate this intuition, we refer to [CS08, Theorem 4.48], which states that the norm associated to the RKHS induce by the Gaussian kernel on a bounded set of $\mathbb{R}^d$ dominates all Sobolev norms. In this setting, a small norm implies that the function is very regular, which corresponds to our intuitive term "simple function".

Before tackling the problem of Eq. (D.I.1), we first study the constraints in the case where a solution exists: we provide a characterisation of the condition that two functions $h_1, h_2 \in H$ be equal on a finite set of points $\{x_1, \dots, x_n\}$.

Proposition D.I.2. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$ For any $h_1, h_2 \in H$, we have:

$$\forall i \in [\![1, n]\!],\ h_1(x_i) = h_2(x_i) \iff h_1 - h_2 \in W^\perp,\ W := \operatorname{Span}\left(k(\cdot, x_i)\right)_{i=1}^n$$

Proof. By the reproducing property of k (Proposition D.I.1 item 4)), we have for each $i \in [\![1, n]\!]$:

$$(h_1 - h_2)(x_i) = 0 \iff \langle h_1 - h_2, k(\cdot, x_i)\rangle_H = 0 \iff h_1 - h_2 \in k(\cdot, x_i)^\perp,$$

concluding the proof by intersecting over $i \in [\![1, n]\!]$. □

Thanks to the characterisation of Proposition D.I.2, we can reformulate the interpolation problem of Eq. (D.I.1) as a problem over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$. This way, the infinite-dimensional problem of Eq. (D.I.1) can be reduced to a finite-dimensional problem over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$.

Theorem D.I.3. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$ and $(y_1, \dots, y_n) \in \mathbb{R}^n$. Consider the matrix $K \in \mathbb{R}^{n \times n}$ of entries $K_{i,j} := k(x_i, x_j)$, and assume that K is invertible. Then there is a unique solution to the exact kernel interpolation problem of Eq. (D.I.1), which is:

$$h^* = \sum_{i=1}^n a_i k(\cdot, x_i), \text{ where } a = K^{-1} y, \tag{D.I.2}$$

with $y := (y_1, \cdots, y_n) \in \mathbb{R}^n$.

Proof. For existence, take $a := K^{-1}y$ and let $h_0 := \sum_{i=1}^n a_i k(\cdot, x_i)$. Since $Ka = y$, the function h_0 satisfies the interpolation conditions $h_0(x_i) = y_i$ for all $i \in [\![1, n]\!]$. Now we see from Proposition D.I.2 that $h \in H$ verifies the constraints if and only if $h - h_0 \in W^\perp$, which can be re-written as $h = h_0 + h_\perp$ for some $h_\perp \in W^\perp$. We then have by orthogonality $\|h\|_H^2 = \|h_0\|_H^2 + \|h_\perp\|_H^2$, showing that for any $h \in H$ verifying the constraints, we have $\|h\|_H^2 \geq \|h_0\|_H^2$, which shows that h_0 is indeed a solution of the interpolation problem Eq. (D.I.1).

For uniqueness, if $h \in H$ verifies the constraints, we have seen that we can write $h = h_0 + h_\perp$ for some $h_\perp \in W^\perp$. If $h \notin W$ then $h_\perp \neq 0$ and thus $\|h\|_H^2 > \|h_0\|_H^2$, which shows that a solution h must be in W. Writing such a solution as $h = \sum_{i=1}^n b_i k(\cdot, x_i)$ for some $b \in \mathbb{R}^n$, using the constraints we observe that $Kb = y$, which implies that $a = b$ by invertibility, showing that the expression in Eq. (D.I.2) is the unique solution. □

The condition of invertibility of the matrix K in Theorem D.I.3 is crucial, as it ensures that the interpolation problem has a unique solution. For some specific kernels, this condition can be guaranteed for distinct points $(x_i)_{i=1}^n$ (see [Mic86, Theorem 2.3] or [Set22, Theorem 3.5] for example). In general, it is natural to resort to a regularisation technique, which consists in replacing the exact interpolation problem Eq. (D.I.1) by the following approximate interpolation problem:

$$\operatorname*{argmin}_{h\in H} \lambda\|h\|_H^2 + \sum_{i=1}^n (h(x_i) - y_i)^2, \tag{D.I.3}$$

which we interpret as a kernel regression problem. Using the same tools as in Theorem D.I.3, we will reduce this problem to optimisation over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$:

Theorem D.I.4. For $(x_1, \dots, x_n) \in \mathcal{X}^n$, $(y_1, \dots, y_n) \in \mathbb{R}^n$ and $\lambda > 0$. A function $h \in H$ is a solution of the kernel regression problem of Eq. (D.I.3) if and only if it can be written $h = \sum_{i=1}^n a_i^* k(\cdot, x_i)$ for some $a^* \in \mathbb{R}^n$ which is a solution to:

$$\operatorname*{argmin}_{a\in\mathbb{R}^n} \lambda a^\top K a + \|Ka - y\|_2^2, \tag{D.I.4}$$

where $K \in \mathbb{R}^{n\times n}$ is the matrix of entries $K_{i,j} := k(x_i, x_j)$ and $y := (y_1, \dots, y_n) \in \mathbb{R}^n$. A vector $a \in \mathbb{R}^n$ is a solution of Eq. (D.I.4) if and only if it verifies:

$$K\left((K + \lambda I)a - y\right) = 0. \tag{D.I.5}$$

Proof. We begin with the second statement about solutions of Eq. (D.I.4). By the symmetry and positivity properties of k (Proposition D.I.1 items 1) and 2)), the matrix K is symmetric positive semi-definite, and thus the energy $J := a \longmapsto \lambda a^\top K a + \|Ka - y\|_2^2$ is convex. Furthermore, it is differentiable and we compute $\nabla J(a) = 2\lambda K a + 2K(Ka - y)$, and Eq. (D.I.5) is equivalent to the condition $\nabla J(a) = 0$. Since K is symmetric positive semi-definite, the matrix $K + \lambda I$ is symmetric positive definite thanks to the assumption $\lambda > 0$, and thus the the vector $a_0 := (K + \lambda I)^{-1} y$ is a solution of Eq. (D.I.4).

Now we consider the element $h_0 := \sum_{i=1}^n a_i^{(0)} k(\cdot, x_i) \in W$. For any $h \in H$, take its orthogonal projection $h_W := P_W(h)$ onto W (using the closedness and convexity of W and the Hilbert projection Theorem ([Rud87, Theorem 4.11])). We have $\|h_W\|_H \leq \|h\|_H$ and since $h - h_W \in W^\perp$, we deduce from Proposition D.I.2 that $h_W(x_i) = h(x_i)$ for all $i \in [\![1, n]\!]$. This shows that h_W has lower cost than h in the kernel regression problem:

$$\lambda\|h_W\|_H^2 + \sum_{i=1}^n (h_W(x_i) - y_i)^2 \leq \lambda\|h\|_H^2 + \sum_{i=1}^n (h(x_i) - y_i)^2,$$

with a strict inequality if $h \notin W$, and thus if a solution of the kernel regression problem exists, it must be in W. Given the definition of W, we conclude that $h \in H$ is a solution of Eq. (D.I.3) if and only if it is in W and is a solution of Eq. (D.I.3), which is equivalent to being of the form $h = \sum_{i=1}^n a_i k(\cdot, x_i)$ for some $a \in \mathbb{R}^n$ solution of Eq. (D.I.4), concluding the proof. □

D.I.3 Kernel Mean Embedding and Maximum Mean Discrepancy

A cornerstone application of RKHS theory in Machine Learning is the Kernel Mean Embedding (KME), which embeds a probability measure μ into an RKHS H. The norm of of the difference between two embeddings is then a measure of discrepancy between the two measures, which is called the Maximum Mean Discrepancy (MMD) [Gre+06; Smo+07; Mua+17]. In this section, we provide a simple definition of the KME without the use of Bochner integrals (see [Hyt+16, Section 1.2.a], or [CS08, Appendix A.5.4] for references on this notion). We begin with a technical lemma that will allow us to define the KME. Given a measurable space $\mathcal{X}$, we say that a kernel $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ is measurable if for every $x \in \mathcal{X}$, the function $k(\cdot, x) : \mathcal{X} \longrightarrow \mathbb{R}$ is measurable. We will write $\mathcal{P}_k(\mathcal{X})$ the set of probability measures μ on $\mathcal{X}$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\mu(\mathrm{d}x) < +\infty$.

Lemma D.I.1. Let $(H, \langle\cdot,\cdot\rangle_H)$ be a RKHS with a measurable kernel k. Let $\mu \in \mathcal{P}_k(\mathcal{X})$, then the linear map defined by:

$$\ell_\mu := \left\{ \begin{array}{ccc} H & \longrightarrow & \mathbb{R} \\ h & \longmapsto & \int_{\mathcal{X}} h \mathrm{d}\mu \end{array} \right. \tag{D.I.6}$$

is continuous.

Proof. First, we show that ℓ_μ is well-defined. Thanks to the Moore-Aronszajn Theorem, we can write any $h \in H$ as a limit (in H and in particular pointwise) of functions $h_n \in H$ of the form $h_n = \sum_{i=1}^{m_n} a_i^{(n)} k(\cdot, x_i^{(n)})$, and thus h is measurable (see [CS08, Lemma 4.24] for additional details).

We now show that for any $h \in H$, we have $h \in L^1(\mu)$, using the reproducing property (Proposition D.I.1 item 4)), the Cauchy-Schwarz inequality and the assumption $\mu \in \mathcal{P}_k(\mathcal{X})$:

$$\begin{aligned} \int_{\mathcal{X}} |h(x)| \mathrm{d}\mu(x) &= \int_{\mathcal{X}} |\langle h, k(\cdot, x)\rangle_H| \mathrm{d}\mu(x) \\ &\leq \int_{\mathcal{X}} \|h\|_H \|k(\cdot, x)\|_H \mathrm{d}\mu(x) \\ &= \|h\|_H \int_{\mathcal{X}} \sqrt{k(x,x)} \mathrm{d}\mu(x) < +\infty. \end{aligned} \tag{D.I.7}$$

The left hand-side term exists in $\mathbb{R}_+ \cup \{+\infty\}$ by measurability, and the computations above show that it is finite. Thus, ℓ_μ is well-defined, and by linearity of integration, it is clearly linear. We now show continuity using Eq. (D.I.7), which yields that for every $h \in H$, we have $|\ell_\mu(h)| \leq \|h\|_H \int_{\mathcal{X}} \sqrt{k(x,x)} \mathrm{d}\mu(x)$, which proves (Lipschitz) continuity. □

Thanks to Lemma D.I.1, we can define the Kernel Mean Embedding of a probability measure μ as the Riesz representation of ℓ_μ.

Definition D.I.5. Let $(H, \langle\cdot,\cdot\rangle_H)$ be an RKHS with a measurable kernel k, and let $\mu \in \mathcal{P}_k(\mathcal{X})$. The Kernel Mean Embedding of μ is the element of H defined by: $\mathrm{M}(\mu) := R\ell_\mu \in H$, where R is the Riesz representation operator and $\ell_\mu \in H'$ is defined in Eq. (D.I.6).

The KME allows comparison of two probability measures using the norm of the difference of their embeddings in H:

Definition D.I.6. Let $(H, \langle\cdot,\cdot\rangle_H)$ be an RKHS with a measurable kernel k and $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$. The Maximum Mean Discrepancy (MMD) between μ and ν is defined as:

$$\mathrm{MMD}(\mu, \nu) := \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H. \tag{D.I.8}$$

We can rewrite the MMD as an Integral Probability Metric [Mül97]:

Proposition D.I.3. For probability measures $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$, we have:

$$\mathrm{MMD}(\mu, \nu) = \sup_{h \in H, \|h\|_H \leq 1} \int_{\mathcal{X}} h(x) \mathrm{d}\mu(x) - \int_{\mathcal{X}} h(x) \mathrm{d}\nu(x). \tag{D.I.9}$$

Proof. By definition of the KME, we have:

$$\sup_{h \in H, \|h\|_H \leq 1} \int_{\mathcal{X}} h(x) \mathrm{d}\mu(x) - \int_{\mathcal{X}} h(x) \mathrm{d}\nu(x) = \sup_{h \in H, \|h\|_H \leq 1} \langle h, \mathrm{M}(\mu) - \mathrm{M}(\nu)\rangle_H,$$

and the supremum of the right hand-side is attained for $h^* = \frac{\mathrm{M}(\mu) - \mathrm{M}(\nu)}{\|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H}$, called the "witness function" of the MMD. We then compute:

$$\langle h^*, \mathrm{M}(\mu) - \mathrm{M}(\nu)\rangle_H = \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H = \mathrm{MMD}(\mu, \nu),$$

concluding the proof. □

It is clear that the MMD verifies the non-negativity, symmetry and triangle inequality axioms for a distance, and that if $\mu = \nu$ then $\mathrm{MMD}(\mu, \nu) = 0$. The converse is not true in general, and is seen in literature as a property of the kernel k used to define the MMD (see [Sri+10; SFL11]):

Definition D.I.7. A measurable kernel k is said to be characteristic if for any $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$, we have $\mathrm{MMD}(\mu, \nu) = 0$ if and only if $\mu = \nu$ (for the MMD associated to k).

We now focus on the case of discrete measures and take $\mu := \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}_k(\mathcal{X})$. By linearity of the Riesz operator we have:

$$\mathrm{M}(\mu) = R\ell_\mu = \sum_{i=1}^n a_i R\ell_{\delta_{x_i}} = \sum_{i=1}^n a_i R\left(h \longmapsto h(x_i)\right) = \sum_{i=1}^n a_i k(\cdot, x_i),$$

where we used Proposition D.I.1 item 3). Given now two empirical probability measures $\mu := \sum_{i=1}^n a_i \delta_{x_i}$, $\nu := \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}_k(\mathcal{X})$, we have by the self-reproducing property (Proposition D.I.1 item 5)):

$$\begin{aligned}
\mathrm{MMD}^2(\mu, \nu) &= \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H^2 \\
&= \left\langle \sum_{i=1}^n a_i k(\cdot, x_i) - \sum_{j=1}^n b_j k(\cdot, y_j), \sum_{i'=1}^n a_{i'} k(\cdot, x_{i'}) - \sum_{j'=1}^n b_{j'} k(\cdot, y_{j'}) \right\rangle_H \\
&= \sum_{i=1}^n \sum_{i'=1}^n a_i k(x_i, x_{i'}) a_{i'} - 2 \sum_{i=1}^n \sum_{j=1}^m a_i k(x_i, y_j) b_j + \sum_{j=1}^m \sum_{j'=1}^m b_j k(y_j, y_{j'}) b_{j'}.
\end{aligned}$$

Writing the vectors $a := (a_1, \cdots, a_n) \in \mathbb{R}^n$ $b := (b_1, \cdots, b_m) \in \mathbb{R}^m$ and the matrices $K_{xx} := [k(x_i, x_{i'})]_{i,i'} \in \mathbb{R}^{n \times n}$, $K_{xy} := [k(x_i, y_j)]_{i,j} \in \mathbb{R}^{n \times m}$ and $K_{yy} := \left[k(y_j, y_{j'})\right]_{j,j'} \in \mathbb{R}^{m \times m}$, we can rewrite the MMD as:

$$\mathrm{MMD}^2(\mu, \nu) = a^\top K_{xx} a - 2 a^\top K_{xy} b + b^\top K_{yy} b.$$

D.II

On Gradients of Convex Functions in RKHS

D.II.1 An Existential Question . . . 329
D.II.2 Heuristic of an RKHS Cone of Convex Functions . . . 330
D.II.3 Impossibility of the Vector-Field-Centric Heuristic . . . 331
D.II.4 Impossibility of the Potential-Centric Heuristic . . . 333
D.II.5 Theoretical Comment . . . 334

Abstract

In this chapter, we investigate the use of kernels to represent gradients of convex functions, which are widespread in Optimal Transport. We propose a heuristic for minimising an objective over an RKHS cone of gradients of convex functions, where the objective depends on a the value of the function at a finite set of points. We present two natural ideas to construct finite-dimension RKHS cones of gradients of convex functions, and provide unfortunate negative existence results for both.

This chapter is based on joint work with Joan Glaunès.

D.II.1 An Existential Question

Kernel Regression [Nad64; Wat64] is a well-established method consisting in minimising a certain energy over functions $\mathbb{R}^d \longrightarrow \mathbb{R}^d$ in a Reproducing Kernel Hilbert Space (RKHS). When the functional depends only on the values of the function at a finite number of points, this problem can be solved efficiently using the kernel interpolation principle, as summarised in Eq. (D.II.1):

$$\operatorname*{argmin}_{f\in H} \lambda\|f\|_H^2+\mathcal{F}(f(x_i)_{i\in[\![1,n]\!]}) \supset \operatorname*{argmin}_{(\alpha_1,\cdots,\alpha_n)\in\mathbb{R}^d} \lambda\left\|\sum_{i=1}^n K_H(\cdot,x_i)\alpha_i\right\|_H^2+\mathcal{F}\left(\left(\sum_{j=1}^n K_H(x_i,x_j)\alpha_j\right)_{i=1}^n\right), \tag{D.II.1}$$

where H is an RKHS of functions $\mathbb{R}^d \longrightarrow \mathbb{R}$ with kernel K_H, and $\mathcal{F}$ is a cost on the values of f at the points $x_1,\cdots,x_n \in \mathbb{R}^d$. The kernel interpolation principle states that the minimisation problem over H is equivalent to a minimisation problem over the finite dimensional space of functions that are linear combinations of the features $K_H(\cdot,x_i)$.

In this chapter, we will discuss a question of existence of suitable RKHS spaces for optimisation over gradients of convex functions. A possible application would be a kernel method for solving a discretised and relaxed Monge formulation of Optimal Transport:

$$\operatorname*{argmin}_{\substack{T:\mathbb{R}^d\longrightarrow\mathbb{R}^d\\ T=\nabla\varphi,\ \varphi\ \text{convex}}} \sum_{i=1}^n a_i\|x_i-T(x_i)\|_2^2+\lambda \mathrm{W}\left(\sum_{i=1}^n a_i\delta_{T(x_i)},\sum_{j=1}^m b_j\delta_{y_j}\right). \tag{D.II.2}$$

Eq. (D.II.2) is a relaxed version of the Monge problem between two measures $\mu=\sum_{i=1}^n a_i\delta_{x_i}$ and $\nu=\sum_{j=1}^m b_j\delta_{y_j}$, where the constraint $T\#\mu=\nu$ is relaxed to a penalisation term $\mathrm{W}(T\#\mu,\nu)$,

where W is a transport cost, for example the 2-Wasserstein distance (squared). Given an absolutely continuous probability measure μ on $\mathbb{R}^d$, Brenier's theorem [Bre91] states that there exists a unique optimal transport map T from μ to ν, and this map is the gradient of a convex function. In Eq. (D.II.2), we seek a gradient of a convex function $T = \nabla\varphi$ that matches the discrete measure μ as as closely as possible to ν (equality $T\#\mu = \nu$ may be impossible). To leverage kernel interpolation techniques, we would like to find a suitable RKHS space in which the constraint that T be the gradient of a convex function can be written as optimisation over a finite-dimensional cone.

To formulate the question over optimising over gradients of convex functions within an RKHS, we will consider two viewpoints: optimising over vector fields $v : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ that are in a RKHS and constrained to be gradients of convex functions, or optimising over potentials $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ that are in a RKHS of $\mathcal{C}^1$ functions and constrained to be convex. Throughout this chapter, we will focus on the case of rotation and translation invariant kernels, i.e. kernels of the form $K(y, x) = k(\|x - y\|)$, since we are interested in applications in Optimal Transport, where such properties are natural.

Vector Field Viewpoint. First, we will look for an RKHS space V of kernel K_V of vector fields $v : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. In Section D.II.2, we will present a heuristic based on kernel interpolation to reduce a problem of the form:

$$\operatorname*{argmin}_{v=\nabla\varphi\in V,\ \varphi \text{ convex}} \mathcal{F}(v(x_i)_{i\in[\![1,n]\!]}),$$

where $\mathcal{F}$ is a cost on the values of v at the points $x_1, \cdots, x_n \in \mathbb{R}^d$, into a problem over finite-dimensional cone:

$$\operatorname*{argmin}_{v\in\mathcal{C}} \lambda\|v\|_V^2 + \mathcal{F}(v(x_i)_{i\in[\![1,n]\!]}), \quad \mathcal{C} := \left\{ \sum_{i=1}^n K_V(\cdot, x_i)\alpha_i \ : \forall i \in [\![1, n]\!],\ \alpha_i \in \mathbb{R}^d_+ \right\}.$$

The question that arises is the existence of an RKHS V of kernel K_V such that for any positive coefficients $\alpha \in \mathbb{R}^d_+$ and point $x \in \mathbb{R}^d$, the feature map $K_V(\cdot, x)\alpha$ is the gradient of a convex function. In Section D.II.3, we show that these conditions cannot yield an RKHS V of nonconstant functions.

Potential Viewpoint. Second, we focus only on the potentials and look for an RKHS H of $\mathcal{C}^1$ functions $\mathbb{R}^d \longrightarrow \mathbb{R}$ and use kernel interpolation to simplify the following problem:

$$\operatorname*{argmin}_{h\in H \text{ convex}} \lambda\|h\|_H^2 + \mathcal{F}(\nabla h(x_i)_{i\in[\![1,n]\!]}).$$

In Section D.II.4, we show that the kernel interpolation method extends to the case of interpolation of gradients, and we propose the heuristic of optimising over the following cone:

$$\mathcal{C} := \left\{ \sum_{i=1}^n \nabla K_H(\cdot, x_i) \cdot \alpha_i \ : \ \forall i \in [\![1, n]\!],\ \alpha_i \in \mathbb{R}^d_+ \right\},$$

and using techniques similar to those of Section D.II.3, we show that the functions $\nabla K_H(\cdot, x_i)\cdot\alpha_i$ cannot be non-constant and convex.

D.II.2 Heuristic of an RKHS Cone of Convex Functions

The idea behind this section is the following: assume the existence of an RKHS space V of functions $\mathbb{R}^d \longrightarrow \mathbb{R}^d$ of kernel $K_V : V^2 \longrightarrow S_d(\mathbb{R})$ such that for coefficients $\alpha \in \mathbb{R}^d_+$ (or, more generally, a cone whose linear span is $\mathbb{R}^d$), the associated elementary functions $K_V(\cdot, x)\alpha$ are gradients of convex functions. Consider problems of the form

$$\operatorname*{argmin}_{v=\nabla\varphi,\ \varphi \text{ convex}} \mathcal{F}(v(x_i)_{i\in[\![1,n]\!]}),$$

where $\mathcal{F}$ is a cost on the values of v at the points $x_1, \cdots, x_n \in \mathbb{R}^d$. This problem can be solved efficiently using the kernel interpolation principle: we consider a regularised problem

$$\underset{\substack{v \in V, \\ v=\nabla\varphi,\ \varphi \text{ convex}}}{\operatorname{argmin}} \quad \lambda \|v\|_V^2 + \mathcal{F}(v(x_i)_{i \in [\![1,n]\!]}). \tag{D.II.3}$$

Consider the following cone of V comprised of gradients of convex functions:

$$\mathcal{C} := \left\{ \sum_{i=1}^{n} K_V(\cdot, x_i)\alpha_i \ : \forall i \in [\![1,n]\!],\ \alpha_i \in \mathbb{R}^d_+ \right\},$$

and $W := \operatorname{Span} \mathcal{C}$ its linear span (which amounts to taking the α_i in $\mathbb{R}^d$ instead). If v^* is a solution of Eq. (D.II.3), then let v_W be its orthogonal projection on W. We shall show that $\forall i \in [\![1,n]\!],\ v^*(x_i) = v_W(x_i)$ by showing the property

$$W^\perp = V_0 := \{ v \in V \ :\ \forall i \in [\![1,n]\!],\ v(x_i) = 0 \}.$$

Indeed,

$$\begin{aligned}
v \in V_0 &\iff \forall i \in [\![1,n]\!],\ v(x_i) = 0 \\
&\iff \forall i \in [\![1,n]\!],\ \forall \alpha \in \mathbb{R}^d,\ v(x_i) \cdot \alpha = 0 \\
&\iff \forall i \in [\![1,n]\!],\ \forall \alpha \in \mathbb{R}^d,\ \delta^\alpha_{x_i} v = 0 \\
&\iff \forall i \in [\![1,n]\!],\ \forall \alpha \in \mathbb{R}^d,\ \langle v, K_V(\cdot, x_i)\alpha \rangle_V = 0 \\
&\iff v \in W^\perp,
\end{aligned}$$

where δ^α_x is the linear form $v \longmapsto v(x) \cdot \alpha$, whose Riesz representation in V is $K_V(\cdot, x)\alpha$ by the kernel reproducing property. Since $v^* - v_W \in W^\perp = V_0$, it holds that $\forall i \in [\![1,n]\!],\ v^*(x_i) - v_W(x_i) = 0$ hence $v^*(x_i) = v_W(x_i)$. As an orthogonal projection, we have $\|v_W\|_V \leq \|v^*\|_V$, and thus, *if* v_W is the gradient of a convex function, it is also a solution of Eq. (D.II.3). The projection v_W is not necessarily the gradient of a convex function, but a sufficient condition for this to be the case is to have $v_W \in \mathcal{C}$. This invites us to solve instead the simpler problem

$$\underset{v \in \mathcal{C}}{\operatorname{argmin}}\ \lambda \|v\|_V^2 + \mathcal{F}(v(x_i)_{i \in [\![1,n]\!]}), \tag{D.II.4}$$

which can be solved with respect to the coefficients (α_i). Note that Problems Eq. (D.II.4) and Eq. (D.II.3) are not equivalent in general.

D.II.3 Impossibility of the Vector-Field-Centric Heuristic

A natural idea to construct V would be to start from an RKHS H of potentials $h : \mathbb{R}^d \longrightarrow \mathbb{R}$, and have $V := \{\nabla h,\ h \in H\}$. This method is presented in [MG14, Section 4 and Proposition 4.1]. They characterise RKHS which are:

- *curl free*: each element $v \in V$ is the gradient of a function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ of an RKHS H;
- *translation invariant*: the map $v \longmapsto v(\cdot - x_0)$ is an isometry of V.
- *rotation invariant*: the map $v \longmapsto Rf(\cdot R^{-1})$ is an isometry of V for any R an orthogonal matrix. [MG14], this implies that $\mathbf{k}_V(y - x) = k_V(\|y - x\|)$.

Under regularity assumptions, they show that kernels K_H satisfying the above conditions are of the form $K_V(y, x) = -D^2\mathbf{k}_H(x - y)$ (the Hessian matrix of the scalar function $\mathbf{k}_H$ at point $x - y$), with $\mathbf{k}_H(x - y) = k_H(\|x - y\|)$. Unfortunately, there are no kernels of this form that have feature maps which are gradients of convex functions, which is our following statement.

Theorem D.II.1. There does **NOT** exist an RKHS space V of **nonconstant functions** such that:

- $V = \{\nabla h,\ h \in H\}$ is a space of functions $\mathbb{R}^d \longrightarrow \mathbb{R}^d$;
- $H \hookrightarrow \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ is an RKHS of inner product $\langle \cdot, \cdot \rangle_H$ and kernel $K_H(y, x) = \mathbf{k}_H(x - y) = k_H(\|x - y\|)$. Each $h \in H$ verifies $h(0) = 0$.
- The inner product of V is $\langle v_1, v_2 \rangle_V = \langle h_1, h_2 \rangle_H$, where $v_1 = \nabla h_1$ and $v_2 = \nabla h_2$ (the conditions $h_1(0) = 0$ and $h_2(0) = 0$ allow us to choose h_1 and h_2 uniquely).
- For any positive coefficients $\alpha \in \mathbb{R}^d_+$, and point $x \in \mathbb{R}^d$, the feature map $K_V(\cdot, x)\alpha$ is the gradient of a convex function.

Note that we justify the existence of the second-order derivatives of K_H using [MG14, Theorem 2.11], thanks to the hypothesis $H \hookrightarrow \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$.

Proof. First, we show that the kernel of V is $K_V(y, x) = -D^2\mathbf{k}_H(x - y)$, following the method from [MG14, Proposition 4.1] (we use the same computation, but their Proposition itself does not apply exactly here for regularity reasons). For $x, \alpha \in \mathbb{R}^d$ and $v = \nabla h \in V$, the kernel K_V's reproducing property reads

$$\langle K_V(\cdot, x)\alpha, v \rangle_V = \alpha \cdot \nabla h(x),$$

which equates to $\sum_{i=1}^d \alpha_i \partial_{x_i} h(x)$. For $i \in [\![1, d]\!]$, the number $\partial_{x_i} h(x)$ equals $\langle \partial_{x_i} K_H(\cdot, x), h \rangle_H$ by [MG14, Theorem 2.11] since $H \hookrightarrow \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$. Continuing the computation, we obtain:

$$\begin{aligned}
\langle K_V(\cdot, x)\alpha, v \rangle_V &= \alpha \cdot \nabla h(x) \\
&= \sum_{i=1}^d \alpha_i \langle \partial_{x_i} K_H(\cdot, x), h \rangle_H \\
&= \left\langle \sum_{i=1}^d \alpha_i \partial_{x_i} K_H(\cdot, x), h \right\rangle_H \\
&= \left\langle \nabla \sum_{i=1}^d \alpha_i \partial_{x_i} K_H(\cdot, x), \nabla h \right\rangle_V . \qquad \text{(D.II.5)}
\end{aligned}$$

Now we use $K_H(\cdot, x) = \mathbf{k}_H(x - \cdot)$. We denote the variable of $\mathbf{k}_H : \mathbb{R}^d \longrightarrow \mathbb{R}$ as z for the partial derivative notation. For $y \in \mathbb{R}^d$, we have $\partial_{x_i} K_H(y, x) = \partial_{z_i} \mathbf{k}_H(x - y)$, thus:

$$\sum_{i=1}^d \alpha_i \partial_{x_i} K_H(y, x) = \nabla \mathbf{k}_H(x - y) \cdot \alpha.$$

then the gradient with respect to y is

$$\nabla_y \sum_{i=1}^d \alpha_i \partial_{x_i} K_H(y, x) = \nabla_y \left[\nabla \mathbf{k}_H(x - y) \cdot \alpha\right] = -D^2\mathbf{k}_H(x - y)\alpha.$$

To conclude, for any $v \in V$, we have from the previous equation and Eq. (D.II.6)

$$\langle K_V(y, x)\alpha, v \rangle_V = \langle -D^2\mathbf{k}_H(x - y)\alpha, v \rangle_V,$$

and thus we obtain:

$$K_V(y, x)\alpha = \nabla_y \left[\nabla \mathbf{k}_H(x - y) \cdot \alpha\right] = -D^2\mathbf{k}_H(x - y)\alpha. \qquad \text{(D.II.6)}$$

In particular, for $\alpha \in \mathbb{R}^d_+$ and $x \in \mathbb{R}^d$, the feature map $K_V(\cdot, x)\alpha$ is the gradient of $\nabla \mathbf{k}_H(x - \cdot)\alpha$. To show that the space V cannot exist, we will show that the convexity of $\nabla \mathbf{k}_H(x - \cdot)\alpha$ for any $\alpha \in \mathbb{R}^d_+$ implies that it is linear.

Let $\alpha := (\delta_{i,j})_{j\in[\![1,d]\!]}$ for $i \in [\![1,d]\!]$ fixed, and $h := \nabla \mathbf{k}_H(x-\cdot)\cdot\alpha = \partial_{z_i}\mathbf{k}_H(x-\cdot)$. First for $z \in \mathbb{R}^d$,

$$\partial_{z_i}\mathbf{k}_H(z) = \partial_{z_i}\left[k_H(\|z\|)\right] = \frac{z_i}{\|z\|}k'_H(\|z\|).$$

The convexity of $h = \nabla\mathbf{k}_H(x-\cdot)\alpha$ implies the convexity of

$$\psi := z_i \longmapsto \frac{z_i}{\|z\|}k'_H(\|z\|), \quad (z_1,\cdots,z_{i_1},z_{i+1},\cdots,z_d) \in \mathbb{R}^{d-1}.$$

Unfortunately, ψ is an odd convex function, hence $-\psi$ is concave, yet $-\psi = \psi(-\cdot)$, which shows that ψ is both convex and concave, thus linear.

By linear combination, we have shown in particular that for any $\alpha \in \mathbb{R}^d$, $y \longmapsto \nabla\mathbf{k}_H(x-y)\cdot\alpha$ is linear, therefore by Eq. (D.II.6), the feature map $K_V(\cdot,x)\alpha$ is constant (as a gradient of a linear map), thus V is the space of constant functions.

□

D.II.4 Impossibility of the Potential-Centric Heuristic

Another idea would be to use an RKHS H of functions $\mathbb{R}^d \longrightarrow \mathbb{R}$, and formulate the problem with respect to the potential $h : \mathbb{R}^d \longrightarrow \mathbb{R}$:

$$\underset{h\in H \text{ convex}}{\operatorname{argmin}} \quad \lambda\|h\|_H^2 + \mathcal{F}(\nabla h(x_i)_{i\in[\![1,n]\!]}) \tag{D.II.7}$$

Assuming that $H \hookrightarrow \mathcal{C}^1(\mathbb{R}^d,\mathbb{R})$, we can use [MG14, Theorem 2.11] to write a reproducing equation for the derivatives of h:

$$\frac{\partial h}{\partial x_j}(x_i) = \left\langle h, \partial_{x_j}K_H(\cdot,x_i)\right\rangle_H. \tag{D.II.8}$$

Just like in Section D.II.2, we start by looking at the space

$$H_0 := \left\{h \in H \ : \ \forall (i,j)\in[\![1,n]\!]\times[\![1,d]\!],\ \partial_{x_j}h(x_i) = 0\right\},$$

which we show to verify $H_0 = W^\perp$, with

$$\mathcal{C} := \left\{\sum_{i=1}^n \nabla K_H(\cdot,x_i)\cdot\alpha_i \ : \ \forall i\in[\![1,n]\!],\ \alpha_i\in\mathbb{R}^d_+\right\}; \quad W := \operatorname{Span}\mathcal{C}.$$

As previously, we choose the cone $\mathbb{R}^d_+$ for the coefficients since we would like to optimise over a cone of convex functions h. For now, we shall **assume** that each function of $\mathcal{C}$ is convex. To show $H_0 = W^\perp$, we use the derivative reproducing property Eq. (D.II.8):

$$\begin{aligned} h\in H_0 &\iff \forall(i,j)\in[\![1,n]\!]\times[\![1,d]\!],\ \left\langle h,\partial_{x_j}K_H(\cdot,x_i)\right\rangle_H = 0\\ &\iff h\in \operatorname{Span}\left\{\partial_{x_j}K_H(\cdot,x_i),\ (i,j)\in[\![1,n]\!]\times[\![1,d]\!]\right\}^\perp\\ &\iff h\in W^\perp. \end{aligned}$$

Given a solution h^* of Eq. (D.II.7), consider h_W its orthogonal projection on $\mathcal{C}$. Since $h^* - h_W \in W^\perp = H_0$, we have in particular $\forall i\in[\![1,n]\!]$, $\nabla h^*(x_i) = \nabla h_W(x_i)$. As an orthogonal projection, we also have $\|h_W\|_H \le \|h^*\|_H$, showing that *if* h_W were convex, then it would be optimal for Eq. (D.II.7). Similarly to Section D.II.2, this gives the idea of looking for solutions $h\in\mathcal{C}$ instead of W. (since by assumption, each function of $\mathcal{C}$ is convex). Again, restricting the variable to $h\in\mathcal{C}$ yields a *different* problem in general.

We would like to consider the natural case of translation and rotation-invariant kernels, which are necessarily of the form $K_H(y,x) = k_H(\|x-y\|)$, as characterised by [MG14, Section 3.2]. (more conditions are required to fully characterise this property). Unfortunately, the computation done in Theorem D.II.1 show that in this case, the function $\partial_{x_j}K_H(\cdot,x_i)$ cannot be convex (more precisely, if it is convex then it is constant).

D.II.5 Theoretical Comment

In order to analyse the heuristic in the problem reduction idea presented in Section D.II.2, we would need to answer the following question: given $\mathcal{K} \subset V$ a closed cone of an RKHS (or just a Hilbert space) V (such as the cone of functions that are gradients of convex functions), and $W \subset V$ a (finite-dimensional) linear subspace, and where $W = \text{Span } \mathcal{C}$, where $\mathcal{C} \subset \mathcal{K}$ is a cone; for which $f \in \mathcal{K}$ does it hold $\text{Proj}_W(f) \in \mathcal{C}$?

D.III

Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces

D.III.1 Introduction . . . 336
D.III.1.1 Kernels in Practice and Related Works 336
D.III.1.2 Elements of RKHS Theory 337
D.III.1.3 Chapter Outline and Contributions 338
D.III.2 Explicit Universal Taylor and Radial Kernels on a Compact Metric Space . . . 339
D.III.2.1 Injection of $\mathcal{X}$ into ℓ^2 339
D.III.2.2 Universal Kernels on $\mathcal{X}$ 340
D.III.3 Approximate Universal Kernels . . . 341
D.III.3.1 Constructing a Smaller RKHS $\hat{H}$ 342
D.III.3.2 Showing that $\hat{H}$ is Approximately Universal 346
D.III.3.3 An Approximate Universal Truncated Kernel 349

Abstract

Universal kernels, whose Reproducing Kernel Hilbert Space is dense in the space of continuous functions are of great practical and theoretical interest. In this chapter, we introduce an explicit construction of universal kernels on compact metric spaces. We also introduce a notion of approximate universality, and construct tractable kernels that are approximately universal. This chapter is based on the paper:

[Tan25] Eloi Tanguy.
"Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces".
arxiv preprint 2506.03661 (Jun. 2025).

D.III.1 Introduction

D.III.1.1 Kernels in Practice and Related Works

Kernels Methods at large are a ubiquitous tool in statistics, starting with Kernel Density Estimation [Ros56; Par62] and Kernel Regression [Nad64; Wat64]. For an overview of the use of kernel methods in statistics and probability, we refer to the monograph [BT11]. In Machine Learning, the first uses of kernels hinged on the "kernel trick" [Aiz64; SSM98], which allows high expressivity of models without the need of an explicit feature map into the underlying infinite-dimensional space. A cornerstone model is the Support Vector Machine [CV95], whose statistical properties have garnered extensive attention, see for example the monograph [CS08]. A useful tool is the Kernel Mean embedding (we refer to the review [Mua+17]) which maps a measure μ to a point $M(\mu)$ in a Hilbert space of features, and can be used to compare measures

with the Maximum Mean Discrepancy defined as $\mathrm{MMD}(\mu,\nu) = \|M(\mu) - M(\nu)\|_H$ which fostered numerous applications [Góm+09; Zha+11; Mua+12; FSG13; Gre+12; Dor+14; Li+17]. Theoretical guarantees for the MMD depending on properties of the kernel have been reviewed in [SFL11].

From a theoretical standpoint, Reproducing Kernel Hilbert Spaces (RKHS) introduced by Aronszajn [Aro50] have been the object of several monographs [SS02; CS08; SS16]. Some questions remain open, in particular constructing suitable kernels on non-euclidean metric spaces is a challenging problem that is the subject of ongoing research. For compact metric spaces, [CS10] show the existence of universal kernels (i.e. such that the associated RKHS is dense in the space of continuous functions) when the space is continuously embedded into a separable Hilbert space, and [SZ21] relate the notions of universality and strictly proper kernel scores. On complete Riemannian manifolds, [Jay+15, Theorem 6.2] observe that the natural Gaussian kernel $k(x,y) = \exp(-s\ d(x,y)^2)$ is indeed a kernel only in the very restricted case where the manifold is isometric to $\mathbb{R}^d$. On Hilbert and Banach spaces, [ZGD24] introduce radial kernels and show universality-adjacent properties. Regarding universality, [MXZ06] study conditions on the feature maps that ensure universality.

Our contribution first consists in an *explicit* construction of universal kernels on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$, in some sense extending [CS10] whose construction is not explicit and relied on the existence of an embedding. The constructed kernels use known kernels known as *Taylor* and *radial* kernels, which are defined on compact subsets of separable Hilbert spaces. Noticing that our kernels are not tractable in practice, we introduce a notion of *approximate universality* and construct other explicit kernels that are approximately universal and tractable.

D.III.1.2 Elements of RKHS Theory

For a set $\mathcal{X}$, a *kernel* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *positive-definite symmetric* function, which is to say a function that verifies $k(x,y) = k(y,x)$ and:

$$\forall n \in \mathbb{N}^*,\ \forall (x_1, \cdots, x_n) \in \mathcal{X}^n,\ \forall a \in \mathbb{R}^n,\ \sum_{i=1}^{n}\sum_{j=1}^{n} a_i k(x_i, x_j) a_j \geq 0.$$

By the Moore-Aronszajn theorem [Aro50], there exists a unique Hilbert space $(H, \langle\cdot,\cdot\rangle_H)$ of functions $\mathcal{X} \longrightarrow \mathbb{R}$, such that H contains all basic functions $k(\cdot, x)$, and its inner product is characterised by the "reproducing property" $\langle k(\cdot,x), k(\cdot,y)\rangle_H = k(x,y)$. Denoting by $\overline{\mathrm{Span}}$ the Hilbertian completion of the linear span of a set, it follows that $H = \overline{\mathrm{Span}}\{k(\cdot,x),\ x \in \mathcal{X}\}$. The space H is referred to as the Reproducing Kernel Hilbert Space (RKHS) associated to the kernel k. The reproducing property of the kernel implies that for any $h \in H$ and $x \in \mathcal{X}$, we have $\langle h, k(\cdot,x)\rangle_H = h(x)$.

If k is continuous (w.r.t. a metric on $\mathcal{X}$), the RKHS H is contained in the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$, denoted $\mathcal{C}(\mathcal{X})$. In this work, we will always consider continuous kernels. Some continuous kernels have an additional property called *universality*:

Definition D.III.1. A continuous kernel k on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ is said to be *universal* if the RKHS H is dense in $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$, the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ equipped with the supremum norm. In other words, for any $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $h \in H$ such that $\|f - h\|_\infty \leq \varepsilon$.

Another equivalent definition of kernels uses the notion of feature map / feature space pairs: through these lens, a kernel is any map $\mathcal{X}^2 \longrightarrow \mathbb{R}$ such that there exists a Hilbert space H_0 and a map $\Phi_0 : \mathcal{X} \longrightarrow H_0$ such that Eq. (D.III.1) holds.

$$\forall x, y \in \mathcal{X},\ k(x,y) = \langle \Phi_0(x), \Phi_0(y)\rangle_{H_0}. \tag{D.III.1}$$

The pair (Φ_0, H_0) is called a *feature map / feature space pair* (or simply *feature pair*) for k, and any kernel can be written in this form ([CS08, Theorem 4.16]). The associated RKHS is then

defined as:

$$H = \{x \longmapsto \langle h_0, \Phi_0(x)\rangle_{H_0},\ h_0 \in H_0\}. \tag{D.III.2}$$

The RKHS H in Eq. (D.III.2) is unique ([CS08, Theorem 4.21]), and equal to $\overline{\text{Span}}\,\{k(\cdot, x),\ x \in \mathcal{X}\}$ as stated above. The *canonical feature map* is defined as $\Phi(x) = k(\cdot, x)$, and the pair (Φ, H) is called the *canonical feature pair* for k.

From the space viewpoint, an RKHS can equivalently be defined as a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$ in which the evaluation $\delta_x : h \longmapsto h(x)$ is continuous for all $x \in \mathcal{X}$, as is done in [CS08, Section 4.2]. The kernel is then defined as $k(x, y) = \langle L\delta_x, L\delta_y\rangle_H$, where $L\delta_x \in H$ is the Riesz representation of $\delta_x \in H'$. In this chapter, we stick to the (equivalent) kernel viewpoint.

For a compact metric space (E, d_E), we will denote by $\text{diam}(E)$ its diameter, which is defined by $\text{diam}(E) := \max_{(x,y)\in E^2} d_E(x, y)$. Throughout this work, $\mathcal{X}$ will be assumed to be a compact metric space, and we denote $D_{\mathcal{X}} := \text{diam}(\mathcal{X})$.

The first type of universal kernels of interest in this work are Taylor kernels (see [CS08, Lemma 4.8 and Corollary 4.57] for their study on compact subsets of $\mathbb{R}^d$).

Definition D.III.2. Let $W \subset \ell^2$ be a non-empty compact set and $D_W^2 := \text{diam}(W)^2 > 0$ the square of its diameter. Take a sequence $(a_n)_{n\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ such that $K(t) := \sum_n a_n t^n$ converges absolutely on $[-D_W^2, D_W^2]$. The Taylor kernel associated to K is the map

$$k_W := \left\{ \begin{array}{ccc} W^2 & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & K(\langle u, v\rangle_{\ell^2}) \end{array} \right. . \tag{D.III.3}$$

Taylor kernels are shown to be universal on compact subsets of ℓ^2 in [CS10, Theorem 2.1]. The second type of universal kernels we will consider are radial kernels[1].

Definition D.III.3. Let $W \subset \ell^2$ be a non-empty compact set and $\mu \in \mathcal{M}([0, +\infty))$ a finite Borel measure on $[0, +\infty)$ with $\text{supp}(\mu) \neq \{0\}$. The associated radial function K and the radial kernel k_W are defined as follows:

$$K := \left\{ \begin{array}{ccc} \mathbb{R}_+ & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \int_0^{+\infty} e^{-st} \mathrm{d}\mu(s) \end{array} \right. , \ k_W := \left\{ \begin{array}{ccc} W^2 & \longrightarrow & \mathbb{R} \\ (u, v) & \longmapsto & K(\|u - v\|_{\ell^2}^2) \end{array} \right. . \tag{D.III.4}$$

The universality of radial kernels on W is a consequence of [ZGD24, Proposition 5.2] combined with [SZ21, Theorem 3.13]. Note that the well-known Gaussian (or RBF) kernel $\exp(-\|\cdot - \cdot\|_{\ell^2}^2/(2\sigma^2))$ is a particular radial kernel with $\mu := \delta_{1/(2\sigma^2)}$.

D.III.1.3 Chapter Outline and Contributions

The objective of this chapter is to construct kernels k on a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ that are *universal* (see Definition D.III.1). We also introduce a notion of approximate universality (Definition D.III.4), and introduce other (tractable) explicit kernels $\hat{k}$ and k_t that verify this property.

Construction of universal kernels in Section D.III.2. To construct universal kernels on $\mathcal{X}$, we first introduce an explicit continuous injection $\varphi : \mathcal{X} \longrightarrow \ell^2$ in Proposition D.III.1. Given any universal kernel k_V on $V := \varphi(\mathcal{X}) \subset \ell^2$ we show in Theorem D.III.1 that $k(x, y) := k_V(\varphi(x), \varphi(y))$, is universal on $\mathcal{X}$.

The construction of φ in Section D.III.2 is based on a countable basis of $\mathcal{X}$, and the associated kernel requires inner products in ℓ^2. In Section D.III.3, we explain how we can use instead a (finite) η-covering of $\mathcal{X}$, yielding a finite-dimensional approximation of the embedding φ, with theoretical guarantees. We also investigate the natural idea of truncating the sequence $\varphi(x)$.

[1] Radial kernels can be defined (and shown to be universal) on separable Hilbert spaces and more [ZGD24], but we will use compactness for other reasons, and thus restrict to compact subsets of ℓ^2 for our purposes.

Approximate universal kernels in Section D.III.3. We introduce a notion of *approximate universal kernels* on $\mathcal{X}$, which are kernels $\hat{k}$ of RKHS $\hat{H}$ such that for all $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $\hat{h} \in \hat{H}$ such that $\|f - \hat{h}\|_\infty \leq \varepsilon + \rho(f)$, where $\rho(f) > 0$ is an error term depending on $\hat{k}$ and f. We construct a simpler map $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ as a surrogate for the embedding $\varphi : \mathcal{X} :\longrightarrow \ell^2$, and embed $\mathbb{R}^J$ into ℓ^2 appropriately to compare φ and $B \circ \hat{\varphi}$. This allows us to introduce the kernel $\hat{k}(x, y) := k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y))$ for a compact set $W \supset \varphi(\mathcal{X}) \cup B(\hat{\varphi}(\mathcal{X}))$ and a Taylor or radial kernel k_W on W. In Corollary D.III.1, we provide a tractable (as in numerically computable) expression for $\hat{k}$. Finally, we show in Theorem D.III.2 that $\hat{k}$ is an approximate universal kernel on $\mathcal{X}$ with an explicit error term ρ depending on discretisation parameters and the "complexity" of the function f. In Section D.III.3.3, we introduce a simple truncation of φ which leads to another approximate universal kernel k_t on $\mathcal{X}$.

D.III.2 Explicit Universal Taylor and Radial Kernels on a Compact Metric Space

As a preliminary to our main constructions, we begin with two elementary general properties of RKHS which will be useful throughout this section. We remind that a *homeomorphism* is a continuous bijection with a continuous inverse.

Lemma D.III.1. i) Let $\mathcal{X}, \mathcal{Y}$ be two sets and $k_\mathcal{Y} : \mathcal{Y}^2 \longrightarrow \mathbb{R}$ a kernel on $\mathcal{Y}$, and $\varphi : \mathcal{X} \longrightarrow \mathcal{Y}$ a function. The map $k_\mathcal{X}$ defined in Eq. (D.III.5) is a kernel on $\mathcal{X}$:

$$k_\mathcal{X} := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x, x') & \longmapsto & k_\mathcal{Y}(\varphi(x), \varphi(x')) \end{array} \right. . \tag{D.III.5}$$

ii) If additionally $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces, φ is a homeomorphism and $k_\mathcal{Y} : \mathcal{Y}^2 \longrightarrow \mathbb{R}$ is universal, the kernel $k_\mathcal{X}$ is universal.

Proof. **Proof of i):** We verify immediately using the definition that $k_\mathcal{X}$ is a positive-definite symmetric function on $\mathcal{X}$ using the fact that $k_\mathcal{Y}$ is such on $\mathcal{Y}$.

Proof of ii): Let $H_\mathcal{Y}$ be the unique RKHS associated to the kernel $k_\mathcal{Y}$ on $\mathcal{Y}$, and $\Phi_\mathcal{Y} : \mathcal{Y} \longmapsto H_\mathcal{Y}$ its canonical feature map (i.e. $\forall y \in \mathcal{Y}$, $\Phi_\mathcal{Y}(y) = k_\mathcal{Y}(\cdot, y)$). Since $k_\mathcal{X}(x, x') = \langle \Phi_\mathcal{Y} \circ \varphi(x), \Phi_\mathcal{Y} \circ \varphi(x') \rangle_{H_\mathcal{Y}}$, the map $\Phi_\mathcal{Y} \circ \varphi$ and the space $H_\mathcal{Y}$ are a feature pair for $k_\mathcal{X}$. In the following, given a set $\mathcal{F}$ of functions and g a function, we write $\mathcal{F} \circ g := \{f \circ g,\ f \in \mathcal{F}\}$. By uniqueness ([CS08, Theorem 4.21]), it follows that the RKHS $H_\mathcal{X}$ associated to $k_\mathcal{X}$ can be written

$$H_\mathcal{X} = \{x \longmapsto \langle h_\mathcal{Y}, \Phi_\mathcal{Y} \circ \varphi(x) \rangle_{H_\mathcal{Y}},\ h_\mathcal{Y} \in H_\mathcal{Y}\} = H_\mathcal{Y} \circ \varphi,$$

where the second equality comes from the reproducing property: for any $x \in \mathcal{X}$ and $h_\mathcal{Y} \in H_\mathcal{Y}$, we have $h_\mathcal{Y} \circ \varphi(x) = \langle h_\mathcal{Y}, \Phi_\mathcal{Y} \circ \varphi(x) \rangle_{H_\mathcal{Y}}$. Since φ is a homeomorphism, we also have $\mathcal{C}(\mathcal{X}) = \mathcal{C}(\mathcal{Y}) \circ \varphi$. Now for $\varepsilon > 0$ and $f_\mathcal{X} \in \mathcal{C}(\mathcal{X})$, take $f_\mathcal{Y} := f_\mathcal{X} \circ \varphi^{-1} \in \mathcal{C}(\mathcal{Y})$. By universality of $k_\mathcal{Y}$, there exists $h_\mathcal{Y} \in H_\mathcal{Y}$ such that $\|f_\mathcal{Y} - h_\mathcal{Y}\|_\infty \leq \varepsilon$. Taking $h_\mathcal{X} := h_\mathcal{Y} \circ \varphi \in H_\mathcal{X}$ yields $\|f_\mathcal{X} - h_\mathcal{X}\|_\infty \leq \varepsilon$, and as a result $k_\mathcal{X}$ is universal. □

Lemma D.III.1 is useful for the construction of universal kernels on compact metric spaces $\mathcal{X}$ using universal kernels on another space. In the following, we will consider a space $\mathcal{Y}$ which is a compact subspace of the Hilbert space ℓ^2 of square-summable sequences.

D.III.2.1 Injection of $\mathcal{X}$ into ℓ^2

Let $(\mathcal{X}, d_\mathcal{X})$ be a non-empty compact metric space, and let $D_\mathcal{X} > 0$ be its diameter. We take a *basis* of $\mathcal{X}$, i.e. a countable sequence $(x_n)_{n \in \mathbb{N}}$ such that for any $x \in \mathcal{X}$ and $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $d_\mathcal{X}(x, x_n) \leq \varepsilon$. Using a basis, we construct an implicit continuous injection φ from $\mathcal{X}$ into ℓ^2 (Proposition D.III.1), then use universal kernels on $V := \varphi(\mathcal{X})$ to build a universal kernel k on $\mathcal{X}$ in Theorem D.III.1. In Fig. D.III.1, we illustrate the injection.

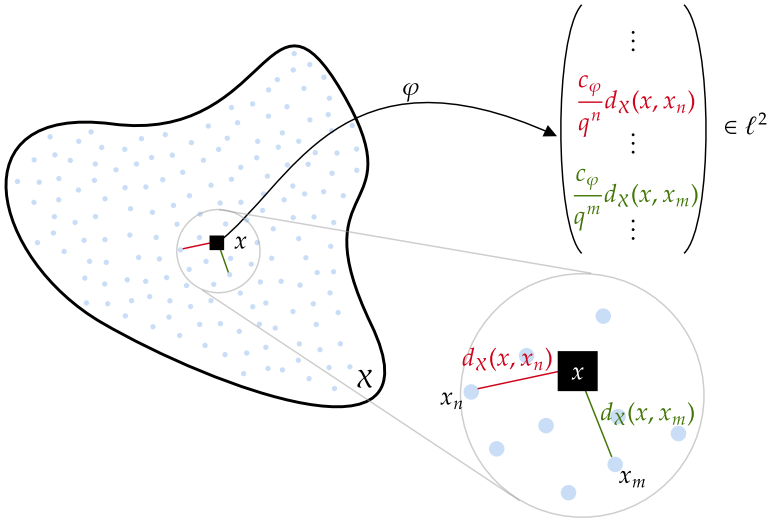


Figure D.III.1: Given a basis $(x_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$, the mapping $\varphi : \mathcal{X} \longrightarrow \ell^2$ maps a point $x \in \mathcal{X}$ to the sequence of its distances to the points of the basis.

Proposition D.III.1. Let (x_n) a basis of $\mathcal{X}$ and $q > 1$. The map

$$\varphi := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \ell^2 \\ x & \longmapsto & \left(\frac{c_\varphi d_{\mathcal{X}}(x, x_n)}{q^n} \right)_{n\in\mathbb{N}} \end{array} \right. , \quad c_\varphi := \frac{\sqrt{q^2-1}}{q}$$

is 1-Lipschitz and injective.

Proof. The fact that $\varphi(\mathcal{X}) \subset \ell^2$ comes from the compactness of $\mathcal{X}$. Take now $x, y \in \mathcal{X}$:

$$\|\varphi(x) - \varphi(y)\|_{\ell^2}^2 = c_\varphi^2 \sum_{n=0}^{+\infty} \frac{|d_{\mathcal{X}}(x, x_n) - d_{\mathcal{X}}(y, x_n)|^2}{q^{2n}} \leq c_\varphi^2 \sum_{n=0}^{+\infty} \frac{d_{\mathcal{X}}(x, y)^2}{q^{2n}} = \frac{c_\varphi^2 q^2}{q^2-1} d_{\mathcal{X}}(x, y)^2,$$

showing 1-Lipschitzness. As for injectivity, consider $x \neq y \in \mathcal{X}^2$ and $\varepsilon := d_{\mathcal{X}}(x, y)/3 > 0$. Since (x_n) is a basis of $\mathcal{X}$, there exists $n \in \mathbb{N}$ such that $d_{\mathcal{X}}(x, x_n) \leq \varepsilon$. The triangle inequality then shows

$$d_{\mathcal{X}}(y, x_n) \geq \underbrace{d_{\mathcal{X}}(y, x)}_{=3\varepsilon} - \underbrace{d_{\mathcal{X}}(x, x_n)}_{\in[0,\varepsilon]} \geq 2\varepsilon,$$

and thus $|\underbrace{d_{\mathcal{X}}(y, x_n)}_{\geq 2\varepsilon} - \underbrace{d_{\mathcal{X}}(x, x_n)}_{\in[0,\varepsilon]}| \geq \varepsilon$, allowing us to conclude

$$\|\varphi(y) - \varphi(x)\|_{\ell^2}^2 \geq c_\varphi^2 \frac{|d_{\mathcal{X}}(y, x_n) - d_{\mathcal{X}}(x, x_n)|^2}{q^{2n}} \geq c_\varphi^2 \frac{\varepsilon^2}{q^{2n}} > 0.$$

□

D.III.2.2 Universal Kernels on $\mathcal{X}$

We can now build universal kernels $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ using φ and a universal kernel $k_W : W^2 \longrightarrow \mathbb{R}$ (for example Taylor or radial) on W a compact subset of ℓ^2 containing $V := \varphi(\mathcal{X})$. The technique follows closely that of [CS10, Theorem 2.2]. Note that thanks to the 1-Lipschitzness of φ, we have $\operatorname{diam}(\varphi(\mathcal{X})) \leq \operatorname{diam}(\mathcal{X}) =: D_{\mathcal{X}}$.

Theorem D.III.1. Let $V := \varphi(\mathcal{X}) \subset \ell^2$ and W be a compact subset of ℓ^2 containing V. Consider $k_W : W^2 \longrightarrow \mathbb{R}$ a universal kernel on W (e.g. Taylor as in Definition D.III.2 or radial as in Definition D.III.3). The kernel

$$k := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & k_W(\varphi(x), \varphi(y)) \end{array} \right.$$

is universal on $\mathcal{X}$.

Proof. We introduce $k_V : V^2 \longrightarrow \mathbb{R}$ the restriction of k_W to V^2. By [CS08, Lemma 4.55 item iii)], k_V remains universal. Since $\mathcal{X}$ is a compact metric space and ℓ^2 is Hausdorff, the co-restriction of φ to V denoted $\varphi_V : \mathcal{X} \longrightarrow V$ is a homeomorphism. Noticing that $k = (x,y) \in \mathcal{X}^2 \longmapsto k_V(\varphi_V(x), \varphi_V(y))$, by Lemma D.III.1, k is thus a universal kernel on $\mathcal{X}$. □

A strictly convex functional on $\mathcal{P}(\mathcal{X})$. We consider the set $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$. As a universal kernel, k is also *characteristic* (see [SFL11] and use the compactness of $\mathcal{X}$), which is to say that the map

$$M := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & H \\ \mu & \longmapsto & \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mu(x) \end{array} \right. ,$$

known as the *kernel mean embedding* [Sri+10], is injective. One can show that the map

$$F := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & \mathbb{R}_+ \\ \mu & \longmapsto & \|M(\mu)\|_H^2 \end{array} \right.$$

is continuous with respect to the weak convergence of measures (apply [HC11, Theorem A.1] using that $x \longmapsto k(\cdot, x)$ is continuous and bounded). Furthermore, by linearity of M and strict convexity of $\|\cdot\|_H^2$, the function F is strictly convex with respect to the "vertical" convex combination of probability measures:

$$\forall \mu, \nu \in \mathcal{P}(\mathcal{X}),\ \forall t \in (0,1),\ F((1-t)\mu + t\nu) < (1-t)F(\mu) + tF(\nu).$$

Note that the fact that M is injective is required to prove *strict* convexity.

D.III.3 Approximate Universal Kernels

In practice, the function φ introduced in Proposition D.III.1 is not tractable (in the sense that computing and storing a full sequence $\varphi(x) \in \ell^2$ is numerically impossible), limiting the use of the kernels proposed in Theorem D.III.1 in their exact formulation. The natural idea of truncating the sequence $\varphi(x)$ to the first N terms is tackled later in Section D.III.3.3, we begin with the main approach of the chapter, which relies on a discretisation of the space $\mathcal{X}$.

We will now introduce a family of tractable kernels which are approximately universal on $\mathcal{X}$. Throughout this section, the kernels k_W on a compact subset W of ℓ^2 that we will consider are Taylor or radial (see Definitions D.III.2 and D.III.3). Our objective is to construct another kernel $\hat{k}$ with a simpler explicit mapping $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$, yielding an RKHS $\hat{H}$ which we will show to be approximately universal in the sense of Definition D.III.4.

Definition D.III.4. Let $\hat{k} : \mathcal{X}^2 \longrightarrow \mathbb{R}$ a kernel on $\mathcal{X}$ of RKHS $\hat{H}$ and $\rho : \mathcal{C}(\mathcal{X}) \longrightarrow \mathbb{R}_+$ an error function. We say that $\hat{k}$ is an *approximate universal kernel* on $\mathcal{X}$ if for all $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $\hat{h} \in \hat{H}$ such that $\|f - \hat{h}\|_\infty \leq \varepsilon + \rho(f)$.

D.III.3.1 Constructing a Smaller RKHS $\hat{H}$

In this section, we provide a principled method to "sub-sample" the sequence $\varphi(x)$: we will begin with a well-chosen finite family $(y_j) \in \mathcal{X}^J$ and construct a well-suited basis $(x_n) \in \mathcal{X}^{\mathbb{N}}$ such that a distance sequence $(d_{\mathcal{X}}(x, x_n))_{n\in\mathbb{N}}$ is adequately approximable by the finite number of distances $(d_{\mathcal{X}}(x, y_j))_{j\in[\![1,J]\!]}$. We illustrate this discretisation concept in Fig. D.III.2.

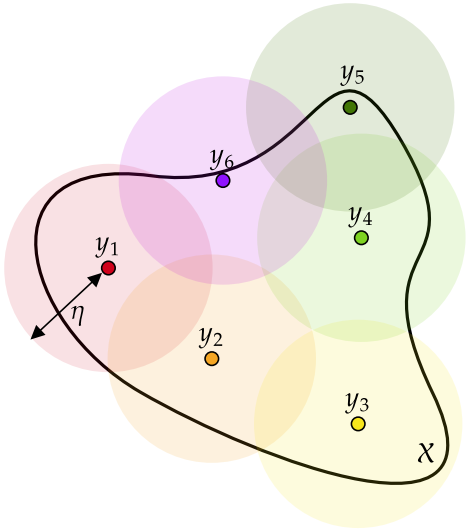


Figure D.III.2: Discretisation of the space $\mathcal{X}$ into a cover of J balls of radius $\eta > 0$ centred at each $(y_j)_{j\in[\![1,J]\!]}$.

Instead of a basis of $\mathcal{X}$, we will now fix $\eta \in (0, D_{\mathcal{X}}]$ and consider $(y_j)_{j\in[\![1,J]\!]}$ a family of distinct points of $\mathcal{X}$ such that the family of (closed) balls $B_{d_{\mathcal{X}}}(y_j, \eta)$ covers $\mathcal{X}$. In Eq. (D.III.6), we introduce a map $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ in the spirit of φ defined in Proposition D.III.1, which we visualise in Fig. D.III.3.

$$\hat{\varphi} := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}^J \\ x & \longmapsto & \left(\dfrac{d_{\mathcal{X}}(x, y_j)}{\sqrt{J}} \right)_{j\in[\![1,J]\!]} \end{array} \right. . \tag{D.III.6}$$

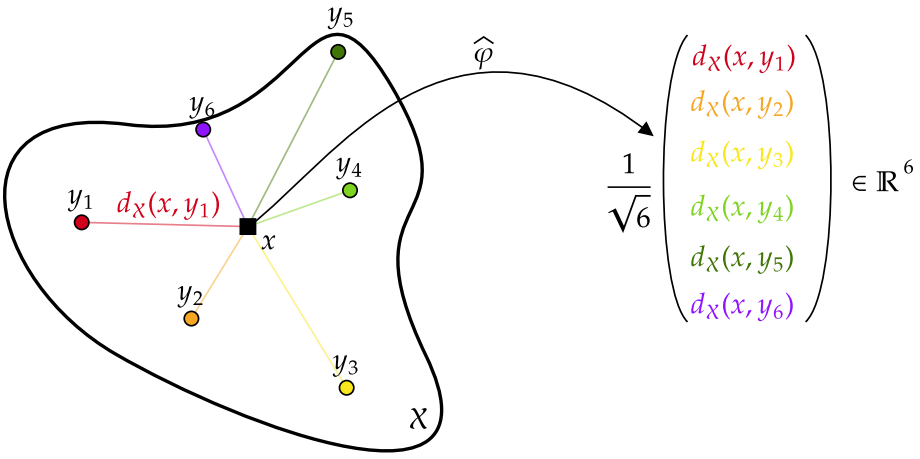


Figure D.III.3: The mapping $\hat{\varphi} : \mathcal{X} \longrightarrow \mathbb{R}^J$ maps a point $x \in \mathcal{X}$ to the vector of normalised distances between x and the centres y_j of the covering.

It is immediate to verify that $\hat{\varphi} : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (\mathbb{R}^J, \|\cdot\|_2)$ is 1-Lipschitz, thanks to the $J^{-1/2}$ normalisation. In Proposition D.III.3, we show how to embed $\mathbb{R}^J \supset \hat{\varphi}(\mathcal{X})$ into ℓ^2 with a mapping B, which will allow us to compare the RKHS induced by $\hat{\varphi}$ and a particular $\varphi : \mathcal{X} \longrightarrow \ell^2$. To construct B, we first begin with a geometric series separation lemma, which will be convenient to deal with the factor $\frac{1}{q^n}$ in φ.

Lemma D.III.2. Let $J \geq 2$, $q \in (1, 1 + \frac{1}{J-1})$ and coefficients $(\lambda_1, \cdots, \lambda_J) \in (0,1)^J$ such that $\sum_j \lambda_j = 1$, there exists $\alpha : \mathbb{N} \longrightarrow [\![1, J]\!]$ with for all $j \in [\![1, J]\!]$, $\alpha^{-1}(\{j\})$ infinite such that:

$$\forall j \in [\![1, J]\!], \quad \sum_{n \in \alpha^{-1}(\{j\})} \frac{1}{q^n} = \lambda_j \frac{q}{q-1}. \tag{D.III.7}$$

Proof. Set $S := \frac{q}{q-1}$. We will construct a sequence $(\alpha(N))_{N \in \mathbb{N}}$ by induction over N, verifying the property

$$P_N : \forall j \in [\![1, J]\!], \quad \sum_{n \in [\![0, N]\!] \,:\, \alpha(n) = j} \frac{1}{q^n} < \lambda_j S.$$

Initialisation: set $\alpha(0)$ the first $j \in [\![1, J]\!]$ such that $1 < \lambda_j S$. Note that such a j exists, otherwise summing over $j \in [\![1, J]\!]$ yields

$$J \geq \frac{q}{q-1} > \frac{1 + \frac{1}{J-1}}{1 + \frac{1}{J-1} - 1} = J,$$

which is a contradiction. We have defined $\alpha(0) := j$ verifying P_0.

Induction step: let $N \in \mathbb{N}$, suppose P_N true. We show that there exists $j \in [\![1, J]\!]$ such that

$$\sum_{n \in [\![0, N]\!] \,:\, \alpha(n) = j} \frac{1}{q^n} + \frac{1}{q^{N+1}} < \lambda_j S \tag{D.III.8}$$

by contradiction. If that were not the case, we would have by summing Eq. (D.III.8) over $j \in [\![1, J]\!]$:

$$\sum_{n=0}^{N} \frac{1}{q^n} + \frac{J}{q^{N+1}} \geq S,$$

which by computation is equivalent to $q \geq 1 + \frac{1}{J-1}$, obtaining a contradiction. Selecting $j \in [\![1, J]\!]$ such that Eq. (D.III.8) holds, we can set $\alpha(N+1) := j$ which satisfies P_{N+1}.

Now that $\alpha : \mathbb{N} \longrightarrow [\![1, J]\!]$ verifying (P_N) is constructed, we introduce the convergent series

$$\forall j \in [\![1, J]\!], \ \forall N \in \mathbb{N}, \ S_N^{(j)} := \sum_{n \in [\![0, N]\!] \,:\, \alpha(n) = j} \frac{1}{q^n}, \ S_\infty^{(j)} := \lim_{N \longrightarrow +\infty} S_N^{(j)}.$$

Thanks to (P_N), for all $j \in [\![1, J]\!]$ taking the limit yields $S_\infty^{(j)} \leq \lambda_j S$, and summing over $j \in [\![1, J]\!]$ gives $\sum_j S_\infty^{(j)} = S$, hence necessarily for all $j \in [\![1, J]\!]$, $S_\infty^{(j)} = \lambda_j S$.

Finally, observing the strict inequality in P_N at each $N \in \mathbb{N}$ shows that $\alpha^{-1}(\{j\})$ has to be infinite, concluding the proof. □

We now turn to constructing an embedding $B : \mathbb{R}^J \longrightarrow \ell^2$, which will allow us to compare $\hat{\varphi}$ and φ. An important property of B will be the correspondence between the inner products in $\mathbb{R}^J$ and ℓ^2 (i.e. B will be an isometry). The construction of this embedding revolves around the construction of an adapted basis $(x_n)_{n \in \mathbb{N}}$ of $\mathcal{X}$ which is balanced with respect to the covering by the balls $B(y_j, \eta)$, as illustrated in Fig. D.III.4.

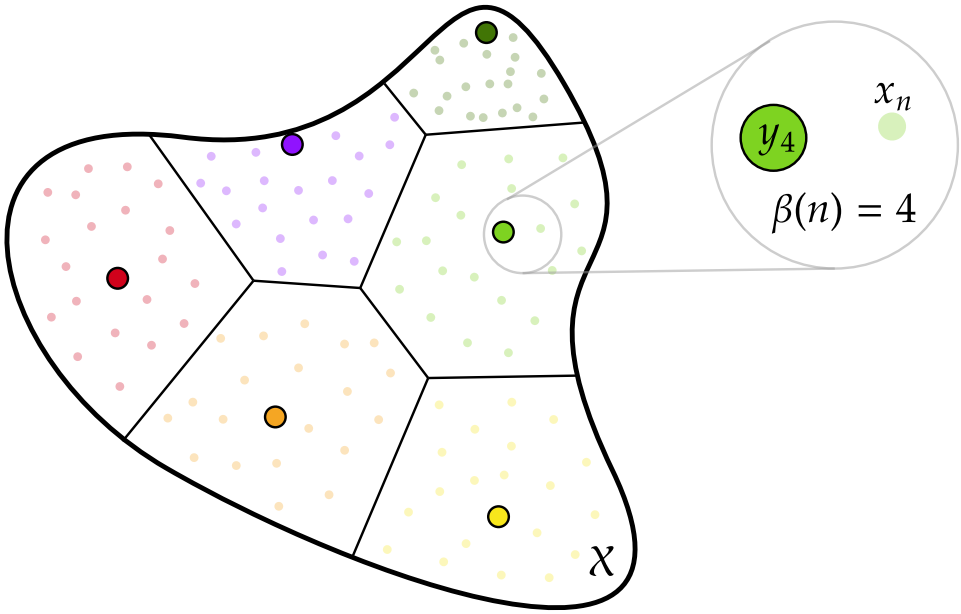


Figure D.III.4: The basis $(x_n)_{n\in\mathbb{N}}$ is such that there equally as many (x_n) in each region $\mathcal{X}_j$ of points closest to y_j. In the figure, we observe a zoom on the region $\mathcal{X}_4$, where the example point x_n is closest to y_4. In mathematical terms, we write this property as $\beta(n) = y_j$, and in Proposition D.III.2 we will construct (x_n) such that the sum $\sum_n q^{-2n}$ is split evenly between the sets $\beta^{-1}(\{j\})$.

Proposition D.III.2. Take $q \in (1, \sqrt{1 + \frac{1}{J-1}})$. There exists a basis $(x_n)_{n\in\mathbb{N}}$ of $\mathcal{X}$ and a mapping $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ with infinite pre-images which verifies $\forall n \in \mathbb{N},\ d_{\mathcal{X}}(x_n, y_{\beta(n)}) = \min_j d_X(x_n, y_j)$, and with the following property:

$$\forall j \in [\![1, J]\!],\ \sum_{n\in\beta^{-1}(\{j\})} \frac{1}{q^{2n}} = \frac{1}{J}\frac{q^2}{q^2-1}. \tag{D.III.9}$$

Proof. Consider for $j \in [\![1, J]\!]$ the set $\mathcal{X}_j := \{x \in \mathcal{X} : \operatorname{argmin}_m\ d_{\mathcal{X}}(x, y_m) = j\}$ (with disambiguation by taking the smallest minimiser if multiple exist). By definition, the sets $\mathcal{X}_j$ are disjoint and cover $\mathcal{X}$. Since $(\mathcal{X}, d_{\mathcal{X}})$ is a compact metric space, each subset $\mathcal{X}_j$ is separable, allowing us to choose a basis $(z_n^{(j)})_{n\in\mathbb{N}}$ of $\mathcal{X}_j$ for each $j \in [\![1, J]\!]$. By Lemma D.III.2, we can choose $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ with infinite pre-images which verifies Eq. (D.III.9). Since for each $j \in [\![1, J]\!]$, the set $\beta^{-1}(\{j\}) \subset \mathbb{N}$ is infinite, we can choose $\omega_j : \beta^{-1}(\{j\}) \longrightarrow \mathbb{N}$ a bijection. We can now define $\forall n \in \mathbb{N},\ x_n := z^{(\beta(n))}_{\omega_{\beta(n)}(n)}$, which is a basis of $\mathcal{X}$ since $\cup_j \mathcal{X}_j = \mathcal{X}$ and

$$\{x_n\}_{n\in\mathbb{N}} = \bigcup_j \{z^{(j)}_{\omega_j(m)}\}_{m\in\beta^{-1}(\{j\})} = \bigcup_j \{z_n^{(j)}\}_{n\in\mathbb{N}},$$

by construction. Furthermore, by definition, we have $\forall n \in \mathbb{N},\ \operatorname{argmin}_j\ d_{\mathcal{X}}(x_n, y_j) = \beta(n)$, which shows that the mapping β satisfies the desired properties. □

Using the adapted basis from Proposition D.III.2, we can finally construct an isometry $B : \mathbb{R}^J \longrightarrow \ell^2$:

Proposition D.III.3. Take a basis (x_n) of $\mathcal{X}$ and $\beta : \mathbb{N} \longrightarrow [\![1, J]\!]$ as in Proposition D.III.2. The mapping B defined below is an isometry:

$$B := \left\{ \begin{array}{ccc} \mathbb{R}^J & \longrightarrow & \ell^2 \\ (u_j)_{j=1}^J & \longmapsto & \left(c_B \frac{u_{\beta(n)}}{q^n}\right)_{n\in\mathbb{N}} \end{array} \right. , \ c_B := \frac{\sqrt{J(q^2-1)}}{q}. \tag{D.III.10}$$

Proof. The mapping B is clearly linear, and for $u, v \in \mathbb{R}^J$ we compute using Eq. (D.III.9):

$$\langle B(u), B(v)\rangle_{\ell^2} = \sum_{n=0}^{+\infty} \frac{c_B^2}{q^{2n}} u_{\beta(n)} v_{\beta(n)} = c_B^2 \sum_{j=1}^{J} u_j v_j \sum_{n \in \beta^{-1}(\{j\})} \frac{1}{q^{2n}} = c_B^2 \frac{1}{J} \frac{q^2}{q^2-1} \langle u, v\rangle_{\mathbb{R}^J} = \langle u, v\rangle_{\mathbb{R}^J},$$

which shows that B is an isometry. □

In the following, we draw a correspondence between a RKHS $\hat{H}$ built with $\hat{\varphi}$ from Eq. (D.III.6) and another RKHS H built using φ from Proposition D.III.1. Let $U := \hat{\varphi}(\mathcal{X})$, which is a compact subset of $\mathbb{R}^J$, then let $\hat{V} := B(U)$, it is a compact subset of ℓ^2. Consider the injection φ introduced in Proposition D.III.1 with basis (x_n) and scale q as in Proposition D.III.3. Define $V := \varphi(\mathcal{X})$, $W := V \cup \hat{V}$, which are also compact subsets of ℓ^2. We now summarise our objects in the following diagram:

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{\varphi} & V \subset W \subset \ell^2 \\ \Big\downarrow{\scriptstyle \hat{\varphi}} & & \\ U \subset \mathbb{R}^J & \xrightarrow[B]{} & \hat{V} \subset W \subset \ell^2 \end{array} \tag{D.III.11}$$

We fix a kernel $k_W : W^2 \longrightarrow \mathbb{R}$ which is of Taylor type or radial (see Definitions D.III.2 and D.III.3) and thus in particular universal on W, and introduce its canonical feature map:

$$\Phi_W := \left\{ \begin{array}{ccc} W & \longrightarrow & H_W \\ u & \longmapsto & k_W(\cdot, u) \end{array} \right. , \tag{D.III.12}$$

where $H_W = \overline{\mathrm{Span}}\,\{k_W(\cdot, u),\ u \in W\} \subset \mathcal{C}(W)$ is the unique RKHS associated to the kernel k_W ([CS08, Theorem 4.21]). Consider the kernels $k, \hat{k}$ on $\mathcal{X}$ defined respectively as:

$$k := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & k_W(\varphi(x), \varphi(y)) \end{array} \right. , \quad \hat{k} := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x,y) & \longmapsto & k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y)) \end{array} \right. . \tag{D.III.13}$$

By definition of the feature pair (H_W, Φ_W) for k_W, we observe that for $x, y \in \mathcal{X}$:

$$k(x,y) = \langle \Phi_W \circ \varphi(x), \Phi_W \circ \varphi(y)\rangle_{H_W},\ \hat{k}(x,y) = \langle \Phi_W \circ B \circ \hat{\varphi}(x), \Phi_W \circ B \circ \hat{\varphi}(y)\rangle_{H_W}. \tag{D.III.14}$$

The RKHS spaces $H, \hat{H}$ associated to $k, \hat{k}$ are both subspaces of $\mathcal{C}(\mathcal{X})$ and can be written with the following respective feature pairs $(H_W, \Phi), (H_W, \hat{\Phi})$ (use Eq. (D.III.14) with [CS08, Theorem 4.21]):

$$H = \{x \longmapsto \langle h_W, \Phi_W \circ \varphi(x)\rangle_{H_W},\ h_W \in H_W\},\ \Phi := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ \varphi(x) \end{array} \right. \tag{D.III.15}$$

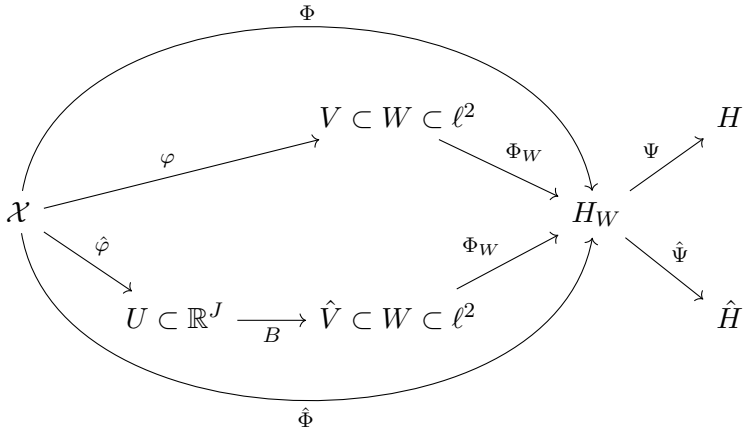
$$\hat{H} = \{x \longmapsto \langle h_W, \Phi_W \circ B \circ \hat{\varphi}(x)\rangle_{H_W},\ h_W \in H_W\},\ \hat{\Phi} := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ B \circ \hat{\varphi}(x) \end{array} \right. . \tag{D.III.16}$$

Notice that the feature space H_W is shared. To finish the diagram, we introduce the "feature-to-map" functionals:

$$\Psi := \left\{ \begin{array}{ccc} H_W & \longrightarrow & H \\ h_W & \longmapsto & x \mapsto \langle h_W, \Phi(x)\rangle_{H_W} \end{array} \right. , \ \hat{\Psi} := \left\{ \begin{array}{ccc} H_W & \longrightarrow & \hat{H} \\ h_W & \longmapsto & x \mapsto \langle h_W, \hat{\Phi}(x)\rangle_{H_W} \end{array} \right. . \tag{D.III.17}$$

By Eqs. (D.III.15) and (D.III.16), Ψ and $\hat{\Psi}$ are surjective. Extending the diagram in Eq. (D.III.11),

we obtain:



(D.III.18)

Using the inner product correspondence induced by the isometry B from Proposition D.III.3, a tractable formula for $\hat{k}$ is obtained immediately for Taylor and radial kernels.

Corollary D.III.1. The kernel $\hat{k}$ on $\mathcal{X}$ is given by, for all $x, y \in \mathcal{X}$:

- if k_W is a Taylor kernel (Definition D.III.2):

$$\hat{k}(x,y) = K(\langle \hat{\varphi}(x), \hat{\varphi}(y) \rangle_{\mathbb{R}^J}) = \sum_{n=0}^{+\infty} a_n \left(\frac{1}{J} \sum_{j=1}^{J} d_{\mathcal{X}}(x, y_j) d_{\mathcal{X}}(y, y_j) \right)^n ; \quad \text{(D.III.19)}$$

- if k_W is a radial kernel (Definition D.III.3):

$$\hat{k}(x,y) = K(\|\hat{\varphi}(x) - \hat{\varphi}(y)\|^2_{\mathbb{R}^J}) = \int_0^{+\infty} \exp\left(-\frac{s}{J} \sum_{j=1}^{J} (d_{\mathcal{X}}(x, y_j) - d_{\mathcal{X}}(y, y_j))^2 \right) \mathrm{d}\mu(s). \quad \text{(D.III.20)}$$

Proof. Let $x, y \in \mathcal{X}$, we remind that from Eq. (D.III.13) that $\hat{k}(x,y) := k_W(B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y))$. Now by Proposition D.III.3, B is an isometry, yielding:

$$\langle B \circ \hat{\varphi}(x), B \circ \hat{\varphi}(y) \rangle_{\ell^2} = \langle \hat{\varphi}(x), \hat{\varphi}(y) \rangle_{\mathbb{R}^J}; \quad \|B \circ \hat{\varphi}(x) - B \circ \hat{\varphi}(y)\|^2_{\ell^2} = \|\hat{\varphi}(x) - \hat{\varphi}(y)\|^2_{\mathbb{R}^J}.$$

Eqs. (D.III.19) and (D.III.20) are then obtained by replacing k_W and K by their definitions in the Taylor and radial cases. □

We refer to the expressions in Eqs. (D.III.19) and (D.III.20) as "tractable" since they can be computed explicitly on a computer or approximated efficiently to numerical precision (note that the measure μ in the radial kernel can be discrete with finite support). The numerical computation of $\hat{\varphi}$ is explicit and tractable in the sense that it can be done with a finite amount of closed-form expressions. For the Taylor kernel, the infinite series can be approximated to numerical precision, which we also refer to as "tractable", as would be said of the exponential function for instance.

D.III.3.2 Showing that $\hat{H}$ is Approximately Universal

In this section, we show that the RKHS $\hat{H}$ introduced in Section D.III.3.1 is approximately universal on $\mathcal{X}$. We use the notation and objects constructed in Section D.III.3.1 extensively, in particular, the mapping $\varphi : \mathcal{X} \longrightarrow \ell^2$ is defined using Proposition D.III.1 with a suitable basis (x_n) and scale q from Proposition D.III.2. The first approximation result we will show concerns a comparison in ℓ^2 between $\varphi(x)$ and $B \circ \hat{\varphi}(x)$:

Proposition D.III.4. For $x \in \mathcal{X}$, we have $\|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2} \leq \eta$.
The diameter of W verifies $D_W := \operatorname{diam}(W) \leq 2D_{\mathcal{X}}$.

Proof. Let $n \in \mathbb{N}$, we look at the terms of the sequences $\varphi(x), B \circ \hat{\varphi}(x) \in W \subset \ell^2$:

$$|[\varphi(x)]_n - [B \circ \hat{\varphi}(x)]_n| = \left| \frac{c_\varphi d_{\mathcal{X}}(x, x_n)}{q^n} - \frac{c_B d_{\mathcal{X}}(x, y_{\beta(n)})}{\sqrt{J} q^n} \right| \leq \frac{1}{q^n} c_\varphi d_{\mathcal{X}}(x_n, y_{\beta(n)}).$$

By construction of the covering $(B_{d_{\mathcal{X}}}(y_j, \eta))_j$ and of β (see Proposition D.III.2), $d_{\mathcal{X}}(x_n, y_{\beta(n)}) \leq \eta$. Summing the squares over $n \in \mathbb{N}$ and replacing $c_\varphi = \frac{c_B}{\sqrt{J}}$ yields:

$$\|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2}^2 \leq \eta^2 \sum_{n=0}^{+\infty} \frac{c_\varphi^2}{q^{2n}} = \eta^2.$$

For the diameter of $W := \varphi(\mathcal{X}) \cup (B \circ \hat{\varphi}(\mathcal{X}))$, we have by 1-Lipschitzness of $\varphi, \hat{\varphi}$ and B: $\operatorname{diam}(\varphi(\mathcal{X})) \leq D_{\mathcal{X}}$ and $\operatorname{diam}(B \circ \hat{\varphi}(\mathcal{X})) \leq D_{\mathcal{X}}$. Using the inequality in the above display and the fact that $\eta \leq D_{\mathcal{X}}$, we conclude:

$$D_W = \max(\operatorname{diam}(\varphi(\mathcal{X})), \sup_{x,y \in \mathcal{X}} \|\varphi(x) - B \circ \hat{\varphi}(y)\|_{\ell^2}) \leq \max(D_{\mathcal{X}}, 2\eta) \leq 2D_{\mathcal{X}}. \qquad \square$$

Using regularity properties of Taylor and radial kernels, we will show that the kernel $\hat{k}$ is approximately universal on $\mathcal{X}$ by relating it to k which is universal by Theorem D.III.1. First, we see in Lemma D.III.3 that the canonical feature map Φ_W is Hölder-continuous for Taylor kernels, and Lipschitz for radial kernels. We introduce the radius of W: $R_W := \max_{w \in W} \|w\|_{\ell^2}$. Using the definition of W and of $\varphi, \hat{\varphi}$ and B with their well-chosen normalisations, it is easy to see that $R_W \leq D_{\mathcal{X}}$.

Lemma D.III.3. The feature map $\Phi_W : (W, \|\cdot\|_{\ell^2}) \longrightarrow (H_W, \|\cdot\|_{H_W})$ has the following regularity:

- If k_W is a Taylor kernel, then Φ_W is $\frac{1}{2}$-Hölder continuous:

  $$\forall u, v \in W,\ \|\Phi_W(u) - \Phi_W(v)\|_{H_W} \leq \sqrt{2D_{\mathcal{X}} C_{K'}} \|u - v\|_{\ell^2}^{\frac{1}{2}},$$

  where $C_{K'} := \max_{t \in [-4D_{\mathcal{X}}^2, 4D_{\mathcal{X}}^2]} |K'(t)|$.

- If k_W is a radial kernel, then Φ_W is $\sqrt{2C_{K'}}$-Lipschitz:

  $$\forall u, v \in W,\ \|\Phi_W(u) - \Phi_W(v)\|_{H_W} \leq \sqrt{2C_{K'}} \|u - v\|_{\ell^2},$$

  where $C_{K'} := \max_{t \in [0, 4D_{\mathcal{X}}^2]} |K'(t)|$.

Proof. First, we remind that by Proposition D.III.4, we have $\operatorname{diam}(W) \leq 2D_{\mathcal{X}}$. For the proof, we take inspiration from [Fie23, Section 4.2]. Using the reproducing property, we begin computations for both kernel types, letting $u, v \in W$:

$$\begin{aligned} \|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 &= k_W(u,u) - 2k_W(u,v) + k_W(v,v) \\ &\leq |k_W(u,u) - k_W(u,v)| + |k_W(v,v) - k_W(u,v)|. \end{aligned}$$

For Taylor kernels, we use the fact that K is $C_{K'}$-Lipschitz on $[-4D_{\mathcal{X}}^2, 4D_{\mathcal{X}}^2]$ and the Cauchy-Schwarz inequality for $\langle \cdot, \cdot \rangle_{\ell^2}$:

$$\begin{aligned} \|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 &\leq C_{K'} \left(|\langle u, u \rangle_{\ell^2} - \langle u, v \rangle_{\ell^2}| + |\langle v, v \rangle_{\ell^2} - \langle u, v \rangle_{\ell^2}| \right) \\ &\leq C_{K'} (\|u\|_{\ell^2} + \|v\|_{\ell^2}) \|u - v\|_{\ell^2} \\ &\leq 2R_W C_{K'} \|u - v\|_{\ell^2}, \end{aligned}$$

and we conclude using $R_W \leq D_{\mathcal{X}}$. For radial kernels, we use the fact that K is $C_{K'}$-Lipschitz on $[0, D_W^2]$ (we remind that K is non-increasing on $[0, +\infty)$):

$$\|\Phi_W(u) - \Phi_W(v)\|_{H_W}^2 = 2(K(0) - K(\|u-v\|_{\ell^2}^2)) \leq 2C_{K'}\|u-v\|_{\ell^2}^2. \qquad \square$$

We now use Lemma D.III.3 to approximate any $h \in H$ with a $\hat{h} \in \hat{H}$ with a certain error, which we approach by comparing the feature-to-map functionals Ψ and $\hat{\Psi}$ from Eqs. (D.III.15) and (D.III.16).

Proposition D.III.5. For $h \in H$, take $h_W \in H_W$ such that $h = x \longmapsto \langle h_W, \Phi(x)\rangle_{H_W} = \Psi(h_W)$. Then let $\hat{h} := x \longmapsto \langle h_W, \hat{\Phi}(x)\rangle_{H_W} = \hat{\Psi}(h_W)$. Denoting $\|\cdot\|_\infty$ the supremum norm on $\mathcal{X}$, we have:

$$\|h - \hat{h}\|_\infty \leq \rho_0 \|h_W\|_{H_W}, \tag{D.III.21}$$

where $\rho_0 = \eta^{\frac{1}{2}}\sqrt{2D_{\mathcal{X}}C_{K'}}$ for a Taylor kernel and $\rho_0 = \eta\sqrt{2C_{K'}}$ for a radial kernel..

Proof. First, we use the regularity of Φ_W from Lemma D.III.3: we have for $x \in X$,

$$\begin{aligned}
|h(x) - \hat{h}(x)| &= \langle h_W, \Phi(x) - \hat{\Phi}(x)\rangle_{H_W} \leq \|h_W\|_{H_W}\|\Phi(x) - \hat{\Phi}(x)\|_{H_W} \\
&= \|h_W\|_{H_W}\|\Phi_W \circ \varphi(x) - \Phi_W \circ B \circ \hat{\varphi}(x)\|_{H_W} \\
&\leq \tilde{c}\|h_W\|_{H_W}\|\varphi(x) - B \circ \hat{\varphi}(x)\|_{\ell^2}^s,
\end{aligned}$$

where $(\tilde{c}, s) = (\sqrt{2D_{\mathcal{X}}C_{K'}}, \frac{1}{2})$ for a Taylor kernel and $(\tilde{c}, s) = (\sqrt{2C_{K'}}, 1)$ for a radial kernel. Combining with Proposition D.III.4, we obtain Eq. (D.III.21). $\square$

Using the universality of the kernel k (thanks to Theorem D.III.1), we can frame the result of Proposition D.III.5 as an approximate universality property of $\hat{k}$. Again, the approximation error functions depend on the type of kernel k_W.

Theorem D.III.2. Let $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, the element $h[\varepsilon, f] \in H$ defined by

$$h[\varepsilon, f] := \operatorname*{argmin}_{h \in H : \|h - f\|_\infty \leq \varepsilon} \|h\|_H^2 \tag{D.III.22}$$

is well-defined, and there exists $\hat{h} \in \hat{H}$ such that:

$$\|f - \hat{h}\|_\infty \leq \varepsilon + \rho_0 \|h[\varepsilon, f]\|_H, \tag{D.III.23}$$

where $\rho_0 = \eta^{\frac{1}{2}}\sqrt{2D_{\mathcal{X}}C_{K'}}$ for a Taylor kernel and $\rho_0 = \eta\sqrt{2C_{K'}}$ for a radial kernel..

Proof. First, we introduce:

$$h_W[\varepsilon, f] := \operatorname*{argmin}_{h_W \in H_W : \|\Psi(h_W) - f\|_\infty \leq \varepsilon} \|h_W\|_{H_W}^2$$

We show that $h_W[\varepsilon, f]$ and $h[\varepsilon, f]$ are well-defined. The triangle inequality ensures that the sets $\mathcal{B}_{H_W} := \{h_W \in H_W : \|\Psi(h_W) - f\|_\infty \leq \varepsilon\}$ and $\mathcal{B}_H := \{h \in H : \|h - f\|_\infty \leq \varepsilon\}$ are convex.

We now show the continuity of Ψ as a mapping $(H_W, \|\cdot\|_{H_W}) \longrightarrow (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$. Fixing $x \in \mathcal{X}$, we upper-bound by the Cauchy-Schwarz inequality:

$$|\Psi(h_W)[x]| = |\langle h_W, \Phi(x)\rangle_{H_W}| \leq \|h_W\|_{H_W}\|\Phi(x)\|_{H_W}, \tag{D.III.24}$$

then we use the definition $\Phi = \Phi_W \circ \varphi$ to show the continuity of $\Phi : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (H_W, \|\cdot\|_{H_W})$: by Proposition D.III.1, $\varphi : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (W, \|\cdot\|_{\ell^2})$ is continuous, and by Lemma D.III.3, $\Phi_W : (W, \|\cdot\|_{\ell^2}) \longrightarrow (H_W, \|\cdot\|_{H_W})$ is also continuous. Combining Eq. (D.III.24) with the continuity of Φ and the compactness of $\mathcal{X}$ ensures that there exists $C > 0$ independent of $x \in \mathcal{X}$ and $h_W \in H_W$ such that $|\Psi(h_W)[x]| \leq C\|h_W\|_{H_W}$, thus Ψ as a mapping $(H_W, \|\cdot\|_{H_W}) \longrightarrow (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ is continuous.

Thanks to the continuity of $\Psi : (H_W, \|\cdot\|_{H_W}) \longrightarrow (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$, we conclude that $\mathcal{B}_{H_W} := \{h_W \in H_W : \|\Psi(h_W) - f\|_\infty \leq \varepsilon\}$ is closed in $(H_W, \|\cdot\|_{H_W})$. Regarding $\mathcal{B}_H$, by [CS08, Lemma 4.23], since the kernel k is bounded on $\mathcal{X}$, the inclusion $\iota : (H, \|\cdot\|_H) \longrightarrow (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ is continuous. We deduce the closedness of $\mathcal{B}_H = \iota^{-1}\left(\mathcal{B}_{\mathcal{C}(\mathcal{X})}(f, \varepsilon)\right)$ in $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$.

Finally, the sets $\mathcal{B}_{H_W}$ and $\mathcal{B}_H$ are non-empty since $H = \Psi(H_W)$ is dense in $(\mathcal{C}(X), \|\cdot\|_\infty)$.

We conclude that $\mathcal{B}_{H_W}$, resp. $\mathcal{B}_H$ is a non-empty closed convex set in the Hilbert space $(H_W, \|\cdot\|_{H_W})$, resp. $(H, \|\cdot\|_H)$, and the Hilbert projection theorem (or directly [Rud87, Theorem 4.10]) ensures that $h_W[\varepsilon, f]$, resp. $h[\varepsilon, f]$ is uniquely defined.

Now, we show that $\|h[\varepsilon, f]\|_H = \|h_W[\varepsilon, f]\|_{H_W}$. By [CS08, Theorem 4.21], we have for all $h \in H$:

$$\|h\|_H = \inf\{\|h_W\|_{H_W},\ h = \Psi(h_W)\}.$$

By the same argument as before (using [Rud87, Theorem 4.10]), we show that the infimum is attained. The equality between norms is then straightforward by separating both inequalities and using $H = \Psi(H_W)$.

To obtain Eq. (D.III.23), we take $h := \Psi(h_W[\varepsilon, f])$ in Eq. (D.III.21) and $\hat{h} := \hat{\Psi}(h_W[\varepsilon, f]) \in \hat{H}$, and apply the triangle inequality for $\|\cdot\|_\infty$, using $\|h - f\|_\infty \leq \varepsilon$ and $\|h[\varepsilon, f]\|_H = \|h_W[\varepsilon, f]\|_{H_W}$. □

The approximation result in Eq. (D.III.23) shows that $\hat{k}$ is ρ-approximately universal (Definition D.III.4) for $\rho(f) := \rho_0\|h[\varepsilon, f]\|_H$. In the case where $\mathcal{X}$ is of dimension d (or has intrinsic dimension d), the number of covering balls scales as $J = \mathcal{O}(\eta^{-d})$, which does not impact the approximation rate, as is commonly the case in kernel methods which do not suffer from the curse of dimensionality (see for example [Gre+12, Section 4.1]). However, as is typically the case for discretisation methods, the rate $J = \mathcal{O}(\eta^{-d})$ is computationally prohibitive for small discretisation step η in high dimension d.

From a functional standpoint, a larger oscillation (a large value for $C_{K'} = \max_{t \in [-D_{\mathcal{X}}^2, D_{\mathcal{X}}^2]} |K'(t)|$ e.g. for the Taylor case), of the function K worsens the error, which could be understood as excessive locality or over-fitting. Finally, the error term $\rho(f)$ is relative in the sense that it depends on $\|h[\varepsilon, f]\|_H$, which is the smallest possible norm of an ε-approximation of f within H, and can be seen as a measure of complexity of f (in loose terms). This term depends on q, and while the exact dependence is unclear, we expect it to grow as q increases.

D.III.3.3 An Approximate Universal Truncated Kernel

In this section, we consider another approximate universal kernel which is obtained by truncation of φ. We will undergo a similar process as the construction of $\hat{H}$ in Section D.III.3.1 and follow closely the proof methods of Section D.III.3.2.

A natural idea is to simply consider a truncation of the mapping φ from Proposition D.III.1: fixing a basis $(x_n)_{n \in \mathbb{N}}$ of $\mathcal{X}$, a discretisation size $N \geq 2$ and scale $q > 1$, consider the mapping:

$$\varphi_t := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}^N \\ x & \longmapsto & \left(\dfrac{c_\varphi d_{\mathcal{X}}(x, x_n)}{q^j} \right)_{n \in [\![0, N-1]\!]} \end{array} \right. .$$

Straightforward computation shows that $\varphi_t : (\mathcal{X}, d_{\mathcal{X}}) \longrightarrow (\ell^2, \|\cdot\|_{\ell^2})$ is $\sqrt{1 - q^{-2N}}$-Lipschitz. We introduce the "padding" isometry:

$$B_t := \left\{ \begin{array}{ccc} \mathbb{R}^N & \longrightarrow & \ell^2 \\ (u_n)_{n=0}^{N-1} & \longmapsto & (u_0, \cdots, u_{N-1}, 0, \cdots) \end{array} \right. ,$$

Similarly to Section D.III.3.1, we take $V := \varphi(\mathcal{X})$, $U_t := \varphi_t(\mathcal{X})$ and $V_t := B_t(U_t)$, allowing us to introduce the compact set $W := V \cup V_t \subset \ell^2$ (we use the same notation as in Section D.III.3.1 to alleviate notation). Take k_W a Taylor or radial kernel on W, and introduce the kernel:

$$k_t := \left\{ \begin{array}{ccc} \mathcal{X}^2 & \longrightarrow & \mathbb{R} \\ (x, y) & \longmapsto & k_W(B_t \circ \varphi_t(x), B_t \circ \varphi_t(y)) \end{array} \right. .$$

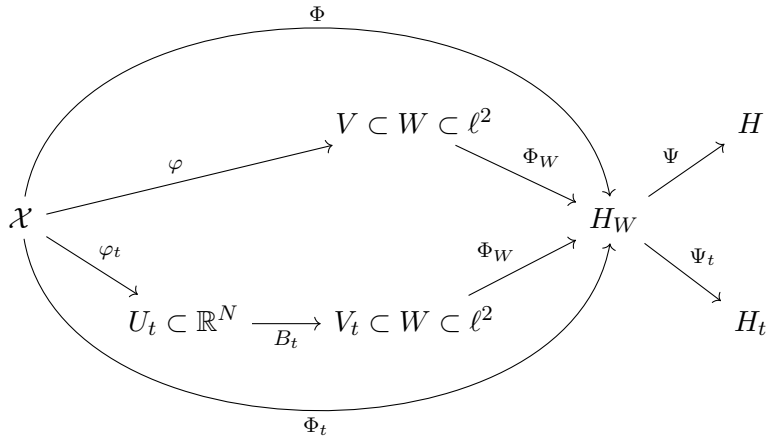
We continue with the feature pair (H_W, Φ_t) for the RKHS H_t associated to k_t, where:

$$\Phi_t := \left\{ \begin{array}{ccc} \mathcal{X} & \longrightarrow & H_W \\ x & \longmapsto & \Phi_W \circ B_t \circ \varphi_t(x) \end{array} \right. .$$

As in Eq. (D.III.17) we introduce the "feature-to-map" functionals:

$$\Psi := \left\{ \begin{array}{ccc} H_W & \longrightarrow & H \\ h_W & \longmapsto & x \mapsto \langle h_W, \Phi(x) \rangle_{H_W} \end{array} \right. , \ \Psi_t := \left\{ \begin{array}{ccc} H_W & \longrightarrow & H_t \\ h_W & \longmapsto & x \mapsto \langle h_W, \Phi_t(x) \rangle_{H_W} \end{array} \right. , \tag{D.III.25}$$

and finish the diagram:



The computation in the proof of Corollary D.III.1 stands, but the coefficients in the expression of φ_t lead to a different expression for k_t, which is a truncated version of k: if k_W is a Taylor kernel, we have:

$$k_t(x,y) = K\left(\langle \varphi_t(x), \varphi_t(y) \rangle_{\mathbb{R}^N}\right) = \sum_{n=0}^{+\infty} a_n \left(\sum_{m=0}^{N-1} \frac{c_\varphi^2 d_{\mathcal{X}}(x, x_m) d_{\mathcal{X}}(y, x_m)}{q^{2m}} \right)^n ,$$

and likewise for radial kernels:

$$k_t(x,y) = K\left(\|\varphi_t(x) - \varphi_t(y)\|^2_{\mathbb{R}^N}\right) = \int_0^{+\infty} \exp\left(-s \sum_{n=0}^{N-1} \frac{c_\varphi^2 (d_{\mathcal{X}}(x, x_n) - d_{\mathcal{X}}(y, x_n))^2}{q^{2n}} \right) \mathrm{d}\mu(s).$$

We now adapt Proposition D.III.4 to k_t:

Proposition D.III.6. For $x \in \mathcal{X}$, we have $\|\varphi(x) - B_t \circ \varphi_t(x)\|_{\ell^2} \leq \frac{D_{\mathcal{X}}}{q^N}$.
The diameter of W verifies $D_W \leq 2D_{\mathcal{X}}$.

Proof. For $n \in [\![0, N-1]\!]$, by construction $[\varphi(x)]_n = [B_t \circ \varphi_t(x)]_n$. For $n \geq N$, we have:

$$|[\varphi(x)]_n - [B_t \circ \varphi_t(x)]_n| = \frac{c_\varphi d_{\mathcal{X}}(x, y_n)}{q^n},$$

and by bounding the distance term by $D_{\mathcal{X}}$, and summing the squares, we obtain:

$$\|\varphi(x) - B_t \circ \varphi_t(x)\|^2_{\ell^2} \leq \sum_{n=N}^{+\infty} \frac{c_\varphi^2 D_{\mathcal{X}}^2}{q^{2n}} = \frac{c_\varphi^2 D_{\mathcal{X}}^2 q^2}{q^{2N}(q^2-1)} = \frac{D_{\mathcal{X}}^2}{q^{2N}}.$$

As for the result on D_W, it follows from 1-Lipschitzness as done in Proposition D.III.4. □

As in Section D.III.3.2, it is easy to verify that $R_W := \max_{w \in W} \|w\|_{\ell^2} \leq D_{\mathcal{X}}$. Following the same steps as in Theorem D.III.2, we show a similar result for k_t, replacing η with $D_{\mathcal{X}} q^{-N}$:

Theorem D.III.3. Let $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, there exists $h_t \in H_t$ such that:

$$\|f - h_t\|_\infty \leq \varepsilon + \rho_t \|h[\varepsilon, f]\|_H, \tag{D.III.26}$$

where $\rho_t = q^{-N/2} D_{\mathcal{X}} \sqrt{2C_{K'}}$ for a Taylor kernel and $\rho_t = q^{-N} D_{\mathcal{X}} \sqrt{2C_{K'}}$ for a radial kernel, with the constants $C_{K'}$ as in Lemma D.III.3.

Proof. We follow the same progression as in the proof of Theorem D.III.2. First, we follow the proof of Proposition D.III.5, applying Lemma D.III.3, then upper-bounding the term $\|\varphi(x) - B_t \circ \varphi_t(x)\|_{\ell^2}^2$ using Proposition D.III.6, and the only difference is the replacement of the term η with $D_{\mathcal{X}} q^{-N}$. Having adapted Proposition D.III.5, the proof of Theorem D.III.3 follows as in Theorem D.III.2. □

To compare with the rate from Theorem D.III.2, we see that the term η is replaced by $D_{\mathcal{X}} q^{-N}$. While q^{-N} becomes rapidly smaller as q increases, we suspect the term $\|h[\varepsilon, f]\|_H$ to grow quickly as q increases, which would favour the kernel $\hat{k}$ from Section D.III.3.1. From an intuitive standpoint, the quality of the truncation approximation depends on how well the truncated basis $(x_n)_{n=0}^{N-1}$ represents $\mathcal{X}$. For example, if $\mathcal{X}$ is a manifold of $\mathbb{R}^d$ with two connected components, and the first N elements are all in the first component, it can be expected that a substantial part of the information about the space is lost, hindering function approximation. This issue can arise because the "basis" property of (x_n) relates to the full sequence, whereas truncation focuses on the first N terms, which are not assumed to satisfy particular conditions. We saw in Sections D.III.3.1 and D.III.3.2 a more principled discretisation approach, which is in some sense a refinement of the truncation principle.

Acknowledgements

We thank Joan Glaunès for carefully proofreading this work and for his valuable insight.

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

Future Directions

In the attempt of answering numerous theoretical and algorithmic questions, we have inevitably opened just as many new ones. We provide here an overview of some possible future directions emanating from the works presented in this thesis. First, we go through natural extensions and technical refinements which are incremental in nature. We then present more ambitious and open-ended research directions.

D.III.4 Extensions and Refinements

Importance of the number of directions in approximations of the Sliced Wasserstein distance. From an optimisation standpoint, the importance of p in the approximation of $\mathcal{E}$ by $\mathcal{E}_p$ remains largely unclear. Guarantees on the quality of local optima of $\mathcal{E}_p$ for $p \gg d$ or on the probability of converging to an unwanted local optimum are still open questions.

Refinements of the analysis of the flow of the Sliced Wasserstein distance. The convergence of SGD for the discrete Sliced Wasserstein distance was established in a relatively weak sense, whereas convergence to the target measure is observed numerically. A proof of convergence or a counter-example thereof seem to be missing, as well as a formal connection to the Wasserstein flow of the Sliced Wasserstein distance, as studied very recently in [CS25; VMK25]. As a possible direction, descent properties such as those established in [VMK25] for $X \longmapsto \mathrm{SW}_2^2(\gamma_X, \rho)$ for an absolutely continuous measure ρ could be studied in the discrete case. Another avenue of research would be to adapt the results on W_2-flows of SW_2^2 for absolutely continuous measures established by [MC23; CS25] to the discrete case. In the light of the importance of the number of directions p in the approximation of $\mathcal{E}$ by $\mathcal{E}_p$, it may be possible to lower-bound the energy of the flow of SW_2^2 at each step $t \in \mathbb{N}$ with respect to t.

Uniqueness in the constrained optimal transport map problem. A natural question when studying $\min_{g \in G} \mathcal{T}_c(g\#\mu, \nu)$ is uniqueness of minimisers. We have seen in Chapter B.I that uniqueness does not hold in many cases, and could only establish it in the very restricted case where G is the set of Lipschitz gradients of convex functions, where $\mu = \nu$ with μ absolutely continuous, and with the square-Euclidean cost. An interesting theoretical question is the study of the relaxation of these conditions for uniqueness, beginning with the case of $\mu = \nu$.

Quantifying the quality of sliced optimal transport plans. In Chapter B.II, we have introduced different notions of plans arising from variants of the Sliced Wasserstein distance. Theoretical guarantees describing the quality of these plans (such as their cost in the Kantorovich problem) with respect to the number of directions are missing and of substantial practical importance. In particular, this would provide a theoretical justification (or otherwise) for the use of these plans in applications such as our sliced Gromov Wasserstein heuristic in Chapter B.III.

Further theoretical study of the Sliced Wasserstein variants. The Pivot-Sliced Wasserstein and Expected Sliced Wasserstein discrepancies studied in Chapter B.II would warrant further theoretical study. It remains unclear whether the min-Pivot-Sliced discrepancy $\min \mathrm{PS}$ verifies the triangle inequality. For the Expected Sliced Wasserstein, due to the failure of the property $\mathrm{ES}_\sigma(\mu, \mu) = 0$, it is natural to study the following "corrected" discrepancy (as in [GPC18]):

$$\overline{\mathrm{ES}}_\sigma(\mu, \nu) := \mathrm{ES}_\sigma(\mu, \nu) - \tfrac{1}{2}\left(\mathrm{ES}_\sigma(\mu, \mu) + \mathrm{ES}_\sigma(\nu, \nu)\right).$$

For $\min \mathrm{PS}$ and $\overline{\mathrm{ES}}_\sigma$, a natural metric question is the equivalence to the 2-Wasserstein distance, as was studied for the Sliced Wasserstein distance in [Bon13; BG21]. Further theoretical ques-

tions about the induced spaces of measures are also of interest, in the line of the considerations of [KT24; PS25].

Refinements of the Sliced Gromov Wasserstein heuristic. Further numerical investigation of the Sliced Gromov Wasserstein heuristic presented in Chapter B.III would be warranted, to validate or disprove its practical computational utility.

Barycentres of degenerate Gaussian Mixture Models. The Mixture Wasserstein distance MW_2 [DGS21] is well-defined between mixtures of *degenerate* Gaussian measures, however the algorithm proposed in Chapter C.II for the computation of barycentres for MW_2 does not apply in this case. This difficulty is due to the fact that barycentres of degenerate Gaussian measures are not covered by [AC11] and not studied in the literature (to our knowledge).

Convergence (speed) of the fixed-point algorithm for OT barycentres. in numerical experiments, the fixed-point algorithm proposed in Chapter C.II converges in very few iterations. A theoretical justification of this observation would be of great interest. Numerous theoretical questions about the problem itself remain open, such as conditions ensuring uniqueness of barycentres, or guaranteeing that fixed points of G are indeed barycentres.

Lifting assumptions on the ground barycentre function for OT barycentres. In Chapter C.II, we have assumed that the ground barycentre B is uniquely defined, which is a relatively strong assumption on the costs c_k. One could adapt our fixed-point algorithm for a multi-valued B, likely using measureable selection (see [Bog07, Section 6.9]). In any case, a characterisation of the costs c_k for which the barycentre B is unique would be useful.

D.III.5 Broad Research Directions

D.III.5.1 A (formally) Riemannian framework for OT barycentres

We provided an informal Riemannian interpretation of the fixed-point algorithm for OT barycentres in Chapter C.II, which was not required for our theoretical analysis. A more formal Riemannian framework (perhaps in the sense of [AGS05; MDC20] and with strong assumptions on the costs and measures) could shed new light on the convergence of the algorithm. As was done for barycentres of Gaussians in [Alt+21], the fixed-point algorithm could be seen as a (Wasserstein) gradient flow of the barycentre energy. For the case of barycentres for the 2-Wasserstein distance on a manifold $(\mathcal{X}, d_{\mathcal{X}})$, the main idea is to see the fixed-point algorithm as (Wasserstein) gradient descent of the barycentre energy $V(\mu) := \sum_k \lambda_k \mathrm{W}_2^2(\mu, \nu_k)$. To see this, we would work by analogy with the Fréchet barycentre algorithm on a (true) Riemannian manifold (M, d) (see [ATV13] for a complete presentation and convergence guarantees). We present the algorithm without discussing technicalities about the definition of the Riemannian exponential (Exp) and logarithmic (Log) maps, referring to [ATV13] for details. The idea is to differentiate the barycentric energy:

$$V(x) := \sum_k \lambda_k d^2(x, y_k), \quad \nabla V(x) = -2 \sum_k \lambda_k \mathrm{Log}_x(y_k),$$

and perform (Riemannian) gradient descent:

$$\forall t \in \mathbb{N};\ x_{t+1} = \mathrm{Exp}_{x_t}\left(-\eta \nabla V(x_t)\right),$$

where $\eta > 0$ is a step size, which in Chapter C.II is chosen as $\eta := \frac{1}{2}$. We illustrate this algorithm visually in Fig. D.III.5. For additional context and formulations in more generality, see also [Pen06; GJS23; MAM25]. To extend these concepts to OT barycentres, a first step would be to see W_2 barycentres of absolutely continuous measures of $\mathcal{P}_2(\mathbb{R}^d)$ through these lenses, as done in [ZP19] (An additional possible reference for inspiration is [CNR25, Exercice

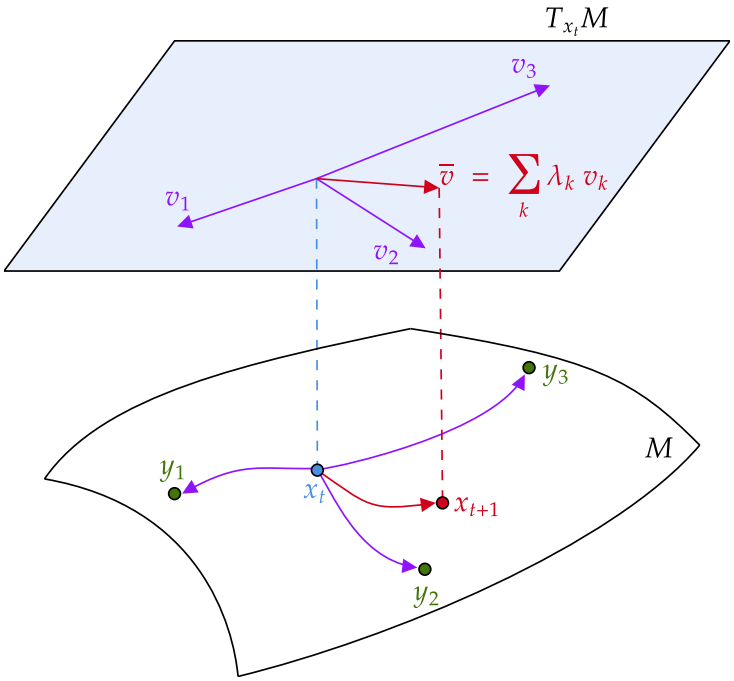


Figure D.III.5: Illustration of the Fréchet barycentre algorithm on a Riemannian manifold. The points y_k are the data, and the point x_t is the current iterate. The Riemannian gradient $\nabla V(x_t)$ is $-2\sum_k \lambda_k v_k$ where $v_k := \mathrm{Log}_{x_t}(y_k)$ and the next iterate is obtained by performing a step in the direction of the gradient by applying the exponential map at x_t.

5, Section 8.5]). More challenging directions to consider include the extension to non-absolute continuous measures, the replacement of $(\mathbb{R}^d, \|\cdot\|_2)$ with a Riemannian manifold $(\mathcal{X}, d_{\mathcal{X}})$, and finally considering more general costs than $d^2_{\mathcal{X}}$.

Given the linearisation interpretation, it also appears natural to draw links to the nu-based Wasserstein distance [NP23]. Finally, insight on the surrogate functional H may be obtained using the properties of the barycentric projection presented in [AGS05, Theorem 12.4.4].

D.III.5.2 RKHS Representations of OT Maps

In Chapter D.II, we presented a negative answer to a specific idea of representing gradients of convex functions in an RKHS cone. Further study of alternative methods, perhaps inspired by the representation of Kantorovich potentials proposed in [Vac+21; Muz+21; Vac+24], could be considered to represent OT maps in RKHS cones. Broadly speaking, the general idea of [Vac+21] is to write the dual of the Kantorovich problem (for the square-Euclidean cost) as:

$$\max_{\substack{u\in H_{\mathcal{X}},\, v\in H_{\mathcal{Y}} \\ A\in S^+(H_{\mathcal{XY}})}} \langle u, M(\mu)\rangle_{H_{\mathcal{X}}} + \langle v, M(\nu)\rangle_{H_{\mathcal{Y}}} \qquad \text{(D.III.27)}$$
$$\text{s.t. } \forall (x,y)\in \mathcal{X}\times\mathcal{Y},\ \|x-y\|_2^2 - u(x) - v(y) = \langle \phi(x,y), A\phi(x,y)\rangle_{H_{\mathcal{XY}}},$$

where μ and ν are respectively supported on the convex, bounded and open sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, and μ, ν admit densities p_μ, p_ν that verify $p_\mu, p_\nu \geq \delta > 0$ on $\mathcal{X}, \mathcal{Y}$ that are sufficiently smooth. The spaces $H_{\mathcal{X}}$ and $H_{\mathcal{Y}}$ are RKHS spaces of Sobolev functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{Y}$ to $\mathbb{R}$ respectively, and $H_{\mathcal{XY}}$ is an RKHS space of Sobolev functions from $\mathcal{X}\times\mathcal{Y}$ to $\mathbb{R}$ and of canonical feature map $\phi : \mathcal{X}\times\mathcal{Y} \longrightarrow H_{\mathcal{XY}}$. The notation $S^+(H_{\mathcal{XY}})$ denotes the cone of positive definite operators on $H_{\mathcal{XY}}$. Finally, $M(\mu)$ and $M(\nu)$ are the kernel mean embeddings of μ and ν in the RKHSs $H_{\mathcal{X}}$ and $H_{\mathcal{Y}}$ respectively.

The problem in Eq. (D.III.27) admits discrete approximations with statistical guarantees, yet the obtained approximate maps are not gradients of convex functions (to our understanding). Another natural direction is the study of the discretisation of the problem without assumptions on the measures, which can be seen as another RKHS variant of the constrained approximate map problem from Chapter B.I.

D.III.5.3 Flows in the GMM Space and Mixture-Wasserstein Geometry

A natural extension of [DD20] and of the results presented in Chapter A.IV would be to study the geometry of the space GMM_d of Gaussian Mixture Models (GMMs) on $\mathbb{R}^d$, equipped with the Mixture Wasserstein distance MW_2. A first possible viewpoint is the one chosen in [CGT18; Lam+22], which consists in seeing GMM_d as probability measures on the space of Gaussian measures $\mathcal{G}_d$, which is to say as a subset of $\mathcal{P}(\mathcal{G}_d)$. The "ground" space $\mathcal{G}_d$ is equipped with the Bures-Wasserstein distance d_{BW} between Gaussian measures (seen as points), and $\mathcal{P}(\mathcal{G}_d)$ with the 2-Wasserstein distance W_2, i.e. the OT cost with ground cost d^2_{BW}. As presented in [Lam+22], this provides a formal Riemannian structure to GMM_d in the sense of Otto-Villani [Ott01; Vil09], and we notice that the 2-Wasserstein distance between finite mixtures (i.e. elements $\mathcal{P}(\mathcal{G}_d)$ that are discrete measures) corresponds to the Mixture Wasserstein distance MW_2. This viewpoint has the drawback that the "infinite-mixture" representation is ill-defined: for example the mesure $\mu := \mathcal{N}(0,1) \in \mathcal{P}_2(\mathbb{R})$ can be represented in $\mathcal{P}(\mathcal{G}_1)$ naturally as $\delta_{\mathcal{N}(0,1)}$, but also by the infinite mixture of degenerate Gaussians:

$$\int_{\mathbb{R}} \varphi(x)\delta_{\mathcal{N}(x,0)}\mathrm{d}x,$$

where $\mathcal{N}(x,0) = \delta_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and φ is the density of $\mathcal{N}(0,1)$ (see [CNR25], Section 5.6). For both practical and theoretical reasons, it appears natural to restrict the space to *finite* mixtures of at most K Gaussians, as proposed in [PLK25]. We would be interested in generalising their study of the flow of the KL divergence for isotropic Gaussians to flows of generic functions $\mathcal{F}$ on GMM_d. Following the construction proposed in [PLK25], we would consider the following modified JKO scheme [JKO98] with respect to the GMM parameters $\theta \in \Theta_d(K) := \triangle_K \times (\mathbb{R}^d)^K \times S_d^{++}(\mathbb{R})^K$:

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta\in\Theta_d(K)} \mathcal{F}(\mu_\theta) + \tfrac{1}{2K\eta}\mathrm{MW}_2^2(\mu_\theta, \mu_{\theta_t}),$$

where μ_θ is the element of $\mathrm{GMM}_d(K)$ associated to $\theta \in \Theta_d(K)$ and $\eta > 0$ is a step size. Furthermore, as in [PLK25], we can consider the following mirror descent linearisation:

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta\in\Theta_d(K)} \langle \nabla_\theta \mathcal{F}(\mu_\theta), \theta - \theta_t\rangle + \tfrac{1}{2K\eta}\mathrm{MW}_2^2(\mu_\theta, \mu_{\theta_t}),$$

(where ∇_θ denotes the gradient w.r.t. θ for the MW_2 geometry), which is more amenable to numerical implementation. For $\eta \longrightarrow 0^+$, the continuous-time limit of these schemes would need to be compared to the continuous-time GMM flow defined in [CNR25, Theorem 5.23]. Given the discussion on fixing weights in Chapter A.IV and the considerations in [CNR25, Section 5.8], it would also be relevant to study alternative geometries which are more adapted to varying the weights of the mixture, such as the Wasserstein Fisher-Rao geometry (introduced by [LMS18], see also [CNR25, Section 5.7]) as done in [Lam+22] for the KL divergence.

Bibliography

[AM24] Anshul Adve and Alpár Mészáros. *On nonexpansiveness of metric projection operators on Wasserstein spaces*. 2024. arXiv: 2009.01370 [math.FA] (cit. on p. 176).

[ATV13] Bijan Afsari, Roberto Tron, and René Vidal. "On the convergence of gradient descent for finding the Riemannian center of mass". In: *SIAM Journal on Control and Optimization* 51.3 (2013), pp. 2230–2260 (cit. on p. 354).

[AC11] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein Space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924 (cit. on pp. 21, 37, 138, 264, 266, 273, 284, 305, 354).

[AC17] Martial Agueh and Guillaume Carlier. "Vers un théorème de la limite centrale dans l'espace de Wasserstein?" In: *Comptes Rendus. Mathématique* 355.7 (2017), pp. 812–818 (cit. on p. 284).

[Aiz64] A Aizerman. "Theoretical foundations of the potential function method in pattern recognition learning". In: *Automation and remote control* 25 (1964), pp. 821–837 (cit. on p. 336).

[AGD19] Hana Alghamdi, Mairead Grogan, and Rozenn Dahyot. "Patch-based colour transfer with optimal transport". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE. 2019, pp. 1–5 (cit. on p. 55).

[AB94] Charalambos D. Aliprantis and Kim C. Border. "Correspondences". In: *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 458–520 (cit. on pp. 289, 291, 292).

[Alt+21] Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. "Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22132–22145 (cit. on pp. 303, 354).

[AB22] Jason M. Altschuler and Enric Boix-Adserà. "Wasserstein Barycenters Are NP-Hard to Compute". In: *SIAM Journal on Mathematics of Data Science* 4.1 (2022), pp. 179–203 (cit. on pp. 21, 37, 138, 284).

[Álv+16] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. "A fixed-point approach to barycenters in Wasserstein space". In: *Journal of Mathematical Analysis and Applications* 441.2 (2016), pp. 744–762 (cit. on pp. 21, 23, 24, 38–40, 138, 283, 285–288, 293, 295, 302, 303, 306).

[AJJ19] David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. "Towards optimal transport with global invariances". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1870–1879 (cit. on p. 164).

[ABS+21] Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*. Vol. 130. Springer, 2021 (cit. on pp. 1, 170).

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005 (cit. on pp. 1–3, 176, 177, 181, 206, 207, 210, 219, 220, 248, 249, 298, 354, 355).

[ABM16] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. "Discrete Wasserstein barycenters: Optimal transport for discrete data". In: *Mathematical Methods of Operations Research* 84 (2016), pp. 389–409 (cit. on pp. 267, 301, 311).

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223 (cit. on pp. 1, 15, 16, 31, 32, 54, 102, 110, 141, 164, 167, 204).

[Aro50] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404 (cit. on pp. 197, 323, 337).

[AG18] Dario Azzimonti and David Ginsbourger. "Estimating orthant probabilities of high-dimensional Gaussian vectors with an application to set estimation". In: *J. Comput. Graph. Statist.* 27.2 (2018), pp. 255–267 (cit. on p. 88).

[Bac+25] Julio Backhoff, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. "Stochastic gradient descent for barycenters in Wasserstein space". In: *Journal of Applied Probability* 62.1 (2025), pp. 15–43 (cit. on pp. 70, 71).

[Bac+22] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. "Bayesian learning with Wasserstein barycenters". In: *ESAIM: Probability and Statistics* 26 (2022), pp. 436–472 (cit. on p. 284).

[BKK19] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "Deep equilibrium models". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 126).

[BE94] TL Bailey and C Elkan. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology*. Vol. 2. 1994, pp. 28–36 (cit. on p. 126).

[BPV24a] Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. "Understanding the training of infinitely deep and wide resnets with conditional optimal transport". In: *Communications on Pure and Applied Mathematics* (2024) (cit. on p. 2).

[BBR06] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. "On minimum Kantorovich distance estimators". In: *Statistics & probability letters* 76.12 (2006), pp. 1298–1302 (cit. on pp. 1, 16, 32).

[BG21] Erhan Bayraktar and Gaoyue Guo. "Strong equivalence between metrics of Wasserstein type". In: *Electronic Communications in Probability* 26 (Jan. 2021) (cit. on pp. 55, 103, 353).

[Bec94] Thomas Beck. "Automatic differentiation of iterative processes". In: *Journal of Computational and Applied Mathematics* 50.1-3 (1994), pp. 109–118 (cit. on pp. 126, 144).

[BB25] Florian Beier and Robert Beinert. "Tangential fixpoint iterations for Gromov–Wasserstein barycenters". In: *SIAM Journal on Imaging Sciences* 18.2 (2025), pp. 1058–1100 (cit. on pp. 285, 300, 302).

[BBS23] Florian Beier, Robert Beinert, and Gabriele Steidl. *Multi-Marginal Gromov-Wasserstein Transport and Barycenters*. 2023. arXiv: 2205.06725 [math.OC] (cit. on p. 285).

[BHS23a] Robert Beinert, Cosmas Heiss, and Gabriele Steidl. "On assignment problems related to Gromov–Wasserstein distances on the real line". In: *SIAM Journal on Imaging Sciences* 16.2 (2023), pp. 1028–1032 (cit. on pp. 252, 258).

[Ben+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138 (cit. on p. 285).

[BT11] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011 (cit. on p. 336).

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Vol. 6. Athena scientific Belmont, MA, 1997 (cit. on pp. 3, 232).

[BM92] Paul J Besl and Neil D McKay. "Method for registration of 3-D shapes". In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. Spie. 1992, pp. 586–606 (cit. on p. 244).

[Bha13] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013 (cit. on p. 155).

[BJL19] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". In: *Expositiones mathematicae* 37.2 (2019), pp. 165–191 (cit. on pp. 103, 306).

[BHS22] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. "Convergence of constant step stochastic gradient descent for non-smooth non-convex functions". In: *Set-Valued and Variational Analysis* 30.3 (2022), pp. 1117–1147 (cit. on pp. 8, 12, 28, 56, 70–75, 78, 87, 101, 103, 105, 106, 108, 110–112).

[BHS23b] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. "Stochastic subgradient descent escapes active strict saddles on weakly convex functions". In: *Mathematics of Operations Research* (2023) (cit. on pp. 8, 78).

[BCP19] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. "Penalization of barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 51.3 (2019), pp. 2261–2285 (cit. on p. 284).

[BBN24] Xin Bing, Florentina Bunea, and Jonathan Niles-Weed. *Estimation and inference for the Wasserstein distance between mixing measures in topic models*. 2024. arXiv: 2206.12768 [math.ST] (cit. on p. 91).

[Bir46] Garrett Birkhoff. "Three observations on linear algebra". In: *Univ. Nac. Tacuman, Rev. Ser. A* 5 (1946), pp. 147–151 (cit. on pp. 3, 20, 36, 206, 225, 226, 229).

[BYF20] Emily Black, Samuel Yeom, and Matt Fredrikson. "Fliptest: fairness testing via optimal transport". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 111–121 (cit. on pp. 2, 165).

[Blo+22] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. "Efficient and modular implicit differentiation". In: *Advances in neural information processing systems* 35 (2022), pp. 5230–5242 (cit. on p. 126).

[Bog07] Vladimir I Bogachev. *Measure theory*. Springer, 2007 (cit. on p. 354).

[Boï+25] Samuel Boïté, Eloi Tanguy, Julie Delon, Agnès Desolneux, and Rémi Flamary. *Differentiable Expectation-Maximisation and Applications to Gaussian Mixture Model Optimal Transport*. 2025. arXiv: 2509.02109 [cs.LG] (cit. on pp. 25, 43, 126).

[BDL07] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems". In: *SIAM Journal on Optimization* 17.4 (2007), pp. 1205–1223 (cit. on p. 8).

[BDL09] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. "Tame functions are semismooth". In: *Mathematical Programming* 117.1 (2009), pp. 5–19 (cit. on pp. 8, 185).

[Bol+07] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. "Clarke subgradients of stratifiable functions". In: *SIAM Journal on Optimization* 18.2 (2007), pp. 556–572 (cit. on pp. 8, 70).

[BLP23] Jérôme Bolte, Tam Le, and Edouard Pauwels. "Subgradient sampling for nonsmooth nonconvex minimization". In: *SIAM Journal on Optimization* 33.4 (2023), pp. 2542–2569 (cit. on pp. 8, 18, 34, 112–114, 191, 192).

[Bol+21] Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. "Nonsmooth implicit differentiation for machine-learning and optimization". In: *Advances in neural information processing systems* 34 (2021), pp. 13537–13549 (cit. on p. 126).

[BP21] Jérôme Bolte and Edouard Pauwels. "Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning". In: *Mathematical Programming* 188 (2021), pp. 19–51 (cit. on pp. 8, 78, 79, 103, 110, 111, 117–120, 184, 185, 192, 269).

[BPV22] Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. "Automatic differentiation of nonsmooth iterative algorithms". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26404–26417 (cit. on p. 126).

[BPV24b] Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. "One-step differentiation of iterative algorithms". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 126, 130, 131).

[Bon+22] Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. "Efficient gradient flows in sliced-Wasserstein space". In: *Transactions on Machine Learning Research* (2022) (cit. on p. 114).

[BDC25] Clément Bonet, Lucas Drumetz, and Nicolas Courty. "Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds". In: *Journal of Machine Learning Research* 26.32 (2025), pp. 1–76 (cit. on p. 205).

[Bon+24] Clément Bonet, Kimia Nadjahi, Thibault Séjourné, Kilian Fatras, and Nicolas Courty. "Slicing Unbalanced Optimal Transport". In: *Transactions on Machine Learning Research* (2024) (cit. on p. 140).

[Bon+11] N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich. "Displacement interpolation using Lagrangian mass transport". In: *Proceedings of SIGGRAPH'Asia*. 2011, pp. 1–12 (cit. on p. 126).

[BD23] Nicolas Bonneel and Julie Digne. "A survey of optimal transport for computer graphics and computer vision". In: *Computer Graphics Forum*. Vol. 42. 2. Wiley Online Library. 2023, pp. 439–460 (cit. on pp. 204, 244).

[BPC16] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. "Wasserstein barycentric coordinates: histogram regression using optimal transport." In: *ACM Trans. Graph.* 35.4 (2016), pp. 71–1 (cit. on p. 284).

[Bon+15a] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. "Sliced and Radon Wasserstein barycenters of measures". In: *Journal of Mathematical Imaging and Vision* 51.1 (2015), pp. 22–45 (cit. on pp. 12, 27, 46, 55, 56, 59, 60, 64, 66, 67, 72, 122, 204, 205, 241).

[Bon+15b] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. "Blind video temporal consistency". In: *ACM Transactions on Graphics (TOG)* 34.6 (2015), pp. 1–9 (cit. on p. 55).

[Bon13] N. Bonnotte. "Unidimensional and Evolution Methods for Optimal Transportation". PhD thesis. Paris 11, 2013 (cit. on pp. 6, 55, 65, 103, 126, 205, 353).

[Bou+17] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. *From optimal transport to generative modeling: the VEGAN cookbook*. 2017. arXiv: 1705.07642 [stat.ML] (cit. on p. 1).

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 149).

[Boy83] Russell A Boyles. "On the convergence of the EM algorithm". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 45.1 (1983), pp. 47–50 (cit. on p. 126).

[Bré+23] Pierre Bréchet, Katerina Papagiannouli, Jing An, and Guido Montúfar. "Critical points and convergence analysis of generative deep linear networks trained with Bures-Wasserstein loss". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 3106–3147 (cit. on p. 103).

[Bre91] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4 (1991), pp. 375–417 (cit. on pp. 4, 165, 175, 177, 330).

[Bre18] Yann Brenier. "The initial value problem for the Euler equations of incompressible fluids viewed as a concave maximization problem". In: *Communications in Mathematical Physics* 364 (2018), pp. 579–605 (cit. on p. 2).

[BFR24] Camilla Brizzi, Gero Friesecke, and Tobias Ried. *h-Wasserstein barycenters*. 2024. arXiv: 2402.13176 [math.AP] (cit. on pp. 21, 37, 284).

[BFR25] Camilla Brizzi, Gero Friesecke, and Tobias Ried. "p-Wasserstein barycenters". In: *Nonlinear Analysis* 251 (2025), p. 113687 (cit. on pp. 21, 37, 284).

[Bun+24] Charlotte Bunne, Geoffrey Schiebinger, Andreas Krause, Aviv Regev, and Marco Cuturi. "Optimal transport for single-cell and spatial omics". In: *Nature Reviews Methods Primers* 4.1 (2024), p. 58 (cit. on pp. 2, 204).

[Bur69] Donald Bures. "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras". In: *Transactions of the American Mathematical Society* 135 (1969), pp. 199–212 (cit. on p. 103).

[BDG12] Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. "Optimal-transport formulation of electronic density-functional theory". In: *Physical Review A—Atomic, Molecular, and Optical Physics* 85.6 (2012), p. 062502 (cit. on p. 204).

[Caf00] Luis A Caffarelli. "Monotonicity properties of optimal transportation and the fkg and related inequalities". In: *Communications in Mathematical Physics* 214 (2000), pp. 547–563 (cit. on p. 165).

[CJJ05] Brian S Caffo, Wolfgang Jank, and Galin L Jones. "Ascent-based Monte Carlo expectation–maximization". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 235–251 (cit. on p. 126).

[Cap11] Olivier Cappé. "Online expectation maximisation". In: *Mixtures: Estimation and applications* (2011), pp. 31–53 (cit. on p. 126).

[CM09] Olivier Cappé and Eric Moulines. "On-line expectation–maximization algorithm for latent data models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.3 (2009), pp. 593–613 (cit. on p. 126).

[CCE24] Guillaume Carlier, Enis Chenchene, and Katharina Eichinger. "Wasserstein Medians: Robustness, PDE Characterization, and Numerics". In: *SIAM Journal on Mathematical Analysis* 56.5 (2024), pp. 6483–6520 (cit. on pp. 21, 37, 284).

[CE10] Guillaume Carlier and Ivar Ekeland. “Matching for teams”. In: *Economic theory* 42 (2010), pp. 397–418 (cit. on pp. 21, 37, 264, 284, 286, 287).

[Cha21] Dajlil Chafaï. “Random Projections, Marginals, and Moments”. In: (Nov. 2021) (cit. on p. 46).

[CTV25] Laetitia Chapel, Romain Tavenard, and Samuel Vaiter. *Differentiable Generalized Sliced Wasserstein Plans*. 2025. arXiv: 2505.22049 [cs.LG] (cit. on pp. 205, 240, 241).

[Cha+21] Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, François-David Collin, and Ghislain Durif. “Kernel operations on the GPU, with autodiff, without memory overflows”. In: *Journal of Machine Learning Research* 22.74 (2021), pp. 1–6 (cit. on p. 190).

[Che+] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. “PLOT: Prompt Learning with Optimal Transport for Vision-Language Models”. In: *The Eleventh International Conference on Learning Representations* (cit. on p. 204).

[CYL20] Xiongjie Chen, Yongxin Yang, and Yunpeng Li. “Augmented Sliced Wasserstein Distances”. In: *International Conference on Learning Representations*. 2020 (cit. on p. 241).

[CGT18] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. “Optimal transport for Gaussian mixture models”. In: *IEEE Access* 7 (2018), pp. 6269–6278 (cit. on p. 356).

[CNR25] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport. École d’Été de Probabilités de Saint-Flour XLIX – 2019*. 1st ed. Vol. 2417. Lecture Notes in Mathematics. Springer Cham, Apr. 2025, pp. XIV + 260 (cit. on pp. 2, 136, 354, 356).

[Chi+20] L. Chizat, P. Roussillon, F. Léger, F.X. Vialard, and G. Peyré. “Faster Wasserstein distance estimation with the Sinkhorn divergence”. In: *Adv. Neural Inf. Process. Syst.* 33 (2020), pp. 2257–2269 (cit. on p. 126).

[Chi+18a] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. “Scaling algorithms for unbalanced optimal transport problems”. In: *Mathematics of computation* 87.314 (2018), pp. 2563–2609 (cit. on p. 140).

[Chi+18b] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. “Unbalanced optimal transport: Dynamic and Kantorovich formulations”. In: *Journal of Functional Analysis* 274.11 (2018), pp. 3090–3123 (cit. on p. 135).

[CM19] Samir Chowdhury and Facundo Mémoli. “The Gromov–Wasserstein distance between networks and stable network invariants”. In: *Information and Inference: A Journal of the IMA* 8.4 (2019), pp. 757–787 (cit. on p. 252).

[CS08] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008 (cit. on pp. 324–326, 336–339, 341, 345, 349).

[CS10] Andreas Christmann and Ingo Steinwart. “Universal kernels on non-standard input spaces”. In: *Advances in neural information processing systems* 23 (2010) (cit. on pp. 337, 338, 340).

[Cla90] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990 (cit. on pp. 7, 8, 70, 117, 118).

[CLP23] Giovanni Conforti, Daniel Lacker, and Soumik Pal. *Projected Langevin dynamics and a gradient flow for entropic optimal transport*. 2023. arXiv: 2309.08598 [math.PR] (cit. on p. 2).

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297 (cit. on p. 336).

[Cos99] M Coste. “An introduction to o-minimal geometry, Inst. Rech”. In: *RAAG Notes, Institut de Recherche Mathématique de Rennes* (1999) (cit. on p. 118).

[Cou+17] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. “Joint distribution optimal transportation for domain adaptation”. In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 1, 126).

[Cou+16] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. “Optimal transport for domain adaptation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.9 (2016), pp. 1853–1865 (cit. on pp. 1, 204).

[Cov67] Thomas M Cover. “The number of linearly inducible orderings of points in d-space”. In: *SIAM Journal on Applied Mathematics* 15.2 (1967), pp. 434–439 (cit. on pp. 233, 234).

[CS25] Giacomo Cozzi and Filippo Santambrogio. “Long-time asymptotics of the sliced-wasserstein flow”. In: *SIAM Journal on Imaging Sciences* 18.1 (2025), pp. 1–19 (cit. on pp. 205, 353).

[CW36] Harald Cramér and Herman Wold. “Some theorems on distribution functions”. In: *Journal of the London Mathematical Society* 1.4 (1936), pp. 290–294 (cit. on p. 46).

[Cua92] CM Cuadras. “Probability distributions with given multivariate marginals and given dependence structure”. In: *Journal of multivariate analysis* 42.1 (1992), pp. 51–66 (cit. on p. 46).

[Cut13] Marco Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in neural information processing systems* 26 (2013) (cit. on pp. 17, 33, 54, 102, 126, 177, 179).

[CD14] Marco Cuturi and Arnaud Doucet. “Fast Computation of Wasserstein Barycenters”. In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research. Beijing, China: PMLR, June 2014, pp. 685–693 (cit. on pp. 21, 24, 37, 38, 40, 138, 265, 267, 269, 285, 296, 302, 317).

[Dal+23] Maxime Dalery, Genevieve Dusson, Virginie Ehrlacher, and Alexei Lozinski. *Nonlinear reduced basis using mixture Wasserstein barycenters: application to an eigenvalue problem inspired from quantum chemistry*. 2023. arXiv: 2307.15423 [math.NA] (cit. on p. 126).

[DKS12] Giorgio Dall’Aglio, Samuel Kotz, and Gabriella Salinetti. *Advances in probability distributions with given marginals: beyond the copulas*. Vol. 67. Springer Science & Business Media, 2012 (cit. on p. 46).

[Dan66] John M Danskin. “The theory of max-min, with applications”. In: *SIAM Journal on Applied Mathematics* 14.4 (1966), pp. 641–664 (cit. on p. 185).

[DD22] Damek Davis and Dmitriy Drusvyatskiy. “Proximal methods avoid active strict saddles of weakly convex functions”. In: *Foundations of Computational Mathematics* 22.2 (2022), pp. 561–606 (cit. on p. 78).

[DDJ23] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. *Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization*. 2023. arXiv: 2108.11832 [math.OC] (cit. on p. 78).

[Dav+20] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. “Stochastic subgradient method converges on tame functions”. In: *Foundations of computational mathematics* 20.1 (2020), pp. 119–154 (cit. on pp. 8, 12, 22, 28, 38, 56, 70, 72, 80, 103, 111, 119, 192, 269, 276).

[DH97] Peter Dayan and Geoffrey E Hinton. “Using expectation-maximization for reinforcement learning”. In: *Neural Computation* 9.2 (1997), pp. 271–278 (cit. on p. 126).

[DF14] Guido De Philippis and Alessio Figalli. “The Monge–Ampère equation and its link to optimal transportation”. In: *Bulletin of the American Mathematical Society* 51.4 (2014), pp. 527–580 (cit. on pp. 2, 165).

[De +16] Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. “BV estimates in optimal transportation and applications”. In: *Archive for Rational Mechanics and Analysis* 219 (2016), pp. 829–860 (cit. on p. 176).

[Del04] Julie Delon. “Midway image equalization”. In: *Journal of Mathematical Imaging and Vision* 21.2 (2004), pp. 119–134 (cit. on p. 139).

[DD20] Julie Delon and Agnes Desolneux. “A Wasserstein-type distance in the space of Gaussian mixture models”. In: *SIAM Journal on Imaging Sciences* 13.2 (2020), pp. 936–970 (cit. on pp. 13, 14, 17, 24, 25, 29, 30, 33, 40, 41, 126, 131, 132, 136, 166, 177, 179, 285, 305, 306, 356).

[DGS21] Julie Delon, Nathaël Gozlan, and Alexandre Saint-Dizier. *Generalized Wasserstein barycenters between probability measures living on different subspaces*. 2021 (cit. on pp. 21, 24, 25, 37, 40, 41, 261, 263–267, 285, 313, 317, 354).

[Dem+20] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. “Gromov-Wasserstein optimal transport to align single-cell multi-omics data”. In: *BioRxiv* (2020), pp. 2020–04 (cit. on p. 2).

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22 (cit. on pp. 126–128).

[Des+19] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. "Max-sliced Wasserstein distance and its use for gans". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10648–10656 (cit. on pp. 1, 103, 114).

[DZS18] Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. "Generative Modeling Using the Sliced Wasserstein Distance". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3483–3491 (cit. on pp. 1, 13, 16, 29, 32, 46, 51, 52, 55–58, 103, 110, 141, 205).

[DF92] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*. Vol. 2. Springer, 1992 (cit. on p. 165).

[Dor+14] G Doran, K Muandet, K Zhang, and B Schölkopf. "A Permutation-Based Kernel Conditional Independence Test". In: *30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*. AUAI Press. 2014, pp. 132–141 (cit. on p. 337).

[Dud69] Richard Mansfield Dudley. "The speed of mean Glivenko-Cantelli convergence". In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 40–50 (cit. on pp. 54, 102).

[DLV24] Théo Dumont, Théo Lacombe, and François-Xavier Vialard. "On the existence of Monge maps for the Gromov–Wasserstein problem". In: *Foundations of Computational Mathematics* (2024), pp. 1–48 (cit. on pp. 167, 215, 258).

[ER24] Matthias J Ehrhardt and Lindon Roberts. "Analyzing inexact hypergradients for bilevel learning". In: *IMA Journal of Applied Mathematics* 89.1 (2024), pp. 254–278 (cit. on p. 126).

[EW22] Ariel Elnekave and Yair Weiss. "Generating natural images with direct patch distributions matching". In: *European Conference on Computer Vision*. Springer. 2022, pp. 544–560 (cit. on p. 205).

[EH13] Paul Embrechts and Marius Hofert. "A note on generalized inverses". In: *Mathematical Methods of Operations Research* 77 (2013), pp. 423–432 (cit. on pp. 196, 221).

[End03] Craig K Enders. "Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data." In: *Psychological methods* 8.3 (2003), p. 322 (cit. on p. 126).

[EN97] Y. M. Ermoliev and V. I. Norkin. "Stochastic Generalized Gradient Method with Application to Insurance Risk Management". In: *IIASA Interim Report, IR-97-021, Laxenburg, Austria* (Apr. 1997) (cit. on pp. 18, 22, 34, 38, 185, 190, 191, 269, 275).

[Eva18] LawrenceCraig Evans. *Measure theory and fine properties of functions*. Routledge, 2018 (cit. on pp. 7, 195).

[Fat+21a] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. "Unbalanced minibatch optimal transport; applications to domain adaptation". In: *International conference on machine learning*. PMLR. 2021, pp. 3186–3197 (cit. on pp. 135, 137, 141).

[Fat+20] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. "Learning with minibatch Wasserstein: asymptotic and gradient properties". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2131–2141 (cit. on pp. 137, 141).

[Fat+21b] Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. *Minibatch optimal transport distances; analysis and applications*. 2021. arXiv: 2101.01792 [stat.ML] (cit. on pp. 70, 103, 137, 141, 166, 204).

[Faz+19] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. "Efficient and accurate estimation of Lipschitz constants for deep neural networks". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 164).

[FJG17] Aingeru Fernandez-Bertolin, Philippe Jaming, and Karlheinz Grochenig. "Determining point distributions from their projections". In: July 2017, pp. 164–168 (cit. on pp. 46, 49, 50).

[FH02] Jeffrey A Fessler and Alfred O Hero. "Space-alternating generalized expectation-maximization algorithm". In: *IEEE Transactions on signal processing* 42.10 (2002), pp. 2664–2677 (cit. on p. 126).

[Fey+19] J. Feydy, P. Roussillon, A. Trouvé, and P. Gori. "Fast and scalable optimal transport for brain tractograms". In: *Proceedings of MICCAI*. Springer. 2019, pp. 636–644 (cit. on p. 126).

[Fey20] Jean Feydy. “Geometric data analysis, beyond convolutions”. In: *PhD Thesis* (2020) (cit. on pp. 198, 199).

[Fey+17] Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. “Optimal transport for diffeomorphic registration”. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer. 2017, pp. 291–299 (cit. on p. 204).

[Fie23] Christian Fiedler. *Lipschitz and Hölder Continuity in Reproducing Kernel Hilbert Spaces*. 2023. arXiv: 2310.18078 [math.FA] (cit. on p. 347).

[Fig09] Alessio Figalli. “Regularity of optimal transport maps (after Ma-Trudinger-Wang and Loeper)”. In: *Séminaire Bourbaki* 2008 (2009), pp. 997–1011 (cit. on p. 165).

[Fig17] Alessio Figalli. *The Monge–Ampère equation and its applications*. European Mathematical Society, 2017 (cit. on pp. 2, 165).

[Fla+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. “POT: Python Optimal Transport”. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8 (cit. on pp. 21, 25, 37, 41, 69, 82, 187–189, 240, 264, 265, 267, 268, 285, 302, 317).

[FW+56] Marguerite Frank, Philip Wolfe, et al. “An algorithm for quadratic programming”. In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110 (cit. on pp. 21, 37, 252).

[Fri98] Nir Friedman. “The Bayesian structural EM algorithm”. In: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. 1998, pp. 129–138 (cit. on p. 126).

[Fro+15] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. “Learning with a Wasserstein loss”. In: *Advances in neural information processing systems* 28 (2015) (cit. on p. 2).

[FSG13] Kenji Fukumizu, Le Song, and Arthur Gretton. “Kernel Bayes’ rule: Bayesian inference with positive definite kernels”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3753–3783 (cit. on p. 337).

[FHP] Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. “Transformers are Universal In-context Learners”. In: *The Thirteenth International Conference on Learning Representations* (cit. on p. 2).

[GLR18] Bruno Galerne, Arthur Leclaire, and Julien Rabin. “A texture synthesis model based on semi-discrete optimal transport in patch space”. In: *SIAM Journal on Imaging Sciences* 11.4 (2018), pp. 2456–2493 (cit. on pp. 15, 31, 127, 141).

[Gal17] Alfred Galichon. “A survey of some recent applications of optimal transport methods to econometrics”. In: *The Econometrics Journal* 20.2 (2017), pp. C1–C11 (cit. on p. 204).

[GTG07] Kuzman Ganchev, Ben Taskar, and João Gama. “Expectation maximization and posterior constraints”. In: *Advances in neural information processing systems* 20 (2007) (cit. on p. 126).

[GM96] Wilfrid Gangbo and Robert J McCann. “The geometry of optimal transportation”. In: *Acta Mathematica* 177 (1996), pp. 113–161 (cit. on pp. 4, 165, 177).

[GG99] Richard J Gardner and Peter Gritzmann. *Uniqueness and complexity in discrete tomography*. 1999 (cit. on p. 46).

[GEB16] Leon Gatys, Alexander Ecker, and Matthias Bethge. “A Neural Algorithm of Artistic Style”. In: *Journal of Vision* 16.12 (Sept. 2016), p. 326 (cit. on p. 140).

[GEB15] Leon Gatys, Alexander S Ecker, and Matthias Bethge. “Texture synthesis using convolutional neural networks”. In: *Advances in neural information processing systems* 28 (2015) (cit. on pp. 15, 31, 127, 141, 158).

[Gen+19] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. “Sample complexity of Sinkhorn divergences”. In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1574–1583 (cit. on pp. 102, 126).

[GPC17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. *GAN and VAE from an Optimal Transport Point of View*. 2017. arXiv: 1706.01807 [stat.ML] (cit. on p. 2).

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. "Learning generative models with Sinkhorn divergences". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1608–1617 (cit. on pp. 54, 353).

[GNB22] Promit Ghosal, Marcel Nutz, and Espen Bernton. "Stability of entropic optimal transport and Schrödinger bridges". In: *Journal of Functional Analysis* 283.9 (2022), p. 109622 (cit. on p. 297).

[Gil92] Jean Charles Gilbert. "Automatic differentiation and iterative processes". In: *Optimization methods and software* 1.1 (1992), pp. 13–21 (cit. on pp. 126, 144).

[GKZ19] Nikita A Gladkov, Alexander V Kolesnikov, and Alexander P Zimin. "On multistochastic Monge–Kantorovich problem, bitwise operations, and fractals". In: *Calculus of Variations and Partial Differential Equations* 58 (2019), pp. 1–33 (cit. on p. 46).

[Góm+09] Luis Gómez-Chova, Gustavo Camps-Valls, Lorenzo Bruzzone, and Javier Calpe-Maravilla. "Mean map kernel methods for semisupervised cloud classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 48.1 (2009), pp. 207–220 (cit. on p. 337).

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 15, 31, 141, 164).

[Gor+19] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. "Obtaining fairness using optimal transport theory". In: *International conference on machine learning*. PMLR. 2019, pp. 2357–2365 (cit. on pp. 2, 284).

[GJ20] Nathael Gozlan and Nicolas Juillet. "On a mixture of Brenier and Strassen theorems". In: *Proceedings of the London Mathematical Society* 120.3 (2020), pp. 434–463 (cit. on p. 2).

[Goz+17] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11 (2017), pp. 3327–3405 (cit. on p. 299).

[GSZ18] Nathael Gozlan, Paul-Marie Samson, and Pierre-André Zitt. "Notes de cours sur le transport optimal". In: *Online resources* (2018) (cit. on p. 198).

[GVS17] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. "Neural expectation maximization". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 126).

[Gre+06] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. "A kernel method for the two-sample-problem". In: *Advances in neural information processing systems* 19 (2006) (cit. on pp. 102, 325).

[Gre+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773 (cit. on pp. 337, 349).

[GW08] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008 (cit. on p. 126).

[GJS23] David Groisser, Sungkyu Jung, and Armin Schwartzman. *A genericity property of Fréchet sample means on Riemannian manifolds*. 2023. arXiv: 2309.13823 [math.PR] (cit. on p. 354).

[Gul+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. "Improved training of Wasserstein GANs". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 54, 204).

[HC11] Robert Hable and Andreas Christmann. "On qualitative robustness of support vector machines". In: *Journal of Multivariate Analysis* 102.6 (2011), pp. 993–1007 (cit. on p. 341).

[Har+20] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020), pp. 357–362 (cit. on p. 81).

[Hei+21] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. "A sliced Wasserstein loss for neural texture synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9412–9420 (cit. on pp. 55, 103, 205).

[Hep56] A Heppes. "On the determination of probability distributions of more dimensions by their projections". In: *Acta Mathematica Hungarica* 7.3-4 (1956), pp. 403–410 (cit. on p. 46).

[HCD25] Johannes Hertrich, Antonin Chambolle, and Julie Delon. *On the Relation between Rectified Flows and Optimal Transport*. 2025. arXiv: 2505.19712 [cs.LG] (cit. on p. 204).

[HHR22] Johannes Hertrich, Antoine Houdard, and Claudia Redenbach. "Wasserstein patch prior for image superresolution". In: *IEEE Transactions on Computational Imaging* 8 (2022), pp. 693–704 (cit. on p. 204).

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851 (cit. on p. 126).

[Ho+17] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. "Multilevel clustering via Wasserstein means". In: *International conference on machine learning*. PMLR. 2017, pp. 1501–1509 (cit. on p. 284).

[Hou+23] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. "A generative model for texture synthesis based on optimal transport between feature distributions". In: *Journal of Mathematical Imaging and Vision* 65.1 (2023), pp. 4–28 (cit. on pp. 127, 141).

[Hua+23] Yu-Jui Huang, Shih-Chun Lin, Yu-Chih Huang, Kuan-Hui Lyu, Hsin-Hua Shen, and Wan-Yi Lin. "On Characterizing Optimal Wasserstein GAN Solutions for Non-Gaussian Data". In: (2023) (cit. on p. 103).

[HPC22] Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini. "Optimal transport improves cell–cell similarity inference in single-cell omics data". In: *Bioinformatics* 38.8 (2022), pp. 2169–2177 (cit. on p. 2).

[Hur08] Glenn Hurlbert. "A short proof of the birkhoff-von neumann theorem". In: *preprint (unpublished)* (2008) (cit. on pp. 225, 227).

[HR21] Jan-Christian Hütter and Philippe Rigollet. "Minimax estimation of smooth optimal transport maps". In: *The Annals of Statistics* 49.2 (2021), pp. 1166–1194 (cit. on p. 165).

[Hyt+16] Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*. Vol. 12. Springer, 2016 (cit. on p. 325).

[Jay+15] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. "Kernel methods on Riemannian manifolds with Gaussian RBF kernels". In: *IEEE transactions on pattern analysis and machine intelligence* 37.12 (2015), pp. 2464–2477 (cit. on p. 337).

[Jin+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. "How to escape saddle points efficiently". In: *International conference on machine learning*. PMLR. 2017, pp. 1724–1732 (cit. on p. 78).

[Jin+21] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. "On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points". In: *Journal of the ACM (JACM)* 68.2 (2021), pp. 1–29 (cit. on p. 78).

[Joe93] Harry Joe. "Parametric families of multivariate distributions with given margins". In: *Journal of multivariate analysis* 46.2 (1993), pp. 262–282 (cit. on p. 46).

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The variational formulation of the Fokker–Planck equation". In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17 (cit. on pp. 2, 356).

[KMK23] Bernard Kamsu-Foguem, Shester Landry Msouobu Gueuwou, and Cheick Abdoul Kadir A Kounta. "Generative Adversarial Networks based on optimal transport: a survey". In: *Artificial Intelligence Review* 56.7 (2023), pp. 6723–6773 (cit. on p. 1).

[Kan42] Leonid V Kantorovich. "On the translocation of masses". In: *Dokl. Akad. Nauk. USSR (NS)*. Vol. 37. 1942, pp. 199–201 (cit. on pp. 1, 2, 126).

[KLA19] T. Karras, S. Laine, and T. Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of IEEE CVPR*. 2019, pp. 4401–4410 (cit. on p. 126).

[Kar+18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*. 2018 (cit. on pp. 46, 51, 55).

[KR19] Nabil Kazi-Tani and Didier Rullière. "On a construction of multivariate distributions given some multidimensional marginals". In: *Advances in Applied Probability* 51.2 (2019), pp. 487–513 (cit. on p. 46).

[Kel64] Hans G Kellerer. "Verteilungsfunktionen mit gegebenen Marginalverteilungen". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 3 (1964), pp. 247–270 (cit. on p. 46).

[Kel17] John L Kelley. *General topology*. Courier Dover Publications, 2017 (cit. on pp. 168, 170).

[Kim22] Minyoung Kim. "Differentiable Expectation-Maximization for Set Representation Learning". In: *International Conference on Learning Representations*. 2022 (cit. on pp. 126, 127).

[KP17] Young-Heon Kim and Brendan Pass. "Wasserstein barycenters over Riemannian manifolds". In: *Adv. Math.* 307 (2017), pp. 640–683 (cit. on p. 285).

[KPS20] Young-Heon Kim, Brendan Pass, and David J Schneider. "Optimal transport and barycenters for dendritic measures". In: *Pure and Applied Analysis* 2.3 (2020), pp. 581–601 (cit. on p. 214).

[KW14] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *Int. Conf. on Learning Representations*. 2014 (cit. on pp. 126, 164).

[KT24] Jun Kitagawa and Asuka Takatsu. *Sliced optimal transport: is it a suitable replacement?* 2024. arXiv: 2311.15874 [math.MG] (cit. on p. 354).

[Kol+19a] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. "Generalized sliced wasserstein distances". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 241).

[Kol+19b] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. "Sliced Wasserstein Auto-Encoders". In: *International Conference on Learning Representations*. 2019 (cit. on pp. 55, 56, 58, 126, 205).

[KRH18] Soheil Kolouri, Gustavo K. Rohde, and Heiko Hoffmann. "Sliced Wasserstein Distance for Learning Gaussian Mixture Models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on p. 205).

[Kor+22] Alexander Korotin, Vage Egiazarian, Lingxiao Li, and Evgeny Burnaev. "Wasserstein iterative networks for barycenter estimation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15672–15686 (cit. on pp. 284, 285).

[KSB22] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. "Neural Optimal Transport". In: *The Eleventh International Conference on Learning Representations*. 2022 (cit. on p. 1).

[Kro+19] Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. "On the complexity of approximating Wasserstein barycenters". In: *International conference on machine learning*. PMLR. 2019, pp. 3530–3540 (cit. on pp. 21, 37).

[Lac16] Simon Lacoste-Julien. *Convergence Rate of Frank-Wolfe for Non-Convex Objectives*. 2016. arXiv: 1607.00345 [math.OC] (cit. on p. 253).

[Lam+22] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 14434–14447 (cit. on p. 356).

[LGL21] Lucas de Lara, Alberto González-Sanz, and Jean-Michel Loubes. *A Consistent Extension of Discrete Optimal Transport Maps for Machine Learning Applications*. 2021. arXiv: 2102.08644 [math.ST] (cit. on p. 165).

[LS22] Hugo Lavenant and Filippo Santambrogio. "The flow map of the fokker–planck equation does not provide optimal transport". In: *Applied Mathematics Letters* 133 (2022), p. 108225 (cit. on p. 2).

[Le+24] Tung Le, Khai Nguyen, Shanlin Sun, Nhat Ho, and Xiaohui Xie. "Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 23188–23198 (cit. on p. 205).

[LDD23] Leclaire, Delon, and Desolneux. "Optimal Transport Between GMM for Texture Synthesis". In: *Proceedings of SSVM 2023*. 2023 (cit. on pp. 15, 31, 126, 127, 141).

[Lee+19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. "Sliced wasserstein discrepancy for unsupervised domain adaptation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10285–10295 (cit. on p. 205).

[Léo12] Christian Léonard. “From the Schrödinger problem to the Monge–Kantorovich problem”. In: *Journal of Functional Analysis* 262.4 (2012), pp. 1879–1920 (cit. on p. 2).

[LHL12] S. Levy, F. Hirsch, and G. Lacombe. *Elements of Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 2012 (cit. on p. 63).

[LBM24] Bruno Lévy, Yann Brenier, and Roya Mohayaee. “Monge-Ampère gravity: From the large deviation principle to cosmological simulations through optimal transport”. In: *Physical Review D* 110.6 (2024), p. 063550 (cit. on p. 2).

[Li+17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. “Mmd gan: Towards deeper understanding of moment matching network”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 337).

[Li+] Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. “A Convergent Single-Loop Algorithm for Relaxation of Gromov-Wasserstein in Graph Data”. In: *The Eleventh International Conference on Learning Representations* (cit. on p. 252).

[LM25] Shiying Li and Caroline Moosmueller. *Measure transfer via stochastic slicing and matching*. 2025. arXiv: 2307.05705 [math.NA] (cit. on pp. 70, 80, 205).

[LMS18] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. “Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures”. In: *Inventiones mathematicae* 211.3 (2018), pp. 969–1117 (cit. on pp. 15, 31, 135, 140, 356).

[Lin23] Johannes von Lindheim. “Simple approximative algorithms for free-support Wasserstein barycenters”. In: *Computational Optimization and Applications* 85.1 (2023), pp. 213–246 (cit. on pp. 285, 293, 317).

[Lin70] S Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors (Doctoral dissertation, Master’s Thesis”. PhD thesis. MA thesis. University of Helsinki, 1970 (cit. on p. 126).

[Liu+24] Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. “Expected Sliced Transport Plans”. In: *The Thirteenth International Conference on Learning Representations*. 2024 (cit. on pp. 18–20, 34, 35, 37, 161, 205, 206, 234–236, 240).

[Liu+19] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. “Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 4104–4113 (cit. on pp. 1, 55, 114).

[Loi+07] Eliane Maria Loiola, Nair Maria Maia De Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. “A survey for the quadratic assignment problem”. In: *European journal of operational research* 176.2 (2007), pp. 657–690 (cit. on p. 252).

[LVD20] Jonathan Lorraine, Paul Vicol, and David Duvenaud. “Optimizing millions of hyperparameters by implicit differentiation”. In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 1540–1552 (cit. on p. 126).

[Lui+18] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. “Differential properties of sinkhorn approximation for learning with wasserstein distance”. In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on p. 126).

[Luz+23] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G Baraniuk, and Michael J Henry. “Evaluating generative networks using Gaussian mixtures of image features”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 279–288 (cit. on p. 126).

[Mah+23] Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. “Fast Optimal Transport through Sliced Generalized Wasserstein Geodesics”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 35350–35385 (cit. on pp. 18–20, 34–36, 161, 205, 212, 214, 215, 232, 233, 244, 245, 257).

[MMM18] Szymon Majewski, Błażej Miasojedow, and Eric Moulines. *Analysis of nonsmooth stochastic approximation: the differential inclusion approach*. 2018. arXiv: 1805.01916 [math.OC] (cit. on pp. 8, 70, 103).

[Man+24] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. “Plugin estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 52.3 (2024), pp. 966–998 (cit. on pp. 165, 182).

[ML18] Haggai Maron and Yaron Lipman. "(Probably) concave graph matching". In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on p. 258).

[MAM25] Simon Mataigne, P. -A. Absil, and Nina Miolane. *On the approximation of the Riemannian barycenter*. 2025. arXiv: 2504.15671 [math.DG] (cit. on p. 354).

[MK07] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007 (cit. on pp. 126, 129).

[MP00] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2000 (cit. on p. 126).

[MO20] Sheheryar Mehmood and Peter Ochs. "Automatic differentiation of some first-order methods in parametric optimization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1584–1594 (cit. on pp. 126, 144).

[Mém11] Facundo Mémoli. "Gromov–Wasserstein distances and the metric approach to object matching". In: *Foundations of computational mathematics* 11 (2011), pp. 417–487 (cit. on pp. 21, 37, 103, 251–253).

[MN22] Facundo Mémoli and Tom Needham. "Distance distributions and inverse problems for metric measure spaces". In: *Studies in Applied Mathematics* 149.4 (2022), pp. 943–1001 (cit. on p. 258).

[MT21] Quentin Merigot and Boris Thibert. "Optimal transport: discretization and algorithms". In: *Handbook of numerical analysis*. Vol. 22. Elsevier, 2021, pp. 133–212 (cit. on pp. 166, 195).

[MDC20] Quentin Mérigot, Alex Delalande, and Frederic Chazal. "Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3186–3196 (cit. on pp. 23, 39, 288, 354).

[MSS21] Quentin Mérigot, Filippo Santambrogio, and Clément Sarrazin. "Non-asymptotic convergence bounds for Wasserstein approximation using point clouds". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12810–12821 (cit. on pp. 56, 166).

[Mi+18] Liang Mi, Wen Zhang, Xianfeng Gu, and Yalin Wang. "Variational wasserstein clustering". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 322–337 (cit. on p. 284).

[Mia+21] Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. "A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention". In: *ICLR 2021-The Ninth International Conference on Learning Representations*. 2021 (cit. on p. 127).

[Mic86] Charles A Micchelli. "Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions". In: *Constr. Approx* 2 (1986), pp. 11–22 (cit. on p. 325).

[MXZ06] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. "Universal Kernels." In: *Journal of Machine Learning Research* 7.12 (2006) (cit. on p. 337).

[MG14] Mario Micheli and Joan A Glaunès. "Matrix-valued kernels for shape deformation analysis". In: *Geometry, Imaging and Computing* 1.1 (2014), pp. 57–139 (cit. on pp. 331–333).

[MGN24] V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin. *Methods of Nonconvex Optimization*. 2024. arXiv: 2406.10406 [math.OC] (cit. on p. 185).

[Miy+18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. "Spectral Normalization for Generative Adversarial Networks". In: *International Conference on Learning Representations* 6 (2018) (cit. on p. 164).

[Mod17] Klas Modin. "Geometry of matrix decompositions seen through optimal transport and information geometry". In: *Journal of Geometric Mechanics* 9.3 (2017) (cit. on p. 2).

[Mon81] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Mem. Math. Phys. Acad. Royale Sci.* (1781), pp. 666–704 (cit. on pp. 1, 2, 126).

[MM21] Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. "Wasserstein barycenter for multi-source domain adaptation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 16785–16793 (cit. on pp. 204, 284).

[MMS24a] Eduardo Fernandes Montesuma, Fred Maurice Ngole Mboula, and Antoine Souloumiac. "Optimal Transport for Domain Adaptation through Gaussian Mixture Models". In: *Transactions on Machine Learning Research* (2024) (cit. on p. 1).

[MMS24b] Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. "Lighter, better, faster multi-source domain adaptation with gaussian mixture models and optimal transport". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2024, pp. 21–38 (cit. on p. 126).

[Moo96] T.K. Moon. "The expectation-maximization algorithm". In: *IEEE Signal Processing Magazine* 13.6 (1996), pp. 47–60 (cit. on p. 128).

[MC23] Caroline Moosmüller and Alexander Cloninger. "Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations". In: *Information and Inference: A Journal of the IMA* 12.1 (2023), pp. 363–389 (cit. on p. 353).

[Mua+12] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. "Learning from distributions via support measure machines". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 337).

[Mua+17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141 (cit. on pp. 325, 336).

[Mül97] Alfred Müller. "Integral probability metrics and their generating classes of functions". In: *Advances in applied probability* 29.2 (1997), pp. 429–443 (cit. on p. 326).

[Muz+21] Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. *Near-optimal estimation of smooth transport maps with kernel sums-of-squares*. 2021. arXiv: 2112.01907 [stat.ML] (cit. on p. 355).

[Nad64] E. A. Nadaraya. "On Estimating Regression". In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142 (cit. on pp. 329, 336).

[NS25] Navid NaderiAlizadeh and Rohit Singh. "Aggregating residue-level protein language model embeddings with optimal transport". In: *Bioinformatics Advances* 5.1 (2025), vbaf060 (cit. on p. 204).

[Nad21] Kimia Nadjahi. "Sliced-Wasserstein distance for large-scale machine learning: theory, methodology and extensions". PhD thesis. Institut polytechnique de Paris, 2021 (cit. on p. 51).

[Nad+20a] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. "Approximate Bayesian Computation with the Sliced-Wasserstein Distance". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 5470–5474 (cit. on pp. 55, 58).

[Nad+20b] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. "Statistical and Topological Properties of Sliced Probability Divergences". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 20802–20812 (cit. on pp. 55, 57, 103, 205).

[Nad+19] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. "Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 55, 58).

[NB06] Nikolaos Nasios and Adrian G Bors. "Variational learning for Gaussian mixture models". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.4 (2006), pp. 849–862 (cit. on p. 126).

[NP23] Luca Nenna and Brendan Pass. "Transport type metrics on the space of probability measures involving singular base measures". In: *Applied Mathematics & Optimization* 87.2 (2023), p. 28 (cit. on pp. 19, 20, 35, 205–212, 214, 355).

[Ng13] Shu-Kay Ng. "Recent developments in expectation-maximization methods for analyzing complex data". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (2013), pp. 415–431 (cit. on p. 126).

[NNH23] Khai Nguyen, Dang Nguyen, and Nhat Ho. "Self-attention amortized distributional projection optimization for sliced wasserstein point-cloud reconstruction". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 26008–26030 (cit. on p. 205).

[Nie+22] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. "Statistical, robustness, and computational guarantees for sliced Wasserstein distances". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28179–28193 (cit. on p. 103).

[NW06] J Nocedal and SJ Wright. "Numerical optimization". In: *Springer Series in Operations Research and Financial Engineering* (2006) (cit. on p. 149).

[Orl97] James B Orlin. "A polynomial time primal network simplex algorithm for minimum cost flows". In: *Mathematical Programming* 78 (1997), pp. 109–129 (cit. on p. 3).

[Ott01] Felix Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: (2001) (cit. on pp. 2, 356).

[PPC10] Nicolas Papadakis, Edoardo Provenzi, and Vicent Caselles. "A variational model for histogram transfer of color images". In: *IEEE Transactions on Image Processing* 20.6 (2010), pp. 1682–1695 (cit. on p. 139).

[PS25] Sangmin Park and Dejan Slepčev. "Geometry and analytic properties of the sliced Wasserstein space". In: *Journal of Functional Analysis* 289.7 (2025), p. 110975 (cit. on p. 354).

[Par62] Emanuel Parzen. "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076 (cit. on p. 336).

[Pas+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on pp. 82, 130, 131, 144, 156, 268).

[PC19a] François-Pierre Paty and Marco Cuturi. "Subspace robust Wasserstein distances". In: *International conference on machine learning*. PMLR. 2019, pp. 5072–5081 (cit. on pp. 103, 114).

[PdC20] François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. "Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1222–1232 (cit. on pp. 16–18, 25, 32–34, 41, 165, 166, 171, 180, 182, 184, 185, 187).

[Pen06] Xavier Pennec. "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements". In: *Journal of Mathematical Imaging and Vision* 25.1 (2006), pp. 127–154 (cit. on p. 354).

[Per+16] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. "Mapping estimation for discrete optimal transport". In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on pp. 165, 166).

[PP08] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Version 20081110. Oct. 2008 (cit. on p. 145).

[PLK25] Marguerite Petit-Talamon, Marc Lambert, and Anna Korba. *Variational Inference with Mixtures of Isotropic Gaussians*. 2025. arXiv: 2506.13613 [stat.ML] (cit. on p. 356).

[PC19b] G. Peyré and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 51.1 (2019), pp. 1–44 (cit. on pp. 1, 3, 51, 54, 92, 102, 126, 127, 166, 179, 187, 198–200, 204, 234, 254, 285, 300).

[Pey19] Gabriel Peyré. "Course notes on Computational Optimal Transport". In: *Couse Notes* (2019) (cit. on pp. 225, 227).

[Pey25] Gabriel Peyré. *Optimal Transport for Machine Learners*. 2025. arXiv: 2505.06589 [stat.ML] (cit. on p. 1).

[PCS16] Gabriel Peyré, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices". In: *International conference on machine learning*. PMLR. 2016, pp. 2664–2672 (cit. on pp. 252, 253).

[PK07] François Pitié and Anil Kokaram. "The linear monge-kantorovitch linear colour mapping for example-based colour transfer". In: *4th European conference on visual media production*. IET. 2007, pp. 1–9 (cit. on p. 139).

[PKD07] François Pitié, Anil C Kokaram, and Rozenn Dahyot. "Automated colour grading using colour distribution transfer". In: *Computer Vision and Image Understanding* 107.1-2 (2007), pp. 123–137 (cit. on p. 139).

[PW14] Yury Polyanskiy and Yihong Wu. "Lecture notes on information theory". In: *Lecture Notes for ECE563 (UIUC) and* 6.2012-2016 (2014), p. 7 (cit. on p. 298).

[Pon+21] Mathieu Pont, Jules Vidal, Julie Delon, and Julien Tierny. "Wasserstein distances, geodesics and barycenters of merge trees". In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), pp. 291–301 (cit. on p. 204).

[PN24] Aram-Alexandre Pooladian and Jonathan Niles-Weed. *Entropic estimation of optimal transport maps*. 2024. arXiv: 2109.12004 [math.ST] (cit. on p. 165).

[Pra07] Aldo Pratelli. "On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. Vol. 43. Elsevier. 2007, pp. 1–13 (cit. on pp. 4, 165, 177).

[RDG09] Julien Rabin, Julie Delon, and Yann Gousseau. "A statistical approach to the matching of local features". In: *SIAM Journal on Imaging Sciences* 2.3 (2009), pp. 931–958 (cit. on p. 204).

[RFP14] Julien Rabin, Sira Ferradans, and Nicolas Papadakis. "Adaptive color transfer with relaxed optimal transport". In: *2014 IEEE international conference on image processing (ICIP)*. IEEE. 2014, pp. 4852–4856 (cit. on p. 139).

[Rab+12] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. "Wasserstein barycenter and its application to texture mixing". In: *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*. Springer. 2012, pp. 435–446 (cit. on pp. 5, 12, 18, 28, 34, 46, 51, 55, 102, 127, 139, 204, 205, 242, 243, 284, 285).

[RR06] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume 1: Theory*. Springer Science & Business Media, 2006 (cit. on p. 1).

[RMC16] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. 2016 (cit. on p. 142).

[RGC17] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2 (2017), p. 47 (cit. on p. 2).

[Ras03] Carl Edward Rasmussen. "Gaussian processes in machine learning". In: *Summer school on machine learning*. Springer, 2003, pp. 63–71 (cit. on p. 126).

[Rei+02] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. "Color transfer between images". In: *IEEE Computer graphics and applications* 21.5 (2002), pp. 34–41 (cit. on p. 139).

[Rén52] Alfréd Rényi. "On projections of probability distributions". In: *Acta Math. Acad. Sci. Hungar* 3.3 (1952), pp. 131–142 (cit. on p. 46).

[RM15] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538 (cit. on p. 164).

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286 (cit. on p. 126).

[Rig22] Philippe Rigollet. *An Optimization Perspective on Sampling Using Optimal Transport*. https://theodumont.github.io/assets/pdf/rigollet_ot.pdf. Lecture notes from a course given at Sorbonne Université, May 31 – June 9, 2022. Notes taken by Théo Dumont. 2022 (cit. on p. 2).

[Roc97] R Tyrrell Rockafellar. *Convex analysis*. Vol. 28. Princeton university press, 1997 (cit. on p. 7).

[RAG05] Steven Roman, S Axler, and FW Gehring. *Advanced linear algebra*. Vol. 3. Springer, 2005 (cit. on p. 322).

[Ros56] Murray Rosenblatt. "Remarks on some nonparametric estimates of a density function". In: *Ann. Math. Stat* 27 (1956), pp. 832–837 (cit. on p. 336).

[RKB21] Litu Rout, Alexander Korotin, and Evgeny Burnaev. "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations*. 2021 (cit. on p. 1).

[Row+19] Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. "Orthogonal estimation of Wasserstein distances". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 186–195 (cit. on p. 205).

[RMB24] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. "Finding global minima via kernel approximations". In: *Mathematical Programming* (2024), pp. 1–82 (cit. on p. 189).

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987 (cit. on pp. 325, 349).

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Cambridge, MA: MIT Press, 1986, pp. 318–362 (cit. on p. 126).

[SG76] Sartaj Sahni and Teofilo Gonzalez. "P-complete approximation problems". In: *Journal of the ACM (JACM)* 23.3 (1976), pp. 555–565 (cit. on p. 252).

[SS16] Saburou Saitoh and Yoshihiro Sawano. *Theory of reproducing kernels and applications*. Vol. 44. Springer, 2016 (cit. on p. 337).

[Sal+18] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. "Improving GANs Using Optimal Transport". In: *International Conference on Learning Representations*. 2018 (cit. on p. 204).

[Sal+22] Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnes Desolneux. "Can Push-forward Generative Models Fit Multimodal Distributions?" In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 10766–10779 (cit. on pp. 164, 167).

[SDD23] Antoine Salmona, Agnes Desolneux, and Julie Delon. "Gromov-Wasserstein-like Distances in the Gaussian Mixture Models Space". In: *Transactions on Machine Learning Research* (2023) (cit. on p. 164).

[SCR12] Rajhans Samdani, Ming-Wei Chang, and Dan Roth. "Unified expectation maximization". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2012, pp. 688–698 (cit. on p. 126).

[San15] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63 (2015), p. 94 (cit. on pp. 1, 2, 4, 5, 54, 59, 102, 165, 169, 175–178, 182, 183, 198, 199, 204, 207, 209, 214, 215, 221, 225, 229, 233, 237, 264, 274, 287, 288, 295).

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural computation* 10.5 (1998), pp. 1299–1319 (cit. on p. 336).

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002 (cit. on p. 337).

[Seg+18] Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. "Large-Scale Optimal Transport and Mapping Estimation". In: *ICLR 2018-International Conference on Learning Representations*. 2018, pp. 1–15 (cit. on pp. 165, 166).

[Set22] Michio Seto. *A Fock space approach to the theory of strictly positive kernels*. 2022. arXiv: 2208.02980 [math.FA] (cit. on p. 325).

[Sha+19] Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. "Truncated back-propagation for bilevel optimization". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1723–1732 (cit. on p. 126).

[Si+21] Nian Si, Karthyek Murthy, Jose Blanchet, and Viet Anh Nguyen. "Testing group fairness via optimal transport projections". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9649–9659 (cit. on p. 2).

[Sil+20] Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. "A general approach to fairness with optimal transport". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 3633–3640 (cit. on p. 2).

[SZ15] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV] (cit. on pp. 15, 31, 140).

[SDT25] Keanu Sisouk, Julie Delon, and Julien Tierny. *A User's Guide to Sampling Strategies for Sliced Optimal Transport*. 2025. arXiv: 2502.02275 [cs.LG] (cit. on p. 205).

[Smo+07] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. "A Hilbert space embedding for distributions". In: *International conference on algorithmic learning theory*. Springer. 2007, pp. 13–31 (cit. on p. 325).

[Sol+15] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics (ToG)* 34.4 (2015), pp. 1–11 (cit. on pp. 244, 284).

[Sol+16] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. "Entropic metric alignment for correspondence problems". In: *ACM Transactions on Graphics (ToG)* 35.4 (2016), pp. 1–13 (cit. on p. 252).

[Son+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-based generative modeling through stochastic differential equations". In: *International Conference on Learning Representations* (2020) (cit. on pp. 164, 167).

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." In: *Journal of Machine Learning Research* 12.7 (2011) (cit. on pp. 327, 337, 341).

[Sri+10] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561 (cit. on pp. 327, 341).

[SZ21] Ingo Steinwart and Johanna F. Ziegel. "Strictly proper kernel scores and characteristic kernels on compact spaces". In: *Applied and Computational Harmonic Analysis* 51 (2021), pp. 510–542 (cit. on pp. 337, 338).

[TCD23] Julián Tachella, Dongdong Chen, and Mike Davies. "Sensing Theorems for Unsupervised Learning in Linear Inverse Problems". In: *Journal of Machine Learning Research* 24.39 (2023), pp. 1–45 (cit. on p. 52).

[Tan23] Eloi Tanguy. "Convergence of SGD for Training Neural Networks with Sliced Wasserstein Losses". In: *Transactions on Machine Learning Research* (2023) (cit. on pp. 25, 43, 101).

[Tan25] Eloi Tanguy. *Explicit Universal and Approximate-Universal Kernels on Compact Metric Spaces*. 2025. arXiv: 2506.03661 [math.FA] (cit. on pp. 26, 319, 336).

[TCD25] Eloi Tanguy, Laetitia Chapel, and Julie Delon. *Sliced Optimal Transport Plans*. 2025. arXiv: 2508.01243 [math.OC] (cit. on pp. 25, 161, 204).

[TDG24] Eloi Tanguy, Julie Delon, and Nathaël Gozlan. *Computing Barycentres of Measures for Generic Transport Costs*. 2024. arXiv: 2501.04016 [math.NA] (cit. on pp. 25, 126, 132, 138, 261, 284).

[TFD24a] Eloi Tanguy, Rémi Flamary, and Julie Delon. "Properties of Discrete Sliced Wasserstein Losses". In: *Mathematics of Computation* (June 2024) (cit. on pp. 25, 43, 54).

[TFD24b] Eloi Tanguy, Rémi Flamary, and Julie Delon. "Reconstructing discrete measures from projections. Consequences on the empirical Sliced Wasserstein Distance". In: *Comptes Rendus. Mathématique* 362 (2024), pp. 1121–1129 (cit. on pp. 25, 43, 45, 52).

[TDD25] Tanguy, Eloi, Desolneux, Agnès, and Delon, Julie. "Constrained Approximate Optimal Transport Maps". In: *ESAIM: COCV* 31 (2025), p. 70 (cit. on pp. 25, 161, 164).

[Tar97] Robert E Tarjan. "Dynamic trees as search trees via euler tours, applied to the network simplex algorithm". In: *Mathematical Programming* 78.2 (1997), pp. 169–177 (cit. on pp. 3, 187, 253).

[TPG16] Guillaume Tartavel, Gabriel Peyré, and Yann Gousseau. "Wasserstein Loss for Image Synthesis and Restoration". In: *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1726–1755 (cit. on pp. 55, 103).

[Tay17] Adrien B Taylor. "Convex interpolation and performance estimation of first-order methods for convex optimization." PhD thesis. Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2017 (cit. on pp. 184, 186).

[THG17] Adrien B Taylor, Julien M Hendrickx, and François Glineur. "Smooth strongly convex interpolation and exact worst-case performance of first-order methods". In: *Mathematical Programming* 161 (2017), pp. 307–345 (cit. on p. 186).

[Ton+24] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. "Improving and generalizing flow-based generative models with minibatch optimal transport". In: *Transactions on Machine Learning Research* (2024), pp. 1–34 (cit. on pp. 2, 137, 141, 204).

[TGR24] William Torous, Florian Gunsilius, and Philippe Rigollet. "An optimal transport approach to estimating causal effects via nonlinear difference-in-differences". In: *Journal of Causal Inference* 12.1 (2024), p. 20230004 (cit. on p. 2).

[Tro12] Joel A Tropp. "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4 (2012), pp. 389–434 (cit. on p. 93).

[Vac+24] Adrien Vacher, Boris Muzellec, Francis Bach, Francois-Xavier Vialard, and Alessandro Rudi. "Optimal estimation of smooth transport maps with kernel SoS". In: *SIAM Journal on Mathematics of Data Science* 6.2 (2024), pp. 311–342 (cit. on p. 355).

[Vac+21] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. "A dimension-free computational upper-bound for smooth optimal transport estimation". In: *Conference on Learning Theory*. PMLR. 2021, pp. 4143–4173 (cit. on p. 355).

[VM96] Lou Van Den Dries and Chris Miller. "Geometric categories and O-minimal structures". In: *Duke Mathematical Journal* 84.2 (1996), pp. 497–540 (cit. on p. 118).

[Van00] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000 (cit. on pp. 88, 89, 218).

[VR08] Ravi Varadhan and Christophe Roland. "Simple and globally convergent methods for accelerating the convergence of any EM algorithm". In: *Scandinavian Journal of Statistics* 35.2 (2008), pp. 335–353 (cit. on p. 126).

[VMK25] Christophe Vauthier, Quentin Mérigot, and Anna Korba. *Properties of Wasserstein Gradient Flows for the Sliced-Wasserstein Distance*. 2025. arXiv: 2502.06525 [stat.ML] (cit. on p. 353).

[Vay20] Titouan Vayer. *A contribution to Optimal Transport on incomparable spaces (PhD thesis)*. 2020. arXiv: 2011.04447 [stat.ML] (cit. on pp. 252, 253).

[Vay+20] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. "Fused Gromov-Wasserstein distance for structured objects". In: *Algorithms* 13.9 (2020), p. 212 (cit. on pp. 21, 37, 252, 253).

[Vay+19] Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. "Sliced gromov-wasserstein". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 252, 258).

[Via83] Jean-Philippe Vial. "Strong and weak convexity of sets and functions". In: *Mathematics of Operations Research* 8.2 (1983), pp. 231–259 (cit. on p. 78).

[Vil09] Cédric Villani. *Optimal transport : old and new*. eng. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009 (cit. on pp. 1, 2, 4, 54, 102, 165, 204, 209, 264, 290, 295, 356).

[Vil21] Cédric Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2021 (cit. on p. 1).

[VS18] Aladin Virmaux and Kevin Scaman. "Lipschitz regularity of deep neural networks: analysis and efficient estimation". In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 164, 167).

[VM19] Cinzia Viroli and Geoffrey J McLachlan. "Deep Gaussian mixture models". In: *Statistics and Computing* 29.1 (2019), pp. 43–51 (cit. on p. 126).

[VH10] Dang Hai Tran Vu and Reinhold Haeb-Umbach. "Blind speech separation employing directional statistics in an expectation maximization framework". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 241–244 (cit. on p. 126).

[Wak08] Seiichiro Wakabayashi. *Remarks on semi-algebraic functions*. Online Notes. Jan. 2008 (cit. on pp. 62, 118, 184).

[Wan+23] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. "Optimal transport for treatment effect estimation". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 5404–5418 (cit. on p. 2).

[Wan+13] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. "A linear optimal transportation framework for quantifying and visualizing variations in sets of images". In: *International journal of computer vision* 101 (2013), pp. 254–269 (cit. on p. 210).

[Wat64] Geoffrey S. Watson. "Smooth Regression Analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), pp. 359–372 (cit. on pp. 329, 336).

[WB19] J. Weed and F. Bach. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In: *Bernoulli* 25.4A (2019), pp. 2620–2648 (cit. on p. 126).

[Wen64] Robert Edwin Wengert. "A simple automatic derivative evaluation program". In: *Communications of the ACM* 7.8 (1964), pp. 463–464 (cit. on p. 126).

[Wu83] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103 (cit. on p. 126).

[Wu+19] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. Paudel, and L. Van Gool. "Sliced Wasserstein Generative Models". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 3708–3717 (cit. on pp. 1, 46, 51, 55, 57, 103, 205).

[XN22] Jiaqi Xi and Jonathan Niles-Weed. "Distributional convergence of the sliced Wasserstein process". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 13961–13973 (cit. on pp. 63, 103).

[Xu+19] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. "Gromov-wasserstein learning for graph matching and node embedding". In: *International conference on machine learning*. PMLR. 2019, pp. 6932–6941 (cit. on p. 252).

[XHM16] Ji Xu, Daniel J Hsu, and Arian Maleki. "Global analysis of expectation maximization for mixtures of two gaussians". In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on p. 126).

[XH22] Xianliang Xu and Zhongyi Huang. *Central limit theorem for the Sliced 1-Wasserstein distance and the max-Sliced 1-Wasserstein distance*. 2022. arXiv: 2205.14624 [math.ST] (cit. on pp. 63, 64, 103).

[YT89] Yinyu Ye and Edison Tse. "An extension of Karmarkar's projective algorithm for convex quadratic programming". In: *Mathematical programming* 44 (1989), pp. 157–179 (cit. on p. 187).

[Yua+20] Wentao Yuan, Benjamin Eckart, Kihwan Kim, Varun Jampani, Dieter Fox, and Jan Kautz. "Deepgmr: Learning latent gaussian mixture models for registration". In: *European conference on computer vision*. Springer. 2020, pp. 733–750 (cit. on p. 126).

[ZAC21] Marco Zaffalon, Alessandro Antonucci, and Rafael Cabañas. *Causal Expectation-Maximisation*. 2021. arXiv: 2011.02912 [cs.AI] (cit. on p. 126).

[ZP19] Yoav Zemel and Victor M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932–976 (cit. on p. 354).

[Zha+11] K Zhang, J Peters, D Janzing, and B Schölkopf. "Kernel-based Conditional Independence Test and Application in Causal Discovery". In: *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press. 2011, pp. 804–813 (cit. on p. 337).

[ZPK18] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. *Open3D: A Modern Library for 3D Data Processing*. 2018. arXiv: 1801.09847 [cs.CV] (cit. on p. 244).

[ZGD24] Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. "Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures". In: *Bernoulli* 30.2 (2024), pp. 1441–1457 (cit. on pp. 337, 338).