PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Convex optimization for data science : continuous-time Lyapunov analyses, geometry dependent matching pursuit and concentration inequalities

Soutenue par

Céline Moucer

Le 1er octobre 2024

École doctorale n°386

Ecole Doctorale de Sciences Mathématiques de Paris Centre

Spécialité

Informatique

Préparée au

Ecole Normale Supérieure

Composition du jury :

Vianney PERCHET Professeur, ENSAE	*Président du jury*
Sylvain ARLOT Professeur, Université Paris Saclay	*Examinateur*
Nelly PUSTELNIK Directeur de recherche, CNRS / ENS Lyon	*Rapporteur*
Joseph SALMON Professeur, Université de Montpellier	*Rapporteur*
Francis BACH Directeur de recherche, INRIA	*Directeur de thèse*
Adrien TAYLOR Chargé de recherche, INRIA	*Co-encadrant*

ENS | PSL

Remerciements

Il me semble essentiel de commencer ce manuscrit en remerciant toutes les personnes qui m'ont accompagnée tout au long de cette expérience ou qui ont contribué à la rendre possible.

Avant tout, je souhaite exprimer ma profonde gratitude à mes encadrants, Francis et Adrien, sans qui cette thèse n'aurait pu voir le jour. Vous m'avez offert l'opportunité de réaliser ce travail de recherche, et je mesure aujourd'hui la chance que j'ai eue de travailler à vos côtés pendant ces trois années. Votre accompagnement bienveillant m'a permis d'avancer sereinement et d'aborder nos réunions avec confiance, même lorsque 'rien ne fonctionnait'. Vous m'avez appris à explorer des horizons parfois éloignés du sujet initial, et à accepter la difficulté de certains problèmes. Adrien, je te remercie particulièrement de m'avoir initiée à l'optimisation, d'avoir partagé ta vision des méthodes du premier ordre et d'avoir été impliqué dans le détail des preuves et du code avec moi. Tes nombreuses relectures m'ont appris l'importance du mot juste et de la fluidité dans l'écriture des articles scientifiques. Francis, merci pour ta constance sans faille pendant ces trois années, tant sur le plan scientifique qu'humain. Ta vision optimiste de la recherche transforme chaque problème en jeu, à explorer de manière astucieuse à travers ses (très) nombreuses pistes. Semaine après semaine, ton subtil mélange d'encadrement et de liberté m'a toujours permis de trouver une direction. J'espère emporter une partie de vos qualités avec moi.

Je tiens également à remercier tous les personnels d'Inria Paris et de l'ENS, qui ont contribué à créer un cadre aussi propice à la recherche. En particulier, merci à Marina, Hélène et Lise-Marie pour votre soutien précieux dans les démarches administratives.

Je souhaite maintenant remercier tous les membres des équipes WILLOW et SIERRA, qui ont animé mon quotidien de thèse, à travers des discussions plus ou moins techniques à la machine à café, le rituel des Friday Beers ou des retraites d'équipe mémorables. Cet esprit d'entraide et de curiosité m'a souvent permis d'avancer et de relativiser. En particulier, je tiens à remercier Bertille, qui a été une véritable 'sœur' de thèse, grâce à notre direction de thèse commune et notre anniversaire partagé ! Nos débriefings après les réunions ont rythmé mes semaines, des anecdotes sur les peaux de kangourou aux mésaventures à vélo, en passant par la permaculture. Cette thèse n'aurait pas été la même sans Benjamin, Quentin et Oumayma, et nos aventures sur le chemin de la cantine, nos critiques des repas végétariens de l'AGRAF, et plus généralement, notre partage des hauts et des bas de la thèse. En espérant n'oublier personne, merci également à Théophile, Gaspard, Sara, Yann, Louis, Ulysse, Eloïse, ainsi qu'à Mathieu et Clément (voisins d'équipes Inria) pour tous ces moments passés au centre ou en conférence. Enfin, une mention spéciale à Baptiste et Sébastien, avec qui nous avons lancé le premier workshop PEP Talks à Louvain-la-Neuve à partir d'une discussion à ICCOPT 2022 en Pennsylvanie!

Je remercie chaleureusement tous mes amis pour leur présence indéfectible à travers des dîners, des séances d'escalade ou des vacances pendant tout ce parcours. Merci à Gwen, d'être mon amie de toutes les aventures et surtout de t'être autant investie dans ce livret de vulgarisation et de jeux sur ma thèse, à un moment où j'étais prête à y renoncer. Un grand merci à Côme, Célia et Caroline d'avoir partagé nos quotidiens de thèse que ce soit en économie de l'environnement, biophysique ou démographie, des prémices à la soutenance. Enfin, un merci tout particulier à mes amis de la Protection Civile de Paris 8&9, grâce à qui j'ai pu garder un pied dans une autre réalité et équilibrer mon quotidien.

Titouan, tu as été à mes côtés tout au long de cette troisième année de thèse et tu as partagé de nombreux doutes. Merci pour ta patience et ta compréhension pendant la rédaction du manuscrit, et surtout lors du choix de l'orientation de sortie de thèse qui m'a conduite dans une autre direction.

Pour finir, je tiens à exprimer ma profonde reconnaissance envers ma famille, qui m'a toujours soutenue. Vous m'avez transmis le goût de la persévérance et de la curiosité intellectuelle. Merci pour vos conseils avisés et votre confiance sans faille.

Résumé

L'optimisation convexe est centrale en science des données, et fournit des outils efficaces pour optimiser des fonctions convexes sur des ensembles convexes. Elle est à la base de nombreux algorithmes d'apprentissage, assurant sufisamment de robustesse pour des tâches telles que la régression, la classification ou le clustering. Par ailleurs, l'optimisation convexe permet de dériver des inégalités de concentration, qui fournissent des bornes probabilistes caractérisant comportement et les performances des algorithmes sur de grands ensembles de données. Cette thèse offre des méthodologies systématiques, utilisant des techniques d'optimisation convexe, pour étudier les modèles d'optimisation en temps continu, aborder les inégalités de concentration, ainsi qu'améliorer la compréhension et le développement des algorithmes de matching pursuit.

La première partie du manuscrit se concentre sur l'étude des garanties de convergence des équations différentielles issues des méthodes de premier ordre, en utilisant des fonctions de Lyapunov. Nous proposons une approche systématique pour trouver et vérifier des fonctions de Lyapunov pour diverses équations différentielles ordinaire et stochastiques, produisant des résultats de convergence avec moins d'hypothèses. Cette fonctionnalité est intégrée dans le package python PEPit. La deuxième partie examine les méthodes de premier ordre gloutonnes, telles que la descente de coordonnées avec la règle de Gauss-Southwell et le matching pursuit, devenues populaires pour leur tendance à proposer des solutions parcimonieuses. Nous proposons une méthode pour générer des algorithmes de matching pursuit (potentiellement régularisés) adaptés à la géométrie du problème en question, ainsi que leurs garanties de convergence. Dans la dernière partie, nous établissons une approche constructive des inégalités de concentration pour des variables aléatoires indépendantes. D'une part, nous introduisons une approche variationnelle qui généralise l'approche par fonction génératrice de moments. Cette approche est adaptée lorsque le premier moment est fixé. D'autre part, nous proposons une approche polynomiale, basée sur une série d'approximations en sommes de carrés, qui est adaptée aux conditions sur les moments d'ordre supérieur.

Mots clés : Optimisation convexe, méthodes du premier ordre, techniques assistées par ordinateur, analyse par pire-case, matching pursuit, analyse par fonctions de Lyapunov, inégalités de concentration, problème généralisé des moments.

Abstract

Convex optimization is a cornerstone of data science, providing efficient tools for optimizing convex functions over convex sets. It serves as the foundation for numerous machine learning algorithms, enabling robust solutions for tasks such as regression, classification, and clustering. It is also crucial for deriving (approximation to) concentration inequalities, which provide probabilistic bounds that help understand the behavior and performance of algorithms on large datasets. This thesis offers comprehensive and principled methodologies for studying continuous-time optimization models, approaching concentration inequalities, as well as enhancing the understanding and development of greedy algorithms.

The first part of the manuscript focuses on studying convergence guarantees of ordinary differential equations arising from first-order methods, using Lyapunov functions. We provide a systematic approach to finding and verifying Lyapunov functions for various ordinary and stochastic differential equations, yielding convergence results with fewer assumptions. This functionality is integrated into the PEPit python package. The second part investigates greedy first-order methods, such as steepest coordinate descent and matching pursuit, which has become popular due to their natural tendency to propose sparse solutions. We propose a principled approach to generating (regularized) matching pursuit adapted to problem geometry at hand, along with their refined convergence guarantees. In the last part, we introduce a constructive approach to concentration inequalities for independent random variables. We propose on the one hand a variational approach that extends the classical approach via moment generating function and is particularly suited to first-order moment assumptions. On the other hand, we propose a polynomial approach, based on a hierarchy of sums-of-squares approximations, suited to higher-order moment conditions.

Keywords : Convex optimization, first-order methods, computer-assisted techniques, worst-case analysis, matching pursuit, Lyapunov analysis, concentration inequalities, generalized problem of moments.

Résumé long en français

1. Introduction

Le machine learning consiste à entraîner des algorithmes sur des données afin de réaliser des prédictions. Selon la nature de la variable à prédire, le problème peut être une régression (pour des valeurs continues, comme l'estimation d'un prix) ou une classification (pour des catégories discrètes, comme la détection de spam). Nous nous plaçons ici dans le cadre de l'apprentissage supervisé : un modèle est appris à partir d'exemples annotés, dans le but de pouvoir généraliser sur de nouvelles données. Concrètement, on définit un risque attendu comme l'espérance d'une fonction de perte, ce qui se traduit, en pratique, par la minimisation du risque empirique sur l'échantillon disponible. L'erreur totale entre le risque attendu et le risque empirique se décompose en une erreur d'approximation (associée à la famille de modèles choisie) et en une erreur d'estimation (liée à l'utilisation d'un échantillon fini).

Pour rendre ce processus efficace, deux grands défis se posent : d'une part, le contrôle de l'erreur d'estimation, qui dépend de la quantité de données et de la complexité du modèle ; d'autre part, l'optimisation numérique du risque empirique, qui nécessite l'emploi de méthodes adaptées (comme par exemple la descente de gradient ou la descente de coordonnées). Cette thèse se concentre sur l'amélioration de la robustesse statistique de ces approches notamment via des inégalités de concentration, ainsi que sur l'efficacité d'une famille d'algorithmes appelée méthodes du premier ordre, particulièrement adaptée aux problèmes convexes.

Ce manuscrit est basé sur des publications qui ont été publiées ou soumises au cours de cette thèse:

- L'analyse des modèles d'optimisation en temps continu à l'aide des fonctions de Lyapunov, correspondant à l'article *A Systematic Approach to Lyapunov Analyses of Continuous-Time Models in Convex Optimization*, C. Moucer, A. Taylor, F. Bach, publié dans SIAM Journal on Optimization 2023, [Moucer et al., 2023] et à l'article *PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python*, B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, A. Dieuleveut, à paraître dans Mathematical Programming Computation (2024), [Goujaud et al., 2024].

- Le développement et l'étude des algorithmes gloutons adaptés à la géométrie des problèmes d'optimisation, tels que les méthodes de *matching pursuit*, correspondant à l'article *Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and LASSO*, C. Moucer, A. Taylor, F. Bach, en révision mineure pour publication dans le Journal of Machine Learning Research (2024), [Moucer et al., 2024b]

- La construction d'inégalités de concentration basées sur des approches variationnelles et

polynomiales, qui correspond à l'article *Constructive approaches to concentration inequalities with independent random variables*, C. Moucer, A. Taylor, F. Bach, *en révision à Mathematics of Operations Research.* [Moucer et al., 2024a]

2. Systematic approach to Lyapunov analyses of continuous-time models in convex optimization

L'optimisation convexe est un outil fondamental dans l'analyse numérique et joue un rôle essentiel dans la formulation de problèmes en science des données et traitement du signal. Le problème de minimisation considéré s'écrit :

$$\min_{x\in\mathbb{R}^d} f(x), \tag{2.1}$$

où $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction convexe différentiable.

Pour résoudre le problème (2.1), on utilise traditionnellement des méthodes de premier ordre telles que la descente de gradient ou le gradient accéléré de Nesterov. Beaucoup d'applications impliquent une part aléatoire et nécessitent une modélisation stochastique de la fonction. La descente de gradient stochastique (SGD) est une variante de la descente de gradient adaptée à de telles fonctions, à un coût réduit.

Ces méthodes peuvent être interprétées en temps continu. La descente de gradient déterministe est approximée par le flot de gradient donné par :

$$\dot{X}_t = -\nabla f(X_t), \tag{2.2}$$

où X_t est la solution d'une équation différentielle ordinaire. Le gradient stochastique est quant à lui décrit par une équation différentielle stochastique:

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t, \tag{2.3}$$

où B_t est un mouvement Brownien standard et $\sigma(X_t)$ caractérise les paramètres de la méthode du premier ordre.

Les flots de gradients (respectivement les méthodes du premier ordre) sont généralement étudiés via des propriétés de convergence dans le pire cas, qui sont vérifiées pour toutes les fonctions d'une certaine classe et toutes les trajectoires générées par cette certaine équation différentielle (respectivement méthode du premier ordre). Dans de nombreux cas, les approches en temps continu semblent permettre des démonstrations plus courtes, plus simples et donc plus intuitives. Elles apportent également des perspectives sur ce que l'on peut attendre des méthodes d'optimisation.

L'analyse des modèles en temps continu repose souvent sur des arguments de stabilité de Lyapunov. Une fonction de Lyapunov $V : \mathbb{R}^d \to \mathbb{R}$ se définit comme une fonction positive, décroissante le long des trajectoires d'une dynamique donnée, et nulle à l'optimum de (2.1). L'existence d'une telle fonction de Lyapunov produit directement une preuve de convergence pour les équations différentielles ordinaires (ODE). Le principal défi de l'approche de Lyapunov est de trouver une fonction appropriée qui décroît le long de toutes les trajectoires générées par une équation différentielle ordinaire (ODE) et pour toutes les fonctions d'une certaine classe.

Du point de vue d'un observateur extérieur, ces analyses sont souvent perçues comme complexes et techniques à maîtriser. Dans ce travail, nous remédions à ce problème en étendant l'approche systématique basée sur la programmation semi-définie (SDP), initialement introduite par Drori and Teboulle [2014], pour certifier la convergence des méthodes d'optimisation. Cette technique est appelée « problèmes d'estimation de performance » (PEPs). La contribution principale de ce travail est de fournir un outil pour analyser la convergence des modèles en temps

continu, en construisant des fonctions de Lyapunov adaptées à une équation différentielle ordinaire (ODE) basée sur le gradient de manière systématique, en utilisant des reformulations en SDP de petite taille. De plus, cette procédure bénéficie de propriétés de précision, ce qui signifie que la faisabilité du SDP permet de conclure qu'il existe une fonction de Lyapunov dans la famille prescrite des fonctions de Lyapunov considérées. Réciproquement, l'infaisabilité du SDP permet de conclure qu'il n'existe pas une telle fonction de Lyapunov valide dans cette famille.

Tout au long de ce travail, les fonctions f considérées sont convexes différentiables (par simplicité), potentiellement L-lisses et μ-fortement convexes.

Temps continu déterministe. Dans le cadre déterministe, nous étendons l'approche d'estimation de performance développée pour les méthodes d'optimisation du premier ordre aux flots de gradient qui proviennent d'une fonction convexe (possiblement fortement convexe). En résumé, nous générons des fonctions de Lyapunov en tant que points faisables de certaines inégalités matricielles linéaires.

Pour étudier le flot de gradient (2.2), on introduit une famille de fonctions de Lyapunov quadratiques paramétrées définies par:

$$\mathcal{V}_{a_t,c_t}(X_t,t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2,$$

où $a_t, c_t \geqslant 0$ sont des fonctions différentiables par rapport au temps, telles que la fonction $\mathcal{V}_{a_t,c_t}$ est non négative et décroissante le long du flot. L'analyse systématique repose sur la formulation des garanties de convergence sous forme de contraintes linéaires matricielles (LMI). Ces contraintes permettent d'une part de vérifier qu'une certaine fonction $\mathcal{V}_{a_t,c_t}$ est une fonction de Lyapunov, et d'autre part, de générer de nouvelles fonctions de Lyapunov.

Par ailleurs, les dynamiques d'ordre supérieur, comme celles introduites pour les méthodes accélérées de Nesterov, sont également considérées. Ces dynamiques sont modélisées par des équations du second ordre :

$$\ddot{X}_t + \beta_t\dot{X}_t + \nabla f(X_t) = 0,$$

où β_t dépend des paramètres de la classe de fonctions et la classe du premier ordre considérée.

Temps continu stochastique. Dans le cadre stochastique, les dynamiques sont modélisées par des équations différentielles stochastiques (2.3). En se basant sur la méthode développée dans le cas déterministe, nous étudions les propriétés des trajectoires générées par des équations différentielles stochastiques (SDE) qui approximent les méthodes de gradient stochastique, incluant des pas dépendants du temps et des techniques de moyennage comme celui de Polyak-Ruppert.

La vérification de fonctions de Lyapunov quadratiques à l'aide de LMI de petites tailles combinées au calcul d'Itô, permet de retrouver des résultats obtenus en temps discret, tout en offrant des analyses plus compactes en temps continu. Cette analyse de convergence met en évidence un équilibre entre l'atténuation des conditions initiales et la réduction de la variance introduite par le bruit stochastique.

Nous étudions enfin les flots accélérés stochastiques d'ordre supérieur qui approchent les méthodes accélérées stochastiques et découlent des techniques de moyennage:

$$d^2X_t + \beta_t dX_t + h_t\nabla f(X_t)dt + h_t\sqrt{\gamma\Sigma(X_t)}dB_t = 0.$$

Cette approche permet de conclure qu'il n'est pas possible d'accélérer le flot de gradient sans hypothèses supplémentaires sur le bruit $\Sigma(X_t)$ ou sur les fonctions f à minimiser.

Ce travail propose un cadre unifié et méthodique pour l'analyse des dynamiques continues en optimisation convexe. En utilisant des fonctions de Lyapunov et une reformulation sous forme de

contraintes linéaires matricielles, nous établissons des garanties de convergence pour les équations différentielles ordinaires et stochastiques. Nos résultats correspondent à ceux obtenus dans le cas discret, mais avec des analyses plus compactes. Les dynamiques d'ordre supérieur, telles que celles des méthodes accélérées, sont également étudiées, mettant en évidence des améliorations du taux de convergence dans le cas déterministe. Cette approche permet de généraliser les résultats existants tout en simplifiant leur analyse théorique et leur implémentation numérique. Parmi les travaux futurs, il serait intéressant d'élargir la famille de fonction de Lyapunov considérée ou d'étudier des problèmes d'inclusion monotone et des équations différentielles faisant appraître des dérivées d'ordre supérieur.

3. Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and LASSO

De nombreux problèmes en machine learning et en science des données prennent la forme d'une minimisation pénalisée par la norme ℓ_1 :

$$\min_{\alpha \in \mathbf{R}^d} \Big\{ F(\alpha) \; + \; H(\alpha) \; \triangleq \; f(P\alpha) \; + \; \lambda \|\alpha\|_1 \Big\}, \tag{2.4}$$

où f est une fonction lisse et fortement convexe, $P \in \mathbf{R}^{n \times d}$ et n, d désignent respectivement le nombre d'échantillons et la dimension du problème. Par exemple, pour la régression linéaire simple, on a $H(\alpha) = 0$ et $F(\alpha) = \frac{1}{2} \|P\alpha - y\|_2^2$. Si $H(\alpha) = \lambda \|\alpha\|_1$, on obtient le LASSO. Ces formulations sont fréquemment utilisées pour induire de la parcimonie et de la régularisation.

L'idée principale de ce travail est de proposer une approche systématique des méthodes de boosting, obtenues en minimisant une certaine borne supérieure quadratique par rapport à la norme ℓ_1 (issue de la lissitude de la fonction à minimiser). Cette approche conduit à une nouvelle stratégie de boosting pour des problèmes régularisés, tout en bénéficiant de propriétés de convergence (sous-)linéaire. Contrairement aux méthodes connues, l'analyse de convergence est effectuée ici sans hypothèses statistiques sur les données. Les garanties de convergence font apparaître des paramètres caractérisant à la fois la classe de fonctions et les géométries des problèmes d'optimisation (1) en $\mathbf{R}^d$ et (3) en $\mathbf{R}^n$, mais ceux-ci restent en grande partie difficiles à calculer. Pour y remédier, nous proposons des estimations *a priori* raffinées des vitesses de convergence pour des méthodes de boosting appliquées à un problème de moindres carrés particulier.

Nous développons deux approches: l'une repose sur des estimations déterministes via des relaxations SDP, l'autre sur des bornes en haute probabilité basées sur la théorie des matrices aléatoires. Il en ressort une phase de transition dans la vitesse de convergence de la descente de gradient (resp. de la descente de coordonnées), selon la valeur de (n, d). En nous appuyant sur ces résultats, nous mettons en évidence expérimentalement une phase de transition pour le gradient proximal et le *matching pursuit* régularisé dans un problème de LASSO, en fonction de la valeur de λ. Enfin, nous définissons une méthode « ultime » bénéficiant d'une convergence linéaire à la fois en régime sous-paramétré ($n \gg d$) et sur-paramétré ($n \ll d$), qui n'est toutefois pas une méthode de boosting (car elle peut en effet ajouter plus d'un atome par itération).

Une phase de transition pour la régression linéaire. Nous commençons par comparer deux algorithmes standard pour le cas simple sans régularisation de la régression linéaire :

$$\min_{\alpha \in \mathbf{R}^d} \Big\{ \tfrac{1}{2n} F(\alpha) \triangleq \|P\alpha - y\|_2^2 \Big\},$$

où $P \in \mathbf{R}^{n \times d}$. D'une part, la descente de gradient en norme ℓ_2 avec un pas fixe $\gamma = 1/L_2^F$, l'itération s'écrit:

$$\alpha_{k+1} \;=\; \alpha_k \;-\; \frac{1}{L_2^F}\nabla F(\alpha_k) \;=\; \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \left\{ F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_2^F}{2}\|\alpha - \alpha_k\|_2^2 \right\},$$

où $L_2^F = \frac{1}{n}\lambda_{\max}(P^\top P)$. La descente de gradient par coordonnées ℓ_1 met à jour une coordonnée à la fois. Considérons en particulier la règle de Gauss-Southwell avec $L_1^F = \frac{1}{n}\max_i \|P_{:,i}\|_2^2$:

$$i_k = \operatorname*{arg\,max}_{1 \le i \le d} |\nabla_i F(\alpha_k)|, \;\; \alpha_{k+1} = \alpha_k - \frac{1}{L_1^F}\nabla_{i_k} F(\alpha_k)\, e_{i_k},$$
$$\text{or} \;\; \alpha_{k+1} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \left\{ F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2}\|\alpha - \alpha_k\|_1^2 \right\},$$

Ces deux méthodes appliquées à la régression linéaire convergent linéairement dans les régimes sur et sous-paramétrés. Sous des hypothèses sur les données P, nous dérivons des estimées déterministes et stochastiques des taux de convergence de la descente de gradient et de la descente par coordonnées. Ces estimées révèlent que la complexité se dégrade lorsque $d \approx n$, puis s'améliore à nouveau lorsque $d \gg n$. Les expériences numériques confirment cette phase de transition autour de $d \simeq n$. La vitesse de convergence de ces algorithmes dépend donc du degré de sous-paramétrisation ou de sur-paramétrisation du problème considéré.

Une phase de transition pour les modèles linéaires pénalisés. Considérons maintenant le problème régularisé (2.4). Parmi les méthodes classiques, le gradient proximal ou la descente de coordonnées proximale convergent linéairement dans le régime surparamétré (F lisse et fortement convexe) et sous-linéairement sinon (F convexe et lisse). Nous introduisons une version régularisée du *matching pursuit*, construite selon la même logique de majoration lisse adaptée à la norme ℓ_1:

$$\alpha_{k+1} = \operatorname*{argmin}_{\alpha \in \mathbf{R}^d} \Big\{ \langle \nabla F(\alpha_k),\, \alpha - \alpha_k \rangle + \tfrac{L_1^F}{2}\,\|\alpha - \alpha_k\|_1^2 + \lambda\,\|\alpha\|_1 \Big\},$$

où $L_1^F = \frac{1}{n}\max_i \|P_{:,i}\|_2^2$. L'itération peut être résolue analytiquement en plusieurs étapes et constitue une procédure de boosting dans la géométrie ℓ_1. À chaque itération, une seule direction est ajoutée, tandis que les autres peuvent être modifiées. Lorsque F est fortement convexe, la méthode de *matching pursuit* régularisée converge de manière linéaire, tandis que dans le cas général, elle converge de manière sous-linéaire. Dans un régime sur-paramétré (lorsque F n'est pas fortement convexe), si la solution est suffisamment parcimonieuse, une phase (sous-)linéaire peut être observée en fonction de l'identification du support. Lorsque le paramètre λ tend vers zéro, on retrouve la régression non pénalisée et sa transition ; lorsque λ est grand, la dimension effective du problème se trouve réduite.

Nous avons proposé une approche basée sur la géométrie pour développer des algorithmes de *matching pursuit* et de descente de coordonnées. Dans le cas non régularisé, la descente de coordonnées (règle de Gauss-Southwell) peut être interprétée précisément comme un algorithme de *matching pursuit* dans une norme spécifique. Dans le cas régularisé par la norme ℓ_1, nous introduisons une version régularisée du *matching pursuit*. Nous avons dérivé des estimées des taux de convergence qui mettent en évidence un phénomène de transition (sous-linéaire ou linéaire) lorsque le rapport $\frac{d}{n}$ ou la pénalité λ varient. Parmi les directions futures, on peut envisager des variantes accélérées ou des méthodes à blocs aléatoires. Une étude du lien de la transition de phase observée avec le phénomène de *double descent* en théorie de l'apprentissage serait également une piste intéressante.

4. Constructive approaches to concentration inequalities with independent random variables

Les inégalités de concentration se sont imposées comme un outil majeur en théorie des probabilités. Elles quantifient la probabilité qu'une variable aléatoire s'écarte de sa valeur moyenne (ou d'une autre référence). Parmi les inégalités classiques, on retrouve celles de Markov (pour les déviations positives) et de Tchebychev (pour les écarts à l'espérance). En apprentissage et en statistique, où l'on suppose souvent les données indépendantes et identiquement distribuées (i.i.d.), des bornes plus fortes, comme les inégalités de Hoeffding, de Bennett ou de Bernstein, s'emploient pour analyser la généralisation des algorithmes. Par exemple, pour $(X_i)_{i=1,\dots,n}$ i.i.d. à valeurs dans $[0,1]$, l'inégalité de Hoeffding [Hoeffding, 1963] montre qu'avec une déviation $t \geq 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \;\geq\; nt\Big) \;\leq\; \exp\Big(-2n\,t^2\Big). \tag{2.5}$$

Nous proposons une méthodologie systématique pour construire et optimiser des majorants adaptés à l'étude d'inégalités de concentration pour des variables aléatoires indépendantes. L'objectif est de résoudre un problème général de moments, formulé comme

$$\begin{aligned}\rho_n &= \sup_{\forall i,\, p_i \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}^n} F(x_1,\dots,x_n)\, dp_1(x_1)\dots dp_n(x_n), \quad \text{tel que} \quad \int_{\mathcal{X}} g_i(x_i)\, dp_i(x_i) \;=\; \mu_i, \\ &= \sup_{\forall i, p_i \in \mathcal{P}(\mathcal{X}^n)} \mathbb{E}_p[F(X)], \quad \text{tel que} \quad \forall i, \mathbb{E}_{p_i} g_i(X_i) = \mu_i,\end{aligned}$$

où F représente généralement une fonction indicatrice et μ_i correspond aux moments imposés à chaque variable X_i. Dans un cadre de variables indépendantes, nous proposons diverses familles de fonctions majorantes à F permettant d'obtenir des bornes quantitatives plus fines que certaines inégalités classiques (par exemple l'inégalité de Hoeffding, Bernstein). Nous explorons deux stratégies pour approcher le problème général des moments de variables aléatoires indépendantes.

Approche basée sur des produits de fonctions. Une première approche pour majorer notre fonction consiste à utiliser des fonctions-produit de la forme $U(x) = \prod_{i=1}^n u_i(x_i)$, telle que $F(x) \leqslant \prod_{i=1}^n u_i(x_i)$ pour tout $x \in \mathcal{X}^n$. Grâce à la dualité, nous pouvons décomposer le problème en (n) sous-problèmes univariés plus simples. Dans le cas particulier de l'inégalité de Hoeffding pour des variables aléatoires i.i.d., la borne obtenue améliore légèrement celle de Hoeffding, et correspond à la loi des grands nombres lorsque le nombre de variables n tend vers l'infini. Pour des variables de moyennes différentes, nous pouvons regrouper les variables par blocs de même moyenne, ce qui améliore significativement les bornes obtenues, surtout lorsque les blocs sont très différents. Bien que cette méthode soit particulièrement adaptée à l'inégalité de Hoeffding, sa généralisation à des contraintes sur des moments d'ordre supérieur, comme celles utilisées dans les inégalités de Bernstein et Bennett, s'avère difficile.

Approche polynomiale et sommes de carrés. Face aux limites de la factorisation exponentielle pour les moments d'ordre supérieur, nous proposons une nouvelle approche basée sur l'approximation polynomiale. En majorant F par un polynome Q de degré d bien choisi, nous obtenons le problème d'optimisation suivant:

$$\sup_{p_1,\dots,p_n \in \mathcal{P}(\mathcal{X})} \int Q(x)\, dp_1 \dots dp_n \quad \text{tel que} \quad \int x_i^k\, dp_i(x_i) = \mu_i^{(k)}.$$

Ce problème peut être relaxé en un problème de sommes de carrés (SOS) et résolu efficacement par des programmes semi-définis (SDP). Nous appliquons cette méthode pour affiner les inégalités classiques de Bernstein et Bennett, en particulier pour des valeurs spécifiques de la déviation t. Cette méthode, bien que plus flexible, présente un coût computationnel plus élevé et ne garantit pas systématiquement des améliorations par rapport aux approches traditionnelles, notamment pour des inégalités comme celle de Bennett.

Des perspectives intéressantes s'ouvrent pour de futurs travaux. Tout d'abord, il serait pertinent d'identifier de nouvelles familles de fonctions produits admettant des expressions analytiques, afin de réduire la complexité numérique tout en conservant leur efficacité. Par ailleurs, l'exploration de techniques de réduction de dimension pour les programmes semi-définis associés aux sommes de carrés pourrait permettre de traiter des problèmes de plus grande taille. Enfin, l'intégration d'arguments analytiques, tels que la convexité ou les développements en série de Taylor, pourrait conduire à des bornes plus précises et à une meilleure compréhension des phénomènes en jeu. L'extension de ces méthodes à des classes de distributions plus générales, comme les distributions unimodales ou sous-gaussiennes, constituerait également une avancée significative dans ce domaine.

Conclusion et perspectives

Cette thèse a exploré trois axes de recherche majeurs : l'analyse de la convergence des méthodes d'optimisation du premier ordre, l'étude des algorithmes de type matching pursuit et l'élaboration d'inégalités de concentration. Nous avons proposé de nouvelles approches pour analyser ces problèmes, en combinant des outils d'analyse fonctionnelle, d'optimisation convexe et de probabilités. Nos résultats contribuent à une meilleure compréhension des mécanismes sous-jacents à ces algorithmes et ouvrent de nouvelles perspectives pour l'optimisation numérique.

Contents

Remerciements i

Résumé iii

Abstract iv

Résumé long en français v

Table of contents xii

Contributions and thesis outline xv

1 General introduction 1
- 1.1 General context . . . 1
  - 1.1.1 ML framework . . . 1
  - 1.1.2 Global outlines . . . 2
- 1.2 First-order methods in convex optimization . . . 3
  - 1.2.1 Basics of convex optimization . . . 3
  - 1.2.2 First-order methods . . . 6
  - 1.2.3 A continuous-time interpretation for gradient descent . . . 10
- 1.3 Computer-assisted worst-case analyses . . . 12
  - 1.3.1 Explaining example: the gradient flow . . . 12
  - 1.3.2 Discrete-time versions . . . 15
- 1.4 Concentration inequalities . . . 17
  - 1.4.1 Standard concentration inequalities . . . 17
  - 1.4.2 Concentration inequalities for the sum of independent random variables . . . 18
  - 1.4.3 Subgaussian properties . . . 20

2 A systematic approach to Lyapunov analyses of continuous-time models in convex optimization 22
- 2.1 Introduction . . . 22
  - 2.1.1 Lyapunov functions . . . 24
  - 2.1.2 Prior works . . . 24
  - 2.1.3 Contributions and organization . . . 25
  - 2.1.4 Assumptions . . . 25
- 2.2 A principled approach to Lyapunov functions for gradient flows . . . 26
  - 2.2.1 The gradient flow . . . 26

2.2.2 Accelerated gradient flows . . . 30
2.2.3 Higher-order convergence and time dilation . . . 34
2.3 SDEs for modeling SGD . . . 36
2.3.1 Lyapunov functions do not always extend to the stochastic setting . . . 37
2.3.2 Diminishing the step size is a key to success . . . 39
2.3.3 Averaging for larger step sizes . . . 40
2.4 Accelerating the gradient flow . . . 43
2.4.1 Averaging does not preserve convergence rates . . . 43
2.4.2 Diminishing step sizes do not help preserve acceleration . . . 44
2.5 Conclusion and future work . . . 45
Appendices . . . 46
2.A Proof for ODEs . . . 46
2.A.1 Proofs for Theorem 11 . . . 46
2.A.2 Proof for Corollary 3 . . . 46
2.B Proofs for SDEs . . . 47
2.B.1 Proof for Theorem 15 . . . 47
2.B.2 Proof for Theorem 17 . . . 48
2.C Strongly convex regularization . . . 48
2.D Using PEPit for computing worst-case guarantees . . . 50

3 Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and LASSO 54
3.1 Introduction . . . 55
3.1.1 Prior works . . . 56
3.1.2 Assumptions . . . 58
3.2 A transition phase for linear regression . . . 59
3.2.1 Gradient descent in the ℓ_2-geometry . . . 60
3.2.2 Gauss-Southwell coordinate descent in the ℓ_1-geometry . . . 62
3.2.3 A transition phase phenomenon: experimental results . . . 64
3.2.4 Coordinate descent is an instance of matching pursuit . . . 66
3.3 A transition phase for penalized linear models . . . 68
3.3.1 Regularized matching pursuit . . . 69
3.3.2 Convergence guarantee . . . 73
3.3.3 A transition phase depending on λ: experimental results . . . 75
3.3.4 An ultimate method adapted to the geometry of regularized models . . . 76
Appendices . . . 79
3.A Mathematical tools . . . 80
3.A.1 The η-tricks: reweighted least-squares formulations . . . 80
3.A.2 Computing the basis of a kernel . . . 81
3.B Proof for Propositions 3, 4 . . . 81
3.C Estimates for the convergence of steepest coordinate descent for least-squares . . . 81
3.C.1 SDP relaxations for μ_1^F and $\mu_{1,L}^F$ . . . 81
3.C.2 High-probability bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F: proof for Proposition 6 . . . 84
3.C.3 Proof for Corollary 7 . . . 88
3.D Matching pursuit in the gauge geometry . . . 88
3.D.1 The gauge is a norm: proof for Lemma 8 . . . 88
3.D.2 Proof for sublinear convergence of matching pursuit . . . 89
3.E Steepest coordinate descent is 'nearly' a matching pursuit algorithm . . . 89
3.F The ultimate method: an inner loop strategy . . . 90
3.F.1 Alternating randomized block coordinate descent . . . 91
3.F.2 Alternating minimization . . . 92

3.F.3 Experimental results . . . 92
3.G Constrained matching pursuit. . . . 93
3.G.1 Constrained matching pursuit . . . 93
3.G.2 Convergence guarantees . . . 96

4 Constructive approaches to concentration inequalities with independent random variables 98
4.1 A variational approach based on product-functions. . . . 102
4.1.1 The separable approach. . . . 103
4.1.2 Optimizing over $\mathcal{U}$. . . . 104
4.2 Revisiting Hoeffding's inequality. . . . 105
4.2.1 One random variable: comparison of the exact and exponential bounds . 105
4.2.2 Generalization to n independent random variables. . . . 107
4.2.3 Reconstruction extremal distributions. . . . 111
4.3 A polynomial approach based on sum-of-square decomposition. . . . 114
4.3.1 A simple linear upper bound for Hoeffding's concentration. . . . 114
4.3.2 A polynomial upper bound: a SoS approximation for polynomial moment assumptions. . . . 115
4.3.3 Applications to Bernstein's and Bennett's inequalities . . . 117
4.3.4 Higher-order polynomial approximations : a variational reformulation. . . 119
4.4 Conclusion and future works. . . . 122
Appendices . . . 123
4.A Exact formulation of generalized problem of moments for independent random variables using optimal transport: . . . 123
4.B Proofs for Hoeffding's inequality . . . 124
4.B.1 Separable approach for univariate distributions . . . 124
4.B.2 Variational approach with equal means: computing the extremal points for $\bar{X}_n$ . . . 124
4.B.3 Closed-form solution to the variational reformulation for $n = 2$ . . . 125
4.B.4 Proof for Lemma 17. . . . 127
4.C Connecting the second-order to the first-order moment: proof for Lemma 18 . . . 128

Conclusion and perspectives 129
5.1 Summary of the thesis . . . 129
5.2 Perspectives . . . 130

Bibliography 131

Contributions and thesis outline

This manuscript is based on publications that were published or submitted during the course of this thesis. This section summarizes the organization of this manuscript and the main contributions of this thesis. Chapters 2, 3 and 4 can be considered separately.

Chapter 1. This opening chapter introduces the main ideas and concepts that are addressed throughout the coming chapters. More precisely, we begin with a section on the basics of convex optimization and a description of standard first-order methods. Then, we present the performance estimation framework, a computer-assisted technique for the worst-case analysis of first-order methods (and their continuous-time limits). Finally, we focus on concentration inequalities and present the main results and assumptions appearing in this manuscript.

Chapter 2. In this chapter, we study differential equations arising from first-order methods, whose convergence guarantees are usually approached via Lyapunov functions. We provide a systematic and principled approach to finding and verifying Lyapunov functions for certain classes of ordinary and stochastic differential equations, extending the performance estimation framework to continuous-time models. We achieve convergence results comparable to those of discrete-time methods while utilizing fewer assumptions and inequalities, and we present new findings for a family of stochastic accelerated gradient flows. This chapter is associated with the publication:

- *A Systematic Approach to Lyapunov Analyses of Continuous-Time Models in Convex Optimization*, C. Moucer, A. Taylor, F. Bach. Published in SIAM Journal on Optimization 2023, [Moucer et al., 2023].

Some appendices and figures of this chapter are derived from the following publication, to which I contributed as part of a global effort, but is not included in this manuscript:

- *PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python*, B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, A. Dieuleveut. To appear in Mathematical Programming Computation (2024), [Goujaud et al., 2024].

Chapter 3. This chapter is devoted to greedy first order-methods, such as coordinate descent with the Gauss-Southwell rule or matching pursuit (or Frank-Wolfe in the constrained case). These methods have become popular in optimization and signal processing due to their natural tendency to propose sparse solutions. In this chapter, we propose a principled approach to derive and analyze (regularized) matching pursuit algorithms adapted to the geometry at hand, together with their refined convergence guarantees. Under subgaussian assumptions, we obtain

approximate values of smoothness and strong convexity parameters for least-square or LASSO problems. Using these approximations, we estimate convergence guarantees for these methods when applied to these problems. Finally, we describe a transition phenomenon in the convergence of (regularized) matching pursuit from underparameterized to overparameterized models. This chapter is associated with the publication:

- *Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and LASSO*, C. Moucer, A. Taylor, F. Bach. Minor revisions for publication in the Journal of Machine Learning Research (2024), [Moucer et al., 2024b].

Chapter 4. This chapter proposes a systematic approach to study and construct concentration inequalities for independent random variables. Specifically, we extend the generalized problem of moments to independent random variables. First, we introduce a variational approach that extends classical moment-generating functions, focusing particularly on first-order moment conditions. Second, we develop a polynomial method based on a hierarchy of sum-of-squares approximations to extend these techniques to higher-order moment conditions. Building on these results, we refine Hoeffding's, Bennett's, and Bernstein's inequalities, providing improved guarantees compared to existing concentration bounds. This chapter is associated with the publication:

- *Constructive approaches to concentration inequalities with independent random variables*, C. Moucer, A. Taylor, F. Bach. *Submitted to Mathematics of Operations Research.*

Conclusion. This chapters summarizes the main results of this manuscript and takes a step back by proposing future works.

CHAPTER 1

General introduction

Objectives

This introductory chapter presents the general context of this manuscript. We begin by discussing the machine learning framework, from which the problems addressed in the following chapters emerge. We then describe standard assumptions and results in convex optimization, computer-assisted techniques for worst-case analysis, and the basics of concentration inequalities. The goal of this introduction is to provide a brief overview of these concepts, with detailed literature reviews elaborated in the introductions of each subsequent chapter.

1.1 General context

1.1.1 ML framework

Machine learning involves training algorithms to generate predictions from data. Two key types of tasks are regression and classification. Regression problems predict continuous outputs, such as forecasting rains or estimating house values, using methods like linear regression. Classification tasks predict discrete labels, such as determining if an email is spam or recognizing handwritten digits, using algorithms like logistic regression, support vector machines, and neural networks. Both types of tasks have become fundamental in various applications, enabling systems to learn from data and improve over time.

Supervised learning is a type of machine learning where a model is trained on labeled data [Bach, 2023b, Mohri et al., 2018]. Its goal is for the model to learn the mapping from inputs to labels, enabling it to make predictions on unseen data. Let $\mathcal{W} \subset \mathbb{R}^d$ be the set of inputs, such as images, sequences of DNA, or times series, and $\mathcal{Y}$ be the set of labels, such as binary labels or regression outputs, and let $(w_i, y_i)_{i=1,\dots,n} \in \mathcal{W} \times \mathcal{Y}$ be n sampled from a distribution p on $\mathcal{W} \times \mathcal{X}$, and $f : \mathcal{W} \to \mathbb{R}$ be a function to learn. Given a loss function $l : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$, the expected risk formulates as

$$\mathcal{R}(f) = \mathbb{E}[l(Y, f(W))] = \int_{\mathcal{Y}\times\mathcal{W}} l(y, f(w))dp(w, y). \tag{1.1}$$

A classical technique for minimizing the risk consists in optimizing a surrogate for $\mathcal{R}(f)$, which

is referred to as the empirical risk minimizer:

$$\min_{\theta\in\Theta}\{\hat{\mathcal{R}}_n(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(y_i, f_\theta(w_i))\}, \tag{1.2}$$

where $\theta \in \Theta$ is associated to a family of predictors f_θ. Linear regression is a typical example for this problem, with $l(y, \tilde{y}) = \|y - \tilde{y}\|^2$ and $f_\theta(w) = \theta^\top w$. In other words, the strategy consists of picking $\hat{\theta} \in \Theta$ by minimizing the empirical risk.

More precisely, the idea is to construct families of predictors such that $\mathcal{R}(f_{\hat{\theta}})$ approximates well the optimal risk $\mathcal{R}(f_\star)$. The risk is usually decomposed into the approximation and estimation error:

$$\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_\star) = \underbrace{\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta_\star})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f_{\theta_\star}) - \mathcal{R}(f_\star)}_{\text{approximation error}},$$

where $f_{\theta_\star} \in \arg\min_{\theta\in\Theta} \mathcal{R}(f_\theta)$. By definition of $f_\star$, the approximation error is always non-negative and entirely depends on the approximation properties of the family $\{f_\theta, \theta \in \Theta\}$. The estimation error is always non-negative as well, and can be upper bounded by the empirical risk minimizer,

$$\begin{aligned}
\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta_\star}) &= \mathcal{R}(f_{\hat{\theta}}) - \hat{\mathcal{R}}_n(f_{\hat{\theta}}) + \hat{\mathcal{R}}_n(f_{\hat{\theta}}) - \hat{\mathcal{R}}_n(f_{\theta_\star}) + \hat{\mathcal{R}}_n(f_{\theta_\star}) - \mathcal{R}(f_{\theta_\star}),\\
&\leqslant 2\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)| + \underbrace{\hat{\mathcal{R}}_n(f_{\hat{\theta}}) - \hat{\mathcal{R}}_n(f_{\theta_\star})}_{\leqslant 0},\\
&\leqslant 2\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)|
\end{aligned} \tag{1.3}$$

where the first term is usually referred to as a uniform deviation and second term $\hat{\mathcal{R}}_n(f_{\hat{\theta}}) - \hat{\mathcal{R}}_n(f_{\theta_\star})$ as the empirical optimization error.

Therefore, minimizing such a surrogate raises both statistical and optimization challenges. On the one hand, controlling $\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)|$ is achieved using statistical properties, and typically relies on concentration inequalities. On the other hand, minimizing the empirical risk $\hat{\mathcal{R}}_n$ in a reasonable amount of time require efficient optimization algorithms. Depending on the properties of the loss function $l(\cdot)$ and the family of predictors, such as convexity, various iterative methods have been developed to optimize such functions in discrete timesteps, along with their a priori convergence guarantees. Both the optimization and statistical part may contain advanced technical arguments. For instance, optimization methods often face high-dimensionality challenges, which arise when dealing with large datasets n or high-dimensional features d, such as genomic datasets or images. Additionally, the uniform deviation usually grows with the size of the family of predictors $\{f_\theta, \theta \in \Theta\}$ under consideration, but must be large enough to ensure a low approximation error.

1.1.2 Global outlines

This thesis focuses on both first-order methods in convex optimization and concentration inequalities in probability theory.

Section 1.2 introduces fundamental concepts of convex optimization and key examples of first-order methods discussed in Chapter 2 and 3. Specifically, Chapter 2 studies first-order methods via their continuous-time limits, while chapter 3 presents a systematic view for generating (regularized) matching pursuit algorithms.

Section 1.3 establishes the foundations of *performance estimation problems*, a computer-assisted technique for analyzing worst-case convergence guarantees of optimization methods. This section serves as a concise tutorial preceding Chapter 2, which extends the performance

estimation framework to continuous-time models. This section outlines how the search for worst-case convergence guarantees can be formulated as optimization problems. Chapter 4 apply similar principles by framing the search for concentration inequalities as optimization problems. In both instances, semidefinite programming plays a central role in reformulating or approximating these optimization problems.

Section 1.4 is dedicated to concentration inequalities, which are relevant when designing the family of predictors for controlling the uniform deviation. These concepts are central in Chapter 3 when approximating convergence guarantees of matching pursuit algorithms. Chapter 4 is devoted to refining standard concentration inequalities for independent random variables, from an optimization viewpoint.

Here is a summary of the topics covered in each chapter:

Topics	**Chapter 2**	**Chapter 3**	**Chapter 4**
First-order methods	✓	✓	
Worst-case analysis	✓		✓
Concentration inequalities		✓	✓

1.2 First-order methods in convex optimization

Many problems emerging from machine learning or statistics can be formulated as convex optimization problems. Among them, linear and logistic regression, support-vector machine or the LASSO regression have been widely studied and can be formulated as unconstrained convex optimization problems:

$$\min_{x \in \mathbb{R}^d} f(x), \tag{1.4}$$

where the objective function f is a convex function and d is the dimension. Under this assumption, local minima are global ones. Throughout this thesis, functions are assumed to be differentiable (mostly for simplicity). We assume in addition that the problem is well-posed, that is that there exists $x_\star \in \mathbb{R}^d$ an optimal point for f such that $f(x_\star)$ is a solution to (1.4).

We focus on first-order methods for solving (1.4), given that we have access to the gradient (we will sometimes assume that we have access to the proximal operator). Gradient descent is the simplest example of first-order methods and can be traced back to Augustin-Louis Cauchy [Cauchy, 1836]. Given a starting point $x_0 \in \mathbb{R}^d$,

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k), \tag{1.5}$$

where $\gamma_k \in \mathbb{R}$ is the step size. In other words, this method moves proportionally to the gradient of f at x_k, essentially following direction indicated by the first-order Taylor expansion. It is natural to wonder under which conditions on f and γ_k gradient descent converges to an optimal point $x_\star$, together with its associated convergence rate. Furthermore, as we will see, this method has given rise to a wide family of first-order methods tailored to different scenarios.

In this section, we outline convex assumptions in modern convex optimization, that ensures, for example, non-asymptotic convergence guarantees for gradient descent. Additionally, we introduce selected first-order methods appearing in Chapter 2 and 3. Recently, first-order methods have been studied via their continuous-time limits. We provide the vanilla example of the gradient flow as limiting differential equation for gradient descent.

1.2.1 Basics of convex optimization

We start this section by considering basic regularity assumptions in convex optimization, and the resulting convergence guarantees for gradient descent, that will be our running example.

These definitions and results are inspired from Nesterov [2004], Nemirovsky and Yudin [1983], Polyak [1987]. A differentiable function f is convex if and only if

$$\forall x, y \in \mathbb{R}^d, f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle.$$

If the function is twice differentiable, its Hessian $\nabla^2 f$ is positive semi-definite. Among others, smoothness and strong convexity are standard assumptions to quantify convergence guarantees.

Smooth functions. Smoothness ensures the function does not grow too fast.

Definition 1. *A function f is L-smooth if its gradient is L-Lipschitz, that is for any $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

If in addition, f is convex, then for any $x, y \in \mathbb{R}^d$, one has,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2.$$

By definition of smoothness, a L-smooth function f is differentiable. If f is twice differentiable, its Hessian $\nabla^2 f \preccurlyeq LI$. Under this assumption, gradient descent converges for fixed step sizes, as stated in Theorem 1 that is taken from Bansal and Gupta [2019, Theorem 3.3].

Theorem 1. *Let f be convex and L-smooth on $\mathbb{R}^d$. Then, gradient descent (1.5) with fixed step sizes $\gamma = \frac{1}{L}$ verifies for $k \geqslant 1$:*

$$f(x_k) - f_\star \leqslant \frac{L\|x_0 - x_\star\|^2}{2k}.$$

Proof. We introduce the potential $\mathcal{V}_k = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2$. If we show that $\mathcal{V}_{k+1} \leqslant \mathcal{V}_k$, we conclude the expected result:

$$k(f(x_k) - f_\star) \leqslant \mathcal{V}_k \leqslant \cdots \leqslant \mathcal{V}_0 = \frac{L}{2}\|x_0 - x_\star\|^2.$$

Let us prove that $\mathcal{V}_k$ is nonincreasing. We have:

$$\begin{aligned}
\mathcal{V}_{k+1} - \mathcal{V}_k &= (k+1)(f(x_{k+1}) - f_\star) + \frac{L}{2}\|x_{k+1} - x_\star\|^2 - k(f(x_k) - f_\star) - \frac{L}{2}\|x_k - x_\star\|^2, \\
&= (k+1)(f(x_{k+1}) - f(x_k)) + f(x_k) - f_\star + \frac{L}{2}\left(\|x_{k+1} - x_\star\|^2 - \|x_k - x_\star\|^2\right).
\end{aligned}$$

Let us handle these three terms differently. First, by definition of smoothness, and for $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, we establish the descent lemma,

$$f(x_{k+1}) - f(x_k) \leqslant -\frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Second, by convexity of f in $(x_k, x_\star)$, we have:

$$f(x_k) - f_\star \leqslant \langle \nabla f(x_k), x_k - x_\star \rangle.$$

Finally, the last term can be handled using the fact that $\|\cdot\|$ is a Euclidean norm:

$$\begin{aligned}
\frac{1}{2}\left(\|x_{k+1} - x_\star\|^2 - \|x_k - x_\star\|^2\right) &= \langle x_{k+1} - x_k, x_k - x_\star \rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2, \\
&= \frac{1}{L}\langle \nabla f(x_k), x_\star - x_k \rangle + \frac{1}{2L^2}\|\nabla f(x_k)\|^2.
\end{aligned}$$

From that, we conclude that:

$$\mathcal{V}_{k+1} - \mathcal{V}_k \leqslant -\frac{k}{2L}\|\nabla f(x_k)\|^2 \leqslant 0.$$

□

Theorem 1 proves sublinear convergence of gradient descent for smooth convex functions. Additional properties, such as strong convexity, allow to improve convergence.

Strong convexity assumption. Strong convexity, recalled in the following definition, ensures the function is not too flat.

Definition 2. *A convex differentiable function f is μ-strongly convex if it verifies,*

$$\forall x, y \in \mathbb{R}^d, \ \|\nabla f(x) - \nabla f(y)\| \geqslant \mu\|x - y\|.$$

Equivalently, for any $x, y \in \mathbb{R}^d$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2.$$

The class of (possibly) L-smooth μ-strongly convex functions is denoted $\mathcal{F}_{\mu,L}$.

A consequence of strong convexity is the unicity of the minimum $x_\star$ in (1.4). If in addition, f is twice differentiable, $\nabla^2 f(x) \succcurlyeq \mu I$. Theorem 2 proves linear convergence for gradient descent for smooth strongly convex functions, and is inspired from Nesterov [2004, Theorem 2.1.14], but other results can be found in Polyak [1987].

Theorem 2. *Let $f \in \mathcal{F}_{\mu,L}$. Then, gradient descent (1.5) with fixed step size $\gamma = \frac{2}{\mu+L}$ satisfies for $k \geqslant 1$:*

$$f(x_k) - f_\star \leqslant \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x_0 - x_\star\|^2 \leqslant \frac{L}{\mu}\left(\frac{L-\mu}{L+\mu}\right)^{2k}(f(x_0) - f_\star).$$

Proof. Let us start by considering the convergence in distance:

$$\|x_{k+1} - x_\star\|^2 = \|x_k - x_\star\|^2 + \gamma^2\|\nabla f(x_k)\|^2 - 2\gamma\langle \nabla f(x_k), x_k - x_\star\rangle.$$

We use a characterization of smooth and strongly convex function established in Nesterov [2004, Theorem 2.1.11], for all $x, y \in \mathbb{R}^d$,

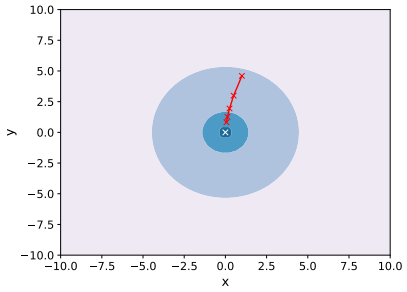
$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

We conclude that,

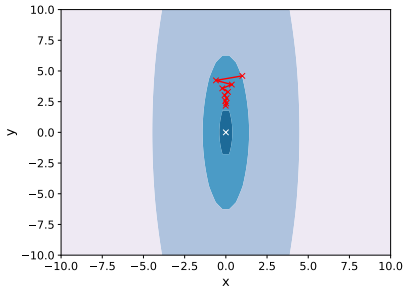
$$\|x_k - x_\star\|^2 \leqslant \left(1 - \frac{2\gamma\mu}{\mu + L}\right)\|x_k - x_\star\|^2 + \gamma\left(\gamma - \frac{2}{\mu + L}\right)\|\nabla f(x_k)\|^2.$$

The results is obtained by evaluating in $\gamma = \frac{2}{\mu+L}$ and by the characterization of smoothness (for all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|$) and strong-convexity (for all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \geqslant \mu\|x - y\|$). □

Strong convexity offers easier proofs since potentials or Lyapunov structure are constant with respect to the iteration number k. Theorem 2 provides a linear convergence guarantee for gradient descent with fixed step size, that let a condition number $\kappa = \frac{\mu}{L}$ appear. This quantity naturally describes the hardness of minimizing certain functions with first-order methods.



(a) Large κ



(b) Small κ

Figure 1.1: Intuition of the influence of the condition number on how difficult a smooth and strongly convex function is to optimize. In colors, the levels lines of the function, the white dot is the optimum, and the red lines correspond to gradient descent. On the left: $\kappa = 0.7$. On the right: $\kappa = 0.05$.

In Figure 1.1, we compare the trajectory followed by gradient descent depending on the condition number. When the condition number is large $\kappa = 0.7$, the gradient step provides a direction roughly towards the optimum. On the contrary, when the condition number is small $\kappa = 0.05$, the gradient step may point to a different direction. Then, gradient descent may oscillate (due to too large steps) before converging.

We have provided basic convexity assumptions for the function to optimize, together with convergence guarantees for gradient descent. Gradient descent has given birth to a large variety of first-order methods, that we briefly review in the next section.

1.2.2 First-order methods

First-order methods are the algorithms of choice for many modern optimization problems. Nemirovsky and Yudin [1983] introduced the notion of first-order black-box oracles, that compute the next iteration by gathering information from the past iterates and gradients. First-order methods have gained significant popularity in machine learning and data science due to their low cost per iteration and the relatively low accuracy requirements of the problems. We provide a brief presentation of first-order methods appearing in this thesis. Continuous-time limits of accelerated methods and stochastic gradient methods are studied in Chapter 2. Chapter 3 focuses on greedy methods (and thereby, coordinate descent techniques) from the the perspective of a proximal minimization technique.

An accelerated gradient method. It is natural to wonder if there exists algorithms improving convergence guarantees for gradient descent in Theorems 1 and 2 under the same assumptions. An accelerated versions is due to Nesterov [1983] for smooth convex functions, improving the worst-case convergence to $O(\frac{1}{k^2})$ compared to the $O(\frac{1}{k})$ complexity of gradient descent. In its simplest form, Nesterov's method takes the form, for $\lambda_0 = 0$, $\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$ and $\alpha_k = \frac{1-\lambda_k}{\lambda_{k-1}}$,

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ x_{k+1} &= (1-\alpha_k)y_{k+1} + \alpha_k y_k. \end{aligned} \tag{1.6}$$

Compared to gradient descent, Nesterov's accelerated method takes previous iterates into account through the so-called *momentum parameter* α_k. This momentum parameter is at the core of accelerated methods. Theorem 3 provides a convergence guarantee for Nesterov accelerated method (1.6).

Theorem 3. *[Nesterov, 1983] Let $f \in \mathcal{F}_{0,L}$. For $k \in \mathrm{N}$, the iterates defined in (1.6) verify:*

$$f(x_k) - f_\star \leqslant \frac{2L\|x_0 - x_\star\|^2}{k^2}.$$

We refer the interested reader to d'Aspremont et al. [2021] for variants of accelerated gradient methods. The convergence guarantee $O(\frac{1}{k^2})$ is nearly optimal: in general, a black-box oracle cannot beat this convergence guarantee for smooth convex functions. We refer to Nesterov [2004] for an overview of lower bounds of black-box oracles, and to Kim and Jeffrey [2016] for an optimized gradient method, that was later proven to match exactly this lower bound [Drori, 2017].

When it comes to strongly convex functions, it is possible to derive accelerated gradient methods with constant parameters:

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ x_{k+1} &= (1-\alpha)y_{k+1} + \alpha y_k, \end{aligned} \tag{1.7}$$

with $\alpha = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$, $\kappa = \frac{\mu}{L}$ and $x_1 = y_1$. This accelerated technique shares similarities with the Heavy-ball acceleration method for quadratics, as introduced by Polyak [1964]. We provide its convergence guarantee in Theorem 4, whose convergence proofs can be derived from a Lyapunov or potential perspective for example in Bansal and Gupta [2019, Section 5.5].

Theorem 4. *Let $f \in \mathcal{F}_{\mu,L}$. The iterates for the accelerated moment method (1.7) verify for $k \geqslant 1$:*

$$f(x_k) - f_\star \leqslant \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{\mu + L}{2}\|x_0 - x_\star\|^2.$$

Again, the convergence guarantee of Theorem 4 for smooth strongly convex functions is linear. Variants, such as the triple momentum method [Van Scoy et al., 2018] (TMM) or the information-theoretic exact method [Taylor and Drori, 2022] (ITEM) let varying momentum appear and improve convergence up to a constant. ITEM and TMM were proven to match exactly the lower bound [Drori and Taylor, 2022].

Stochastic gradient descent. Gradient descent (1.5) or accelerated methods (1.6) and (1.7) require a full call for gradient at each iteration, which may be expansive to compute (indeed, it requires computing d coordinates). In many machine learning cases however, computing stochastic estimates of the gradient may be more convenient. Stochastic gradient descent (SGD) was introduced by Robbins [1951] and performs a step in such a stochastic direction at each iteration:

$$x_{x+1} - x_k - \gamma_k \nabla \bar{f}(x_k, \xi_{i_k}), \tag{1.8}$$

where $\gamma_k > 0$ and ξ_{i_k} are uniformly drawn in $(\xi_1, \dots, \xi_n)$ and where $\nabla \bar{f}(x_k, \xi_{i_k})$ is an unbiased estimate of the full gradient $\nabla f(x_k)$, that is $\mathbb{E}_\xi[\nabla \bar{f}(x_k, \xi)] = \nabla f(x_k)$. In supervised machine learning, the function to optimize is typically the empirical risk minimizer $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ as defined in (1.2). One could for example choose the stochastic estimate $\nabla f_{i_k}(x)$, where i_k is drawn uniformly at random in $\{1, \dots, n\}$.

Different convergence results can be obtained, depending on the control on the gradient and the regularity of the function under consideration [Bottou et al., 2018, Bach and Moulines, 2011, Schmidt and Le Roux, 2013]. The following lemma is an example of typical guarantee for stochastic gradient descent.

Lemma 1. *Let $f \in \mathcal{F}_{\mu,L}$ and let x_k be generated in (1.8). We assume in addition that estimates to the gradient are unbiased $\mathbb{E}_{\xi_{i_k}}[\nabla \bar{f}(x_k, \xi_{i_k})] = \nabla f(x_k)$, a control on the gradients*

$\mathbb{E}[\|\nabla \bar{f}(x_k)\|^2] \leqslant \sigma^2 + \|\nabla f(x_k)\|^2$ *and fixed step sizes* $\gamma \leqslant \frac{1}{L}$. *Then, it holds for* $k \geqslant 1$,

$$\mathbb{E}_{\xi_{i_k}}[f(x_k)] - f_\star \leqslant \frac{\gamma\sigma^2 L}{2\mu} + (1-\gamma\mu)^k \left(f(x_0) - f_\star - \frac{\gamma\sigma^2 L}{2\mu} \right).$$

Proof. By smoothness and convexity of f, the iterates verify:

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\leqslant \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2, \\
&= -\gamma_k \langle \nabla f(x_k), \nabla \bar{f}(x_k, \xi_{i_k}) \rangle + \gamma_k^2 \frac{L}{2} \|\nabla \bar{f}(x_k, \xi_{i_k})\|^2, \\
\mathbb{E}_{\xi_{i_k}}[f(x_{k+1})] - f(x_k) &\leqslant -\gamma_k \|\nabla f(x_k)\|^2 + \gamma_k^2 \frac{L}{2} \mathbb{E}_{\xi_{i_k}}[\|\nabla \bar{f}(x_k, \xi_{i_k})\|^2], \\
&\leqslant -\left(\gamma_k + \gamma_k^2 \frac{L}{2} \right) \|\nabla f(x_k)\|^2 + \gamma_k^2 \frac{L}{2} \sigma^2, \quad \text{(control over the gradients)}.
\end{aligned}$$

By strong convexity and using the fact that the step size is constant $\gamma_k = \gamma$, it holds that:

$$\mathbb{E}_{\xi_{i_k}}[f(x_{k+1})] - f(x_k) \leqslant -2\mu\gamma \left(1 - \frac{L\gamma}{2} \right) (f(x_k) - f_\star) + \frac{L\gamma^2}{2}\sigma^2.$$

Since $\gamma \leqslant \frac{1}{L}$, we have $\mathbb{E}_{\xi_{i_k}}[f(x_{k+1})] - f(x_k) \leqslant -\mu\gamma(f(x_k) - f_\star) + \frac{L\gamma^2}{2}\sigma^2$, from which we conclude by subtracting $f_\star + \frac{L\gamma}{2\mu}\sigma^2$. □

Lemma 1 illustrates the fact that SGD converges at a a certain precision level for fixed step sizes (in a certain range). The larger the step size, the better the convergence but the larger the bias.

Numerous improvements and analyses can be found in the literature. On the one hand, selecting varying step sizes can lead to unbiased convergence guarantees, as suggested by the proof for Lemma 1. We refer to the Robbins-Monroe condition Robbins [1951] for general conditions on step sizes ($\sum_\infty \gamma_k = \infty$ and $\sum_\infty \gamma_k^2 < \infty$) and to Bottou et al. [2018], Nemirovski et al. [2009], Bach and Moulines [2011], Dieuleveut and Bach [2016] for refined choices of step sizes. On the other hand, averaging over the trajectory (the so-called Polyak-Ruppert averaging) instead of the sequence (x_k) itself [Ruppert, 1988, Polyak and Juditsky, 1992] can significantly improve convergence [Polyak and Juditsky, 1992, Bach and Moulines, 2011, Nemirovski et al., 2009] and achieves the nearly optimal complexity for stochastic first-order oracles [Agarwal et al., 2009]. Finally, techniques including SAGA or SVRG have been widely studied to reduce the variance by taking advantage of past gradients (see SAGA [Defazio et al., 2014] or SVRG [Johnson and Zhang, 2013]).

Coordinate descent. We have seen that SGD relied on a cheap stochastic estimate of the full gradient particularly suited to objective functions that are expressed as the sum of functions. Coordinate descent methods were developed to provide cheaper approximate to $\nabla f(x_k)$ by calling only for the gradient at a coordinate (or blocks of coordinates):

$$x_{k+1} = x_k - \frac{1}{L} \nabla_{i_k} f(x_k) e_{i_k}, \tag{1.9}$$

where $i_k \in \{1, \dots, d\}$ and e_{i_k} is a vector with a 1 at position i_k, and 0 otherwise. Randomized coordinate descent corresponds to coordinates i_k chosen at random.

By construction, coordinate descent techniques benefit from both a low-cost per iteration and sparse representations of the solution. These techniques have gained interest since the seminal work of Nesterov [2012b], from which the following convergence theorem is extracted.

Theorem 5. *[Nesterov, 2012b, Theorem 2] Let $f \in \mathcal{F}_{\mu,L}$ and x_k be generated by randomized coordinate descent (1.9) with i_k taken uniformly at random in $\{1, \dots, d\}$ at each iterates. Then for $k \geqslant 1$,*

$$\mathbb{E}[f(x_{k+1})] - f_\star \leqslant \left(1 - \frac{\mu}{Ld}\right)(f(x_k) - f_\star).$$

Proof. Applying smoothness to f in (x_{k+1}, x_k), we have:

$$\begin{aligned}
f(x_{k+1}) &\leqslant f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2, \\
&= f(x_k) - \frac{1}{L}(\nabla_{i_k} f(x_k))^2 + \frac{1}{2L}(\nabla_{i_k} f(x_k))^2, \\
&= f(x_k) - \frac{1}{2L}(\nabla_{i_k} f(x_k))^2, \\
\mathbb{E}_{i_k}[f(x_{k+1})] &\leqslant f(x_k) - \mathbb{E}_{i_k}[\frac{1}{2L}(\nabla_{i_k} f(x_k))^2], \\
&= f(x_k) - \frac{1}{d}\sum_{i=1}^{d} \frac{1}{2L}(\nabla_i f(x_k))^2, \text{ by uniform sampling}, \\
&= f(x_k) - \frac{1}{2dL}\|\nabla f(x_k)\|^2, \\
&\leqslant f(x_k) - \frac{\mu}{dL}(f(x_k) - f_\star), \text{by strong convexity}.
\end{aligned}$$

We conclude by subtracting $f_\star$. □

Compared to gradient descent convergence rate from Theorem 2, randomized coordinate descent benefits from a cheap call per iteration but suffers a factor $\frac{1}{d}$ in the linear convergence rate. The larger the dimension, the worse the convergence speed. Likewise, randomized coordinate descent has been extended to (block)-composite minimization [Richtárik and Takáč, 2014, Nesterov, 2012b, Fercoq and Richtárik, 2015] and benefits from accelerated (proximal) versions [Fercoq and Richtárik, 2015, Nesterov, 2012b, Lee and Sidford, 2013] with the same dependence in the dimension.

Other deterministic rules were developed, such as coordinate-dependent step sizes (that let local Lipschitz constants appear) [Nesterov, 2012b, Section 1], cyclic coordinate descent [Beck and Tetruashvili, 2013] or steepest coordinate descent [Nutini et al., 2018b]. Let us provide a convergence guarantee for steepest coordinate descent.

Theorem 6. *Let $f \in \mathcal{F}_{\mu,L}$ and x_k be generated by steepest coordinate descent (1.9) with $i_k = \arg\max_{i=1,\dots,d} |\nabla_i f(x_k)|$. Then for $k \geqslant 1$,*

$$f(x_k) - f_\star \leqslant \left(1 - \frac{\mu}{Ld}\right)(f(x_k) - f_\star).$$

Proof. The proof follows from the proof for Theorem 5 and $\max_{i=1,\dots,d} |\nabla_i f(x_k)|^2 \leqslant \frac{1}{d}\|\nabla f(x_k)\|^2$ by definition. □

Without further refinements on the choice for L, μ, steepest and randomized coordinate descent converges at the exact same rate. Yet, computing the $\max_{i=1,\dots,d} |\nabla_i f(x_k)|$ has a cost in $O(d)$, which is comparable to gradient descent. Nutini et al. [2018b] have refined the analysis for steepest coordinate descent, improving the convergence guarantee over randomized coordinate descent. As detailed in Nutini et al. [2018b, Appendix 2], steepest coordinate descent is particularly suited to functions f for which the computation for $\max_{i=1,\dots,d} |\nabla_i f(x_k)|$ is efficient, such as for sparse least-squares, logistic regression or graph-based label propagation.

Proximal gradient for composite minimization. Many problems in machine learning takes the following form:

$$\min_{x\in\mathbb{R}^d}\{g(x) = f(x) + \lambda h(x)\}, \tag{1.10}$$

where f is L-smooth (possibly strongly) convex and differentiable on $\mathbb{R}^d$, $\lambda > 0$, but where h is a convex function which lacks regularity assumptions. We assume that h is closed, convex and proper. The well-known LASSO introduced by Tibshirani [1996] indeed takes the form $h(x) = \|x\|_1$ and $f(x) = \|Wx - x\|_2$, where (w_i, y_i) are data under consideration.

The proximal operator $\mathrm{prox}_h : \mathbb{R}^d \to \mathbb{R}$ is defined by:

$$\mathrm{prox}_h(v) = \arg\min_x h(x) + \frac{1}{2}\|x - v\|^2, \tag{1.11}$$

where $\|\cdot\|$ is a usual euclidean norm. It can be seen as a counterpart of the gradient oracle for nonsmooth functions. Typically, given a composite minimization problem (1.10), the idea is to construct an algorithm that call for the gradient in f and to the proximal operator in h. The proximal gradient method performs exactly these two calls. Given a starting point $x_0 \in \mathbb{R}^d$, the iterates are obtained by minimizing the smooth-upper bound applied to f:

$$g(x) \leqslant f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|^2 + \lambda\|x\|_1.$$

This defines the proximal gradient method, for $x_0 \in \mathbb{R}^d$,

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d} \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|^2 + \lambda\|x\|_1. \tag{1.12}$$

Theorem 7 provides an example of convergence guarantee for proximal gradient descent that can be found in [Beck and Teboulle, 2009, Theorem 3.1].

Theorem 7. *Let $f \in \mathcal{F}_{0,L}$, h be closed, convex and proper and x_k be the sequence generated in (1.12). Then, for $k \geqslant 1$,*

$$g(x_k) - g_\star \leqslant \frac{L\|x_0 - x_\star\|^2}{2k}.$$

As gradient descent, proximal gradient descent converges sublinearly for smooth convex objectives. Among others, we refer to the works of Taylor et al. [2018b, Theorem 2.1], [Polyak, 1987, Section 2] or Bredies and Lorenz [2008] for linear convergence guarantees for smooth strongly convex objectives, to Ryu and Boyd [2016, Section 5.1] or Tseng [2000, Theorem 3.4] for convergence proofs in the case of monotone operator. Accelerated versions can be found for example in Beck and Teboulle [2009, Section 4.], Tseng [2008] or Nesterov [2013]. More details and references can be found in the monograph [Combettes and Pesquet, 2011].

1.2.3 A continuous-time interpretation for gradient descent

First-order methods often rely on simple ideas while their convergence analyses often look more complicated than what they should. The convergence analysis of first-order methods trough the viewpoint of differential equations was initiated by Polyak [1964] for the Polyak's heavy ball method, and is part of a general quest for simple interpretation and intuition of these methods.

Interplay between gradient descent and the gradient flow. We first explore the connection between gradient descent and its limiting differential equation. Recall the definition of gradient descent starting from $x_0 \in \mathbb{R}^d$ with a step size $\gamma > 0$,

$$x_{k+1} = x_k - \gamma\nabla f(x_k).$$

Consider that this trajectory in sampled from a continuous trajectory X_t at time $t_k = k\gamma$ such that $X_{t_k} = x_k$. Then, gradient descent can be formulated as

$$\frac{1}{\gamma}\left(X_{t_k+\gamma} - X_{t_k}\right) = -\nabla f(X_{t_k}).$$

As $\gamma \to 0$, the limiting ordinary differential equation (ODE) corresponds exactly to the gradient flow:

$$\dot{X}_t = -\nabla f(X_t), \;\; X_0 = x_0 \in \mathbb{R}^d, \tag{1.13}$$

with the notation $\dot{X}_t = \frac{d}{dt}X_t$. It is relatively straightforward to observe that minimizing a differentiable function is equivalent to finding a stationary point to the gradient flow. Reversely, gradient descent can be seen as an integration scheme for the gradient flow, namely explicit Euler [Gautschi, 2012, Section 5.6]. Similarly, implicit Euler (a.k.a. the proximal point) corresponds to a discretized version of the gradient flow:

$$x_{k+1} = x_k - \gamma \nabla f(x_{k+1}).$$

Deriving continuous-time versions of first-order methods leads to a unique ordinary differential equation, whereas wide families of integration schemes emerge from the gradient flow. A connection between optimization and integration methods has been highlighted by Scieur et al. [2017], demonstrating that optimization methods can be recovered as a discretization of the gradient flow for quadratics. Figure 1.2 provides an example of the trajectories of gradient descent with discrete time steps compared to the trajectory of the gradient flow in continuous-time.

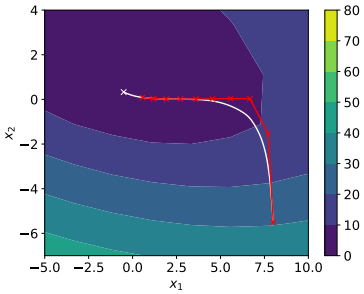


Figure 1.2: Comparison of the trajectories of gradient descent (in red) and of the gradient flow (in white), applied on a logistic regression problem.

When it comes to analyzing the gradient flow, proofs are generally much shorter. Let us first consider a general setting, where f is differentiable. The following sequence verifies:

$$\frac{d}{dt}f(X_t) = \langle \dot{X}_t, \nabla f(X_t)\rangle = -\|\nabla f(X_t)\|^2 \leqslant 0.$$

The trajectory X_t decays in function values, but the convergence rate cannot be specified. In Lemma 2, we improve this convergence guarantee by assuming that f is a strongly convex function.

Lemma 2. *Let $f \in \mathcal{F}_\mu^\infty$ and X_t be generated by the gradient flow (1.13). For any $t \geqslant 0$,*

$$f(X_t) - f_\star \leqslant e^{-2\mu t}\left(f(X_0) - f_\star\right).$$

Proof. First, it holds that $\frac{d}{dt}(f(X_t) - f_\star) \leqslant -\|\nabla f(X_t)\|^2$ By strong convexity, $-\|\nabla f(X_t)\|^2 \leqslant -2\mu\left(f(X_t) - f_\star\right)$. The result follows from integrating between 0 and t. □

The gradient flow originating from strongly convex functions converges to the optimum $x_\star$

at a linear convergence rate $e^{-2\mu}$. This convergence rate is to be compared to the convergence guarantee of gradient descent originating from L-smooth μ-strongly convex functions, that we recall from Theorem 2 $f(x_k) - f_\star \leqslant \frac{L}{\mu}\left(\frac{L-\mu}{L\mu}\right)^{2k}(f(x_0) - f_\star)$. The smoothness assumption does not play a role in the convergence of the gradient flow, while it was at the core of convergence guarantees for first-order methods.

Recently, Su et al. [2016] paved the way to continuous-time interpretation of accelerated, by interpreting Nesterov's accelerated method as a second-order differential equation. When it comes to strongly convex functions, Shi et al. [2022] refined the analysis of Nesterov's accelerated method and Polyak's heavy ball by introducing high-resolution differential equations that incorporate $O(\sqrt{\gamma})$-terms in the limiting process. Finally, stochastic first-order methods, whose convergence proofs may be challenging (see Lemma 1), were analyzed thanks to approximating stochastic differential equations [Li et al., 2017, 2019, Orvieto and Lucchi, 2020].

We have introduced several first-order methods for solving convex optimization problems relevant to certain machine learning problems, along with their convergence guarantees. Specifically, we have presented variants of gradient descent adapted to diverse challenges such as minimal assumptions on the optimization function (including accelerated methods for smooth convex functions), high-dimensional scenarios (such as stochastic gradient descent and coordinate gradient descent) and to the proximal oracle (proximal gradient descent).

1.3 Computer-assisted worst-case analyses

Optimization algorithms are often approached via worst-case convergence rates. Their analyses mostly consist in combining inequalities characterizing the class of functions and the optimization method. Deriving such proofs often relies on tedious and non-intuitive steps. The automated search for combinations of inequalities was initiated by Drori and Teboulle [2014] and later formalized by Taylor et al. [2017c,b] through the notion of convex interpolation, leading to the notion of *performance estimation problems* (PEP). Taylor et al. [2017a] developed a MATLAB package including several oracles and classes of functions. These notions are related to the notion of *Integral quadratic constraints* [Lessard et al., 2016], that appear in control theory. In the following, we outline a few key components of performance estimation problems by determining the worst-case guarantee for the gradient flow (1.13) applied to strongly convex functions. This example introduces the basic concepts central to Chapter 2.

1.3.1 Explaining example: the gradient flow

Let us start by recalling the gradient flow starting from $x_0 \in \mathbb{R}^d$,

$$\dot{X}_t = -\nabla f(X_t), X_0 = x_0 \in \mathbb{R}^d,$$

where f is μ-strongly convex. Given μ-strongly convex functions, Lemma 2 ensures $f(X_t) - f_\star \leqslant e^{-2\mu t}\left(f(x_0) - f_\star\right)$. This guarantee directly followed from the derivative with respect to time:

$$\frac{d}{dt}(f(X_t) - f_\star) = \langle \nabla f(X_t), \dot{X}_t\rangle \leqslant -2\mu\left(f(X_t) - f_\star\right).$$

Considering the class of functions $\mathcal{F}_\mu^\infty$, the gradient flow (1.13), the *performance measure* $\frac{d}{dt}(f(X_t) - f_\star)$ and the *initial condition* $f(X_t) - f_\star$, it is natural to wonder the worst possible value τ verifying,

$$\frac{d}{dt}(f(X_t) - f_\star) = \langle \nabla f(X_t), \dot{X}_t\rangle \leqslant -\tau\left(f(X_t) - f_\star\right).$$

The main idea of this framework is to formulate the smallest possible value $\tau \geqslant 0$ as the solution of an optimization problem:

$$\begin{aligned} -\tau_\star = \max_{f\in\mathcal{F}_{\mu,\infty}, X_t\in\mathbb{R}^d, d\in\mathbb{N}} & \frac{\langle \nabla f(X_t), \dot{X}_t\rangle}{f(X_t)-f_\star}, \\ \text{such that } & \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

Equivalently, via an homogeneity argument,

$$\begin{aligned} -\tau_\star = \max_{f\in\mathcal{F}_{\mu,\infty}, X_t\in\mathbb{R}^d, d\in\mathbb{N}} & \langle \nabla f(X_t), \dot{X}_t\rangle, \\ \text{such that } & f(X_t)-f_\star = 1, \\ & \dot{X}_t = -\nabla f(X_t). \end{aligned} \tag{1.14}$$

Due to the constraint '$f \in \mathcal{F}_{\mu,\infty}$', this problem is infinite dimensional. Taylor et al. [2017c] reformulated this problem as an SDP.

Reformulation using a sampled version of f. Let us first reformulate the condition '$f \in \mathcal{F}_{\mu,\infty}$' as a feasibility condition over the class of functions.

$$\begin{aligned} -\tau_\star = \max_{(X_i,g_i,f_i)_{t,\star}, d\in\mathbb{N}} & -\|g_t\|^2, \\ \text{subject to such that } & f_t - f_\star = 1, \\ & \exists f\in\mathcal{F}_{\mu,\infty}: \begin{cases} g_t = \nabla f(X_t) & f_t = f(X_t), \\ g_\star = \nabla f(x_\star) = 0 & f_\star = f(x_\star). \end{cases} \end{aligned}$$

By strong convexity, it holds that $\forall x, y, f(y) \geqslant f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|x-y\|^2$, which is in particular verified in $(x,y) = (X_t, x_\star)$ and $(x,y) = (x_\star, X_t)$. It holds that:

$$\begin{aligned} -\tau_\star = \max_{(X_i,g_i,f_i)_{t,\star}, g_\star=0, d\in\mathbb{N}} & -\|g_t\|^2, \\ \text{subject to such that } & f_t - f_\star = 1, \\ & f_t \geqslant f_\star + \frac{\mu}{2}\|x_t - x_\star\|^2, \\ & f_\star \geqslant f_t + \langle g_t, x_\star - x_t\rangle + \frac{\mu}{2}\|x_t + x_\star\|^2. \end{aligned} \tag{1.15}$$

Equality is due to an interpolation theorem [Taylor et al., 2017c, Theorem 4].

From a quadratic program to a semi-definite program. Problem (1.15) is a non-convex quadratic program. Indeed, its objective and constraints are quadratic in $x_t, x_\star, g_t$, and linear in $f_t, f_\star$. It turns out that this problem can be reformulated as a semidefinite program (SDP). We introduce the Gram matrix that is positive semi-definite by definition:

$$G = \begin{pmatrix} \|x_t - x_\star\|^2 & \langle g_t, x_t - x_\star\rangle \\ \langle g_t, x_t - x_\star\rangle & \|g_t\|^2 \end{pmatrix} \succcurlyeq 0,$$

together with the vector $F = \begin{pmatrix} f_t \\ f_\star \end{pmatrix}$. We can now reformulate Problem (1.15) into an SDP that is linear in F and in the Gram matrix G:

$$\begin{aligned} \tau_\star = \max_{G \succeq 0, F} \; & b_0 + \mathrm{Tr}(A_0 G), \\ \text{such that } & b_1^T F + \mathrm{Tr}(A_1 G) \geqslant 0, \\ & b_2^T F + \mathrm{Tr}(A_2 G) \geqslant 0, \\ & b_3^T F + \mathrm{Tr}(A_3 G) = 1, \end{aligned} \tag{1.16}$$

where $A_0 = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$, $A_1 = \begin{pmatrix} -\mu/2 & 0 \\ 0 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, $b_0 = 0$, $b_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $b_2 = -b_1$, $b_3 = b_1$. The generic formulation of the objective function and constraints as $\mathrm{Tr}(A_i G) + b_i^\top B \geqslant 0$ allows dealing with other types of objectives, such as $\|\nabla f(X_t)\|^2$, $\|X_t - x_\star\|^2$, $f(X_t) - f_\star$ and combinations of them. This will be particularly usefull in Chapter 2.

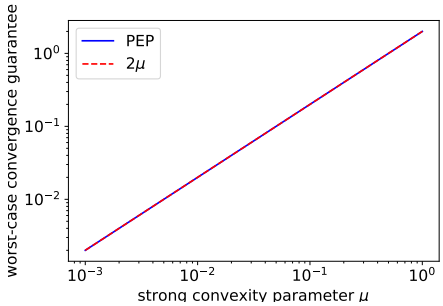


Figure 1.3: Comparison of $\tau_\star$ defined in (1.16) and of the known bound 2μ from Lemma 2, as a function of μ, solved using the solver MOSEK.

We formulated the problem of finding the worst-case scenario as an SDP, that can be efficiently computed thanks to solvers such as CVXPY [Diamond and Boyd, 2016], MOSEK [ApS, 2022] or SCS [O'Donoghue et al., 2016], as for example in Figure 1.3. Figure 1.3 ensures that the bound -2μ derived in Lemma 2 cannot be improved (and is even tight, as we can see on quadratics). In the following, we deduce convergence proofs from the dual of Problem (1.16).

Reformulation in the dual. The Lagrange dual function associated to the primal problem (1.16) takes the form:

$$\max_{G \succeq 0, F} \min_{\lambda_1, \lambda_2 \geqslant 0, \tau \in \mathbb{R}} \{\mathcal{L}(F, G, \lambda, \tau) = \sum_{i=0,1,2} \lambda_i \left(b_i^\top F + \mathrm{Tr}(A_i G) \right) + \tau \left(b_3^\top + \mathrm{Tr}(A_3 G) - 1 \right)\},$$

where $\lambda_1, \lambda_2 \geqslant 0$ and $\tau \in \mathbb{R}$ are dual variables associated with each constraint. The dual problem is obtained by maximizing the Lagrange dual function over $G \succcurlyeq 0, F \in \mathbb{R}^2$:

$$\begin{aligned} \min_{\lambda_1, \lambda_2 \geqslant 0, \tau \in \mathbb{R}} \; & -\tau, \\ \text{such that } & S = A_0 + \lambda_1 A_1 + \lambda_2 A_2 + \tau A_3 \preccurlyeq 0, \\ & b_0 + \lambda_1 b_1 + \lambda_2 b_2 + \tau b_3 = 0. \end{aligned} \tag{1.17}$$

By weak duality, each feasible point for the dual corresponds to an upper bound for $-\tau_\star$. In addition, strong duality holds (which can be proven by showing that there exists a Slater point in the primal [Taylor et al., 2017c, Theorem 6] [Boyd and Vandenberghe, 2004, Section 5.2.3])).

Let $\lambda_1, \lambda_2, \tau$ be feasible points of the dual (1.17) corresponding to the primal problem (1.16) (that is, the gradient flow originating from strongly convex functions). It holds that $\mathrm{Tr}(SG) \leqslant 0$ for any Gram matrix $G \succcurlyeq 0$, that is:

$$-\|g_t\|^2 + \lambda_1 \left(\frac{-\mu}{2} \|x_t - x_\star\|^2 \right) + \lambda_2 \left(\frac{-\mu}{2} \|x_t - x_\star\|^2 + \langle g_t, x_t - x_\star \rangle \right) \leqslant 0.$$

The condition '$b_0 + \lambda_1 b_1 + \lambda_2 b_2 + \tau b_3 = 0$' in the dual (1.17) implies $\lambda_1 - \lambda_2 + \tau = 0$. Multiplying this inequality by $f_t - f_\star$ and subtracting it to the inequality, we have:

$$\begin{aligned} -\|g_t\|^2 &\leqslant -\tau(f_t - f_\star) + \lambda_1 \underbrace{\left(f_\star - f_t + \frac{\mu}{2} \|x_t - x_\star\|^2 \right)}_{\leqslant 0} \\ &\quad + \lambda_2 \underbrace{\left(f_t - f_\star + \frac{\mu}{2} \|x_t - x_\star\|^2 + \langle g_t, x_\star - x_t \rangle \right)}_{\leqslant 0}, \\ &\leqslant -\tau(f_t - f_\star). \end{aligned} \tag{1.18}$$

In other words, if $\lambda_1, \lambda_2, \tau$ are feasible points for (1.17), then any function $f \in \mathcal{F}_\mu^\infty$ verifies $-\|\nabla f(x_t)\|^2 \leqslant -\tau \left(f(x_t) - f_\star \right)$ and thus, $-\tau_\star \leqslant -\tau$. To rephrase it, the proof for convergence consists in combining inequalities characterizing strong convexity with weights λ_1, λ_2. Let us explicit the conditions from the dual for the specific values of A_i, b_i, for $i = 0, 1, 2$:

$$\begin{aligned} & S = \begin{pmatrix} -\frac{\mu}{2}(\lambda_1 + \lambda_2) & \frac{1}{2}\lambda_2 \\ \frac{1}{2}\lambda_2 & -1 \end{pmatrix} \preccurlyeq 0, \\ & \lambda_2 = \lambda_1 + \tau, \\ & \lambda_1, \lambda_2 \geqslant 0. \end{aligned}$$

The constraint $S \preccurlyeq 0$ is equivalent with a nonpositive trace ('$-\frac{\mu}{2}(\lambda_1 + \lambda_2) - 1 \leqslant$'), which holds for $\lambda_1, \lambda_2 \geqslant 0$, and a nonnegative determinant. The dual (1.17) in case of the gradient flow originating from strongly convex functions reformulates as:

$$\begin{aligned} & \inf_{\lambda_1, \lambda_2 \geqslant 0} \lambda_1 - \lambda_2, \\ & \qquad \text{such that } \lambda_2^2 \leqslant 2\mu(\lambda_1 + \lambda_2). \end{aligned}$$

Considering the KKT conditions for this problem, we conclude that $\lambda_1 = 0$ and $\lambda_2 = \tau = 2\mu$, which corresponds exactly to Lemma 2. We leave the complete computation of the KKT conditions to the reader.

Therefore, obtaining convergence proofs can be viewed as finding analytical feasible points of the dual (1.17). In the case of Problem (1.14), optimal dual values are unique and can be computed analytically. However, dual values are often not unique and challenging to derive.

1.3.2 Discrete-time versions

Originally, performance estimation problems were developed to facilitate the convergence analysis of first-order methods. We provide briefly the example of gradient descent below originating from smooth strongly convex functions.

Let us first recall gradient descent (1.5) starting from $x_0 \in \mathbb{R}^d$:

$$x_{k+1} = x_k - \gamma \nabla f(x_k),$$

where f is L-smooth and μ-strongly convex. It holds from Theorem 2 for $\gamma = \frac{2}{\mu+L}$, for $k \geqslant 1$:

$$f(x_k) - f_\star \leqslant \frac{L}{\mu}\left(\frac{L-\mu}{L+\mu}\right)^{2k} (f(x_0) - f_\star)\,. \tag{1.19}$$

Again, the goal is to determine whether this convergence guarantee is tight or not. Given a class the class $\mathcal{F}_{\mu,L}$ and x_k a sequence generated by gradient descent (1.5), the performance measure $f(x_1) - f_\star$ and the initial condition $f(x_0) - f_\star$, we aim at computing the smallest possible value τ verifying:

$$f(x_{k+1}) - f_\star \leqslant \tau\,(f(x_k) - f_\star)\,.$$

This can be rephrased as the following optimization problem in one iteration:

$$\begin{aligned} \tau_\star = \max_{f\in\mathcal{F}_{\mu,L}, x_0\in\mathbb{R}^d, d\in\mathbb{N}} & f(x_1) - f(x_\star), \\ \text{such that } & f(x_0) - f_\star = 1, \\ & x_1 = x_0 - \frac{1}{L}\nabla f(x_0), \\ & \nabla f(x_\star) = 0. \end{aligned} \tag{1.20}$$

The class $\mathcal{F}_{\mu,L}$ benefits from interpolation conditions [Taylor et al., 2017c, Theorem 4.], allowing to apply the analysis performed for the gradient flow to gradient descent. In Figure 1.4, we compare the solution of (1.20) to the bound provided in Theorem 2. It turns out that the known bound can be improved, as already proven by Taylor et al. [2017c, Section 4.1.2].

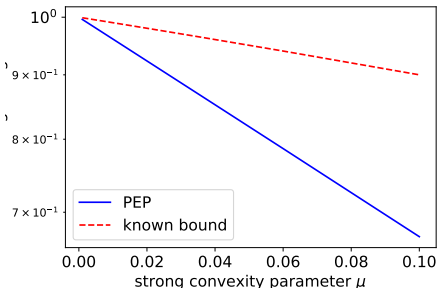


Figure 1.4: Comparison of $\tau_\star$ defined in (1.20) and of the bound (1.19) from Theorem 2, as a function of μ and for $L = 1$, solved using the solver MOSEK.

We presented the PEP technique, which is a computer-assisted technique for computing worst-case scenario of first-order methods and their continuous-time limits. Interpolation theorems are a key component for this problem, ensuring that these worst-case guarantees are tight, meaning there exists a function f that matches these bounds. They also allow for the reconstruction of corresponding convergence proofs by computing the dual. The framework includes a large number of oracles, such as (sub)-gradient steps, proximal steps or linear minimization oracle [Taylor et al., 2017b], as well as standard classes of functions [Taylor et al., 2017c,b] (convex functions within different class of assumptions, convex indicator functions, smooth nonconvex functions, ect.), classes of monotone operators [Ryu et al., 2020, Lieder, 2021] and more recently, quadratics [Bousselmi et al., 2024a] allowing for instance the study of primal-dual algorithms involving a linear operator [Bousselmi et al., 2024b]. However, the technique suffers from several limits. Among them, it may difficult to deduce analytical guarantees from the numerical analysis.

1.4 Concentration inequalities

In machine learning, we care about estimation error (1.3) which is typically upper bounded through:

$$\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)|,$$

or its deviation from a certain quantity $t \geqslant 0$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)| \geqslant t\right). \tag{1.21}$$

Controlling the quantity (1.21) can be achieved via concentration inequalities, which typically ensure that empirical estimates (like sample means) are close to their true values with high probability, which corresponds exactly to (1.21). Throughout this section, we mean by concentration inequalities how a certain random variables deviates from a quantity t, that is:

$$\mathbb{P}\left(X \geqslant t\right) \leqslant \phi(t).$$

For instance, when Θ is finite of size d and the loss l is bounded such that $\mathcal{R} = \mathbb{E}[f(y, f(w))]$ and $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(w_i))$, we conclude by Hoeffding's inequality (that is detailed below, or see, e.g., [Boucheron et al., 2013, Theorem 2.8]) that for any $t \geqslant 0$,

$$\begin{aligned}
\mathbb{P}(\sup_{\theta\in\Theta} |\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)| \geqslant t) &\leqslant d\mathbb{P}\left(|\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)| \geqslant t\right), \text{ union bound} \\
&\leqslant 2d\mathbb{P}\left(\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta) \geqslant t\right), \\
&\leqslant 2d\exp\left(-\frac{t^2 n}{2R^2}\right),
\end{aligned}$$

where for any $\sup_{y,\tilde{y}} l(y,\tilde{y}) = R$. In other words, '$|\hat{\mathcal{R}}_n(f_\theta) - \mathcal{R}(f_\theta)| \leqslant R\sqrt{\frac{2}{n}\log(\frac{2d}{\delta})}$' holds with probability $1-\delta$ with $\delta > 0$. Such an exponential decay in the number of samples n allows to consider broader families of predictors of size d diminishing the approximation error. For infinite function spaces, techniques like covering numbers (used for bounded function spaces as detailed for example in [Wainwright, 2019]) and Rademacher complexity (discussed, e.g., in Bousquet et al. [2004], Bartlett and Mendelson [2002]) have been developed to measure the richness of the class.

Below is a short overview of basic concentration inequalities and (sub)-Gaussian assumptions in machine learning, focusing on those appearing in Chapters 3 and 4. We first provide standard concentration inequalities applicable to a single random variable X. Subsequently, we present examples of concentration inequalities on the sum of independent random variables. In particular, we detail an example of the Cramer-Chernoff technique for Hoeffding's inequality, which sets the stage for Chapter 4. Lastly, we focus on sub-Gaussian random variables, which play a central role in machine learning and are particularly pertinent in Chapter 3 for establishing approximate convergence guarantees of greedy methods. For more details on general concentration inequalities, we refer to Boucheron et al. [2013, Section 2], Vershynin [2018, Chapter 2], Ledoux and Talagrand [1991, Chapter 6] and references therein.

1.4.1 Standard concentration inequalities

Let X be a random variable. We provide two elementary inequalities requiring few assumptions on X, namely fixed moments.

Markov's inequality. Let X be a random variable with finite mean $\mathbb{E}[X] < +\infty$ and almost surely nonnegative. Then, Markov's inequality provides an upper bound to the probability that X is greater or equal to a certain quantity $t > 0$

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}[X]}{t}.$$

The proof for this result can be found among others in Vershynin [2018, Proposition 1.2.4.]. If φ denotes a nonnincreasing and nonnegative function, it holds that $\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}[\varphi(X)]}{\varphi(t)}$. Given well-chosen functions φ such that $\mathbb{E}[\varphi(X)] < \infty$, it is possible to improve Markov's inequality. A standard example consists in considering the parametrized moment generating function $\varphi_\lambda(t) = e^{\lambda t}$ for $\lambda \in \mathbb{R}$ for $\lambda \in \mathbb{R}$. Then, it holds that $\mathbb{P}(X \geqslant t) \leqslant e^{-\lambda t}\mathbb{E}[e^{\lambda X}]$. Minimizing with respect to λ, it holds that:

$$\mathbb{P}(X \geqslant t) \leqslant \inf_{\lambda \in \mathbb{R}} \{e^{-\lambda t}\mathbb{E}[e^{\lambda X}]\}. \tag{1.22}$$

This result is also known as *Chernoff's bound*.

Chebyshev's inequality. Let random variable X with mean μ and variance σ^2. Chebyshev's inequality provides an upper bound to the probability that a random variable with finite variance deviates from its mean. It holds for any $t \geqslant 0$,

$$\mathbb{P}\left(|X - \mu| \geqslant t\sigma^2\right) \leqslant \min\left(1, \frac{1}{t^2}\right). \tag{1.23}$$

The proof for this result can be found for instance in Vershynin [2018, Corollary 1.2.5]. Chebyshev's inequality offers a guarantee of $O(\frac{1}{t^2})$, which is more precise than the $O(\frac{1}{t})$ guarantee provided by Markov's inequality. However, Chebyshev's inequality requires the additional assumption of knowing the variance (a second-order moment).

1.4.2 Concentration inequalities for the sum of independent random variables

This section is devoted to concentration inequalities on a sum of random variables, such as the empirical risk minimizer (1.2). Typically, data are assumed to be independent. We review three standard inequalities for the tail distribution of the sum of (real-valued) independent random variables: Hoeffding's, Bennett's, and Bernstein's inequalities. These inequalities use the same proof technique, known as the Cramer-Chernoff method, which we explain in detail for Hoeffding's inequality.

Hoeffding's inequality. Hoeffding's inequality is the most basic inequality for the sum of bounded independent random variables with fixed variance.

Theorem 8. *[Hoeffding, 1963, Theorem 2] Let $X_1, ..., X_n$ be independent random variables such that X_i takes it value in $[a_1, b_i]$ almost surely. Then, for every $t \geqslant 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_n - \mathbb{E}[X_n]) \geqslant t\right) \leqslant \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) = \rho_n^{\text{Hoeffding}}. \tag{1.24}$$

Proof. The X_i's are bounded almost surely and therefore have finite moments $E[X_i^q]$ for any

$q \in \mathrm{N}$ and its moment-generating function exists. We apply Chernoff's bound (1.22) for $\lambda \in \mathbb{R}$:

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_n - \mathbb{E}[X_n]) \geqslant t\right) &\leqslant e^{-n\lambda t}\mathbb{E}[e^{-\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])}], \\ &\leqslant e^{-n\lambda t}\prod_{i=1}^{n} e^{-\lambda(X_i - \mathbb{E}[X_i])}, \text{ by independence.}\end{aligned}$$

We now need to control the quantity $e^{-\lambda(X_i - \mathbb{E}[X_i])}$. First, notice that $|X - \frac{a+b}{2}| \leqslant \left(\frac{b-a}{2}\right)$ and thus that $\mathrm{Var}(X) = \mathrm{Var}(X - \frac{a+b}{2}) \leqslant \left(\frac{b-a}{2}\right)^2$. In addition, let $\varphi(\lambda) = \log\left(e^{-\lambda(X_i - \mathbb{E}[X_i])}\right)$. Its second-order derivative is equal to:

$$\begin{aligned}\frac{d^2\varphi}{d\lambda^2}(\lambda) &= e^{-\varphi(\lambda)}\mathbb{E}[(Y - \mathbb{E}[Y])^{\varphi(\lambda)}] + e^{-2\varphi(\lambda)}\mathbb{E}[(Y - \mathbb{E}[Y])^{\varphi(\lambda)}]^2, \\ &= \mathrm{Var}(Z),\end{aligned}$$

where Z is a random variable with density $x \mapsto e^{-\varphi(\lambda)}e^{\lambda x}$, that lies therefore almost surely in $[a, b]$. We then have $\mathrm{Var}(Z) \leqslant \left(\frac{b-a}{2}\right)^2$. We conclude by second-order Taylor expansion, by noting that $\varphi(0) = \frac{d\varphi}{d\lambda}(0) = 0$, leading to the overall inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_n - \mathbb{E}[X_n]) \geqslant t\right) \leqslant e^{-n\lambda t + \frac{2\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2}.$$

We then optimize with respect to λ and conclude the result above. □

Hoeffding's inequality states that the sum of independent bounded random variables has sub-Gaussian tails. The proof for Hoeffding's inequality relies heavily on Chernoff's inequality and the independence of the random variables. The key idea is to convert the evaluation of the expectation of the sum of random variables into a product of expectations of the individual random variables, as highlighted by Boucheron et al. [2013].

Other standard inequalities. Hoeffding's inequality requires minimal assumptions on random variables, specifically boundedness and fixed first-order moments. Like Markov's and Chebyshev's inequalities, Hoeffding's inequality can be refined by assuming fixed second-order moments, such as a fixed variance. When the random variables have finite variance, the weak law of large numbers states that the sample mean converges in probability to the population mean. Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ and variance σ^2. Then, it holds from Chebyshev's inequality (1.23) and independence of random variables that for any $t \geqslant 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant t\right) \leqslant \frac{\sigma^2}{nt^2}.$$

We review two standard (subgaussian) inequalities appearing in Chapter 4 and requiring finite variances. First, Bernstein's proposition controls the sum of independent bounded random variables with finite second-order moment (and thereby mean). Proposition 1 corresponds to Boucheron et al. [2013, Corollary 2.11], but more general versions are derived in Bernstein [1927, Section 2.8].

Proposition 1. *Let $X_1, \ldots, X_n$ be independent random variables in $[-c, c]$ with means $\mathbb{E}[X_i] =$*

μ_i, and variance $v = \sum_{i=1}^n \mathbb{E}[X_i^2]$. Then, for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geqslant nt + \sum_{i=1}^n \mu_i\right) \leqslant \exp\left(-\frac{n^2\frac{t^2}{2}}{v + c\frac{nt}{3}}\right).$$

Second, Bennett's inequality provides an exponential decay for the sum of semi-infinite random variables with finite second-order moment (and thereby finite mean).

Theorem 9. *[Bennett, 1968] Let $X_1, \cdots, X_n$ be i.i.d. random variables such that $X_i \leqslant a$ almost surely and $\sigma^2 = \sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}X_i)^2]$, then for any $t \geqslant 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\} \geqslant nt\right) \leqslant \exp\left(-\frac{\sigma^2}{a^2} h\left(\frac{atn}{\sigma^2}\right)\right),$$

where $h(t) = (1+t)\log(1+t) - t$. It implies that $\mathbb{P}(\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\} \geqslant nt) \leqslant \exp\left(-\frac{t^2n^2}{2(\sigma^2 + atn/3)}\right)$.

We have reviewed standard concentration inequalities that rely on minimal assumptions about random variables, specifically finite moments and a defined range of values.

1.4.3 Subgaussian properties

Subgaussian random variables play a significant role in machine learning, particularly in the analysis of algorithms and the derivation of probabilistic guarantees. For example, Gaussian assumptions are frequently employed to model data or noise, since it can often be viewed as the limiting behavior of the sample average of random variables (due to the central limit theorem). This section presents standard concentration results on the maximum of subgaussian random variables, and serves as a tutorial for Chapter 3 for deriving approximate convergence guarantee of (regularized) matching pursuit algorithms under subgaussian data.

A random variable X is considered subgaussian if its tail probabilities decay at least as fast as those of a Gaussian distribution. Usually, a random variable X with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = \sigma^2$ is said to be subgaussian if there exists $\tau > 0$ such that for any $t \in \mathbb{R}$,

$$\mathbb{E}[e^{tX}] \leqslant \exp\left(\frac{t^2\tau^2}{2}\right). \tag{1.25}$$

We recall below some useful characterization of subgaussian variables, which can be found for example in Vershynin [2018, Definition 2.5.2, Proposition 2.5.2].

Definition 3. *Let X be a subgaussian random with $\tau > 0$, and $\mathbb{E}[X] = 0$. Then, there exists absolute constants $c_1, c_2 > 0$ such that,*

$$\forall t \in \mathbb{R},\ \mathbb{E}[e^{tX}] \leqslant e^{c_1 \|X\|_{\Psi_2}^2 t^2},$$
$$\forall t \in \mathbb{R},\ \mathbb{P}(|X| \geqslant t) \leqslant 2\exp\left(-\frac{c_2}{\|X\|_{\Psi_2}^2} t^2\right),$$

In addition, $\|X\|_{\Psi_2} = \inf\{t > 0, \mathbb{E}[e^{X^2/t^2}] \leqslant 2\}$ is the subgaussian norm of X.

Typically, $\tau = \sqrt{2c_1\|X\|_{\Psi_2}}$. In other words, subgaussian random variables are characterized by their tail of probability bound as well as in terms of their moment generating function. Hoeffding's inequality ensures for example that the sum of independent bounded random variables has a subgaussian tail.

Subgaussian random variables offer several significant advantages when computing estimation errors (1.3) in machine learning and statistical analysis, especially when estimating the maximum

of such random variables. The following proposition offers a simple example of a concentration result on the maximum of subgaussian random variables.

Proposition 2. *Let $X_1, \dots, X_d$ be subgaussian random variables with parameter $\tau > 0$ as in Definition 3, then it holds that:*

$$\mathbb{E}\left[\max_{i=1,\dots d} X_i\right] \leqslant \sqrt{2\tau^2 \log(d)}.$$

In addition, it holds for any $t \geqslant 0$:

$$\mathbb{P}(\max_{i=1,\dots,d} X_i \geqslant \sqrt{2\tau^2 \log(d)} + t) \leqslant e^{-\frac{t^2}{2\tau^2}}.$$

Proof. By Jensen's inequality, we have for $t \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{E}[\max_{i=1,\dots d} X_i] &\leqslant \frac{1}{t} \log(\mathbb{E}[e^{t \max_{i=1,\dots,d} X_i}]), \\
&= \frac{1}{t} \log(\max(\mathbb{E}[e^{tX_i}])], \\
&\leqslant \frac{1}{t} \log(de^{t^2\tau^2/2}).
\end{aligned}$$

We optimize with respect to t, which leads to $t_\star = \sqrt{\frac{2}{\tau^2} \log(d)}$.

We now compute the concentration guarantee. Let $s \geqslant 0$, it holds that:

$$\begin{aligned}
\mathbb{P}(\max_{i=1,\dots,d} X_i \geqslant s) &\leqslant d\mathbb{P}(X_1 \geqslant s) \text{(by union bound)}, \\
&\leqslant d \inf_{\lambda \in \mathbb{R}} e^{-\lambda s} \mathbb{E}[e^{\lambda X}] \text{ (by Chernoff's bound)}, \\
&\leqslant de^{-\frac{s^2}{2\tau^2}}.
\end{aligned}$$

By taking $s = \sqrt{2\tau^2 \log(d)} + t$, we obtain the expected result. □

We conclude from Proposition 2 that the maximum of subgaussian random variables verifies a one-sided subgaussian tail, and does not require independence of random variables (as in Hoeffding's inequality). For a detailed approach to the supremum of empirical processes involving subgaussian variables, refer to Boucheron et al. [2013, Chapter 12] and Vershynin [2018, Section 2.5]. Similarly, Bennett's and Bernsteint's inequality exhibit subexponential tails (that is, for all $t \geqslant 0$, $\mathbb{P}(|X| \geqslant t) \leqslant 2\exp(-t/K_1)$ with $K_1 > 0$), for which it is possible to derive similar concentration results [Vershynin, 2018, Section 2.7].

We have reviewed a set of examples of concentration inequalities relevant to Chapters 3 and 4. The key takeaway from this section can be summarized as follows: *'given $X_1, \dots, X_n$, independent random variables with finite moments and/or taking their values in a certain range, it is possible to derive (subgaussian) tails for certain quantities of $X_1, \dots, X_n$'*. Their proofs happen to heavily rely on the Chernoff's bound, more precisely on the moment generating function. Chapter 4 focuses on generalizing this approach to other classes of upper-bounds. In Chapter 3, we derive examples of refined concentration inequalities for subgaussian random variables.

Throughout this introduction, we have presented fundamental results concerning the convergence of first-order methods and concentration inequalities, and reviewed techniques for computing these bounds. The core of this thesis contributes to these areas by introducing new greedy methods in Chapter 3, developing principled approaches to continuous-time models in Chapter 2, and addressing concentration inequalities through systematic approaches in Chapter 4.

CHAPTER 2

A systematic approach to Lyapunov analyses of continuous-time models in convex optimization

Objectives

First-order methods are often analyzed via their continuous-time models, where their worst-case convergence properties are usually approached via Lyapunov functions. In this work, we provide a systematic and principled approach to find and verify Lyapunov functions for classes of ordinary and stochastic differential equations. More precisely, we extend the performance estimation framework, originally proposed by Drori and Teboulle [2014], to continuous-time models. We retrieve convergence results comparable to those of discrete-time methods using fewer assumptions and inequalities, and provide new results for a family of stochastic accelerated gradient flows.

This chapter is published in *SIAM Journal on Optimization* [Moucer et al., 2023].

Codes. All codes for reproducing numerical results can be found at `https://github.com/CMoucer/PEP_ODEs`. We have also added continuous-time models to the Python package PEPit and detailed an example in Appendix 2.D, which is extracted from Goujaud et al. [2024] to come in *Mathematical Programming Computation*.

Contents

2.1 Introduction

Convex optimization is an important tool in the numerical analyst toolbox. It serves, among others, for framing modeling problems in data science and signal processing. We consider optimization problems of the form:

$$\min_{x \in \mathbf{R}^d} f(x), \tag{2.1}$$

where f is convex and differentiable. First-order methods (that gather information about f by evaluating its gradient at past iterates) are very popular to solve these problems, due to their attractive low cost per iteration, and to the fact that data science applications typically do not

require very accurate solutions [Bottou and Bousquet, 2007]. Gradient descent is a common first-order method, which starts from a point $x_0 \in \mathbf{R}^d$ and whose iterates are given by the simple recursion

$$x_{k+1} = x_k - \gamma \nabla f(x_k), \tag{2.2}$$

where $\gamma > 0$ is a step size. Gradient descent with small step sizes is directly related to the so-called gradient flow:

$$\dot{X}_t = -\nabla f(X_t), \;\; X_0 = x_0 \in \mathbf{R}^d, \tag{2.3}$$

with the notation $\dot{X}_t \triangleq \frac{d}{dt} X_t$ and where the solution X_t of the ordinary differential equation (ODE) verifies $X_{t_k} \approx x_k$ with the identification $t_k = \gamma k$. In numerical integration, gradient descent (2.2) is also known as the explicit Euler scheme for integrating gradient flows. Recently, Su et al. [2016] have interpreted Nesterov's accelerated gradient [Nesterov, 1983] in a similar fashion through its continuous-time version, paving the way to several continuous-time analyses of accelerated methods [Shi et al., 2022, Wilson et al., 2021, Wibisono et al., 2016].

Many applications entail some randomness and require a stochastic modeling of the function f, which is often defined in terms of an expectation $f(x) = \mathbf{E}_\xi[\tilde{f}(x, \xi)]$. The function f is the expectation over some random variable ξ, and accounts for some stochastic modeling. When ξ is drawn uniformly from a finite set of possible samples $(\xi_1, ..., \xi_n)$, we have a finite sum $f(x) = \frac{1}{n} \sum_{k=1}^n \tilde{f}(x, \xi_k)$. As soon as the number of data points n is large, computing the gradient of a finite sum, as it is done in gradient-based methods, is possibly expensive (computing the gradient of each element of the sum, which is possibly very large). Stochastic gradient descent (SGD) provides an alternative with lower computational burden per iteration, by evaluating only the gradient of a single $\tilde{f}(\cdot, \xi_{i_k})$ per iteration:

$$x_{k+1} = x_k - \gamma \nabla \tilde{f}(x_k, \xi_{i_k}),$$

where $\gamma > 0$ is the step size, ξ_{i_k} is drawn uniformly at random in $(\xi_1, ..., \xi_n)$. Thereby $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of the full gradient: $\mathbf{E}_{i_k}[\nabla \tilde{f}(x_k, \xi_{i_k})] = \nabla f(x_k)$. Li et al. [2017, 2019] derived stochastic differential equations (SDE) approximating SGD:

$$dX_t = -\nabla f(X_t) dt + \sigma(X_t) dB_t,$$

where $\sigma(X_t)$ is a noise parameter connected to parameters of the method, and these were further developed by Shi et al. [2020], Xu et al. [2018a]. Relying on approximate theorems between SDEs and original stochastic gradient algorithms, SDEs have thus become a tool for analyzing convergence speeds of discrete-time methods. Usually, gradient flows (resp. first-order methods) are studied via worst-case convergence properties, that hold for any function of a given class, and any trajectory generated by the ODE (resp. optimization method). In many cases, continuous-time approaches seem to allow for shorter, simpler, and thereby more intuitive proofs. They also bring insights on what can be expected from optimization methods.

The analysis of continuous-time models often relies on Lyapunov stability arguments, as in system theory and physics, where energy dissipation plays a crucial role. The existence of such Lyapunov functions provides direct convergence proofs for ODEs under consideration. The main challenge in the Lyapunov approach is to find a suitable function that is decreasing along all trajectory generated by an ODE.

From an outsider point of view, these analyses are often seen as complicated and technical to reach. In this work, we remedy this problem by extending the systematic approach based on semidefinite programming (SDP) originally coined by Drori and Teboulle [2014] for certifying convergence of optimization methods. This technique is referred as "performance estimation problems" (PEPs). The main contribution of this work is to provide a tool for analyzing convergence of continuous-time models, by constructing Lyapunov functions suited to a gradient-based

ODE in a systematic way, using small-sized SDPs reformulations. Furthermore, this procedure benefits from tightness properties, meaning that the feasibility of the SDP allows to conclude that there exists a Lyapunov function within the prescribed family of Lyapunov functions under consideration. Reciprocally, infeasibility of the SDP allows to conclude that there exists no such valid Lyapunov function within the family.

2.1.1 Lyapunov functions

Lyapunov functions are a standard tool for dealing with convergence properties of gradient flows. Such functions are also more generally used for studying stability properties of dynamical systems [Kalman and Bertram, 1960]. More precisely, consider a differentiable function f within a class $\mathcal{F}$, and an ODE aiming at minimizing f. For such a continuous-dynamical system with a stationary point $x_\star \in \operatorname{argmin}_x f(x)$, we call $\mathcal{V} : \mathbf{R}^d \times \mathbf{R}^+ \to \mathbf{R}$ a Lyapunov function if it is differentiable and satisfies the following conditions for all trajectory X_t generated by the ODE:

- $\mathcal{V}(x,t) = 0 \iff x = x_\star$,
- $\mathcal{V}(X_t,t) \geqslant 0$,
- $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$,

for all $t \geqslant 0$.

Lyapunov functions are suited for deriving both linear (or exponential) and sublinear convergence rates. When looking for linear convergence rates (as we may expect for strongly convex functions), we typically use $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\tau\mathcal{V}(X_t)$ instead of the third condition, where τ depends on the class of functions and on the ODE (more details Remark 2). In this ad-hoc definition, we enforced nonnegativity along the trajectory X_t, but definitions often require nonnegativity on $\mathbf{R}^d$ [Taylor et al., 2018a]. There exist similar definitions of Lyapunov functions for discrete-time optimization methods [Taylor et al., 2018a, Sanz Serna and Zygalakis, 2021, Lessard et al., 2016]

With this approach, convergence guarantees highly depend on how rich is the family of Lyapunov functions under consideration. In this work, we use a family of quadratic Lyapunov functions that is popular and natural for studying both discrete-time [Nesterov, 1983, d'Aspremont et al., 2021] and continuous-time [Su et al., 2016] optimization schemes.

2.1.2 Prior works

Lyapunov functions are common for analyzing continuous-time and discrete-time models in convex optimization. For example, convergence proofs for Nesterov's accelerated gradient method typically rely on such Lyapunov analyses [Nesterov, 1983] [d'Aspremont et al., 2021, Theorem 4.8]. In the recent work of Bansal and Gupta [2019], the authors proposed Lyapunov-based analyses for many first-order methods, for linear and sublinear convergence rates. Continuous-time versions of optimization methods also often involve Lyapunov arguments, such as Nesterov's accelerated gradient flow introduced in Su et al. [2016], and its high-resolution ODEs for strongly convex functions proposed in Shi et al. [2022], or accelerated mirror descent whose continuous-time dynamics was analyzed in Krichene et al. [2015].

Different techniques were developed to compute suitable Lyapunov functions. The authors of Wibisono et al. [2016] put forward an approach based on Bregman Lagrangian for accelerated methods in potentially non-Euclidean settings, further developed in Wilson et al. [2021]. Diakonikolas and Jordan [2021] directly derived Lyapunov functions from Hamiltonian equations describing dynamics of ODEs. Using similar conservation laws in a dilated coordinate system, Suh et al. [2022] also generated Lyapunov functions in a principled way.

Given a class of functions and an optimization method, proving a convergence rate mostly consists in combining inequalities characterizing the class of functions at hand. Recently, the automated search for combination of inequalities formulated as semidefinite programs was pioneered by Drori and Teboulle Drori and Teboulle [2014], and led to the notion of performance estimation problems. Their work was followed up in Taylor et al. [2017b,c] to provide worst-case bounds in a principled way, and extended to the Lyapunov framework [Taylor et al., 2018a]. A competing strategy inspired by control theory was developed by Lessard et al. [2016], Hu and Lessard [2017], where Lyapunov functions for discrete-time models are constructed using integral quadratic constraints (IQCs) and semidefinite programming; a similar approach was applied to continuous-time models in Fazlyab et al. [2018]. Connections between Lyapunov functions obtained via the IQC framework in continuous-time and discrete-time were highlighted by Sanz Serna and Zygalakis [2021].

For stochastic differential equations (SDEs), convergence proofs can also be obtained through the Lyapunov approach, together with Ito's calculus. For some well-chosen Lyapunov functions, Orvieto and Lucchi [2020] analyzed both SGD, SAGA [Defazio et al., 2014], and SVRG [Johnson and Zhang, 2013]. Xu et al. [2018a,b] extended the framework of Bregman Lagrangian Lyapunov functions [Wibisono et al., 2016] to the stochastic setting. To the best of our knowledge, a systematic way of verifying a Lyapunov function for SDEs has not been developed yet.

2.1.3 Contributions and organization

In this work, we are concerned with worst-case convergence analyses of ODEs and SDEs for modeling (stochastic) gradient based optimization methods. We propose a principled approach to worst-case analyses based on Lyapunov functions, SDPs and Ito's calculus.

In Section 2.2, we extend the performance estimation approach developed for optimization methods to gradient flows, which originates from a (possibly strongly) convex function. In short, we find Lyapunov functions as feasible points of certain linear matrix inequalities (LMIs). After that, building on the first part of this work for gradient-based ODEs, we analyze continuous-time versions of stochastic optimization algorithms.

Section 2.3 studies properties of trajectories generated by SDEs that approximates stochastic gradient methods. We obtain a simple version of the trade-off between forgetting the initial conditions and diminishing the noise, with and without averaging. It appears that decreasing step sizes, together with a nonuniform version of averaging, allows to reach an optimal trade-off between the two terms. Our results match those obtained for the stochastic gradient method, but with more compact analyses than those of the discrete-time setting.

In Section 2.4, we prove that stochastic accelerated gradient flows require diminishing step sizes to converge in our setting. In contrast to first-order stochastic gradient flows, averaging does not preserve convergence for SDEs approximating accelerated gradient methods with constant step size.

2.1.4 Assumptions

Throughout this work, functions to be minimized are convex (see Problem (2.1)). Under this assumption, stationary points are global minimizers. We restrict ourselves to continuous-time versions of gradient descent, accelerated gradient descent and stochastic gradient descent.

Let us recall a few basic definitions and properties characterizing the classes of functions under consideration within the next sections. A function $f : \mathbf{R}^d \to \mathbf{R}$ is convex if for all $x, y \in \mathbf{R}^d$, and for all $\lambda \in [0, 1]$, $f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y)$. Such functions (i.e. with full domain, $\mathrm{dom}(f) = \mathbf{R}^d$) are convex closed proper (CCP) (i.e., their epigraphs are non-empty closed convex sets). For simplicity, we assume in addition differentiability of f. Such a

differentiable function f is convex if and only if for all $x, y \in \mathbf{R}^d$, $f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle$. A differentiable function f is L-smooth if its gradient is L-Lipschitz, that is if for any $x, y \in \mathbf{R}^d$:

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

Smoothness is a common assumption for analyzing optimization methods, that limits the growth rate of the function. A convex differentiable function f is μ-strongly convex if for any $x, y \in \mathbf{R}^d$ it satisfies:

$$\|\nabla f(x) - \nabla f(y)\| \geqslant \mu\|x - y\|.$$

Strong convexity ensures the function is not too flat, and the unicity of the minimizer $x_\star$. We denote by $\mathcal{F}_{\mu,L}$ the family of L-smooth μ-strongly convex functions from $\mathbf{R}^d$ to $\mathbf{R}$, with $0 \leqslant \mu \leqslant L \leqslant +\infty$. Weaker assumptions than strong convexity are also encountered in the literature for analyzing gradient algorithms, and leads to similar convergence guarantees. Among them, Bolte et al. [2010, Appendix 5.2] proved linear convergence of gradient descent under Łojasiewicz inequality, first introduced by Lojasiewicz [1963]. Other relaxed versions of strong convexity followed [Necoara et al., 2019].

2.2 A principled approach to Lyapunov functions for gradient flows

In this section, we study convergence properties of the gradient flow and its accelerated versions, via quadratic Lyapunov functions. In this context, we show that verifying such a Lyapunov function can be formulated as verifying the feasibility of an LMI. This framework allows to search for Lyapunov functions, and to derive convergence bounds for non-autonomous gradient flows.

2.2.1 The gradient flow

Let us consider the gradient flow:

$$\dot{X}_t = -\nabla f(X_t), \ X_0 = x_0 \in \mathbf{R}^d.$$

Let $x_\star$ be a global minimizer of f. Without further assumptions, the function f is decreasing along the trajectory X_t solution to the gradient flow. The Lyapunov function $\mathcal{V}(X_t) = f(X_t) - f(x_\star)$ is indeed nonnegative, equal to zero at $x_\star$ and has a nonpositive derivative with respect to time $\frac{d}{dt}\mathcal{V}(X_t) = \dot{X}_t^\top \nabla f(X_t) = -\|\nabla f(X_t)\|^2$.

In the next section, we show how to obtain and verify such Lyapunov functions, and their corresponding convergence rates in the case of gradient flow originating from a strongly convex function.

2.2.1.1 Minimizing strongly convex functions

Let f be μ-strongly convex (i.e., $f \in \mathcal{F}_{\mu,\infty}$) with $\mu > 0$, and $x_\star$ be its unique minimizer such that $f(x_\star) = f_\star$. In this context, it is possible to prove linear convergence of the gradient flow to its stationary point. Scieur et al. [2017, Proposition 1.1] the following convergence bound in function values for the gradient flow:

$$f(X_t) - f_\star \leqslant e^{-2\mu t}(f(x_0) - f_\star). \tag{2.4}$$

This convergence guarantee follows directly from the derivative with respect to time of the Lyapunov function $\mathcal{V}(X_t) = f(X_t) - f_\star$, together with strong convexity (or Łojasiewicz inequality):

$\frac{d}{dt}\mathcal{V}(X_t) = \dot{X_t}^\top \nabla f(X_t) = -\|\nabla f(X_t)\|^2 \leqslant -2\mu(f(X_t) - f_\star) = -2\mu\mathcal{V}(X_t)$.

Given the specific gradient flows studied in this work, it is reasonable to search for Lyapunov functions made of linear combinations of function values, and a quadratic form in the trajectory X_t. We simply refer to them as quadratic Lyapunov functions:

$$\mathcal{V}_{a,c}(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2, \tag{2.5}$$

where $a, c \geqslant 0$ are constants that do not depend on t, such that the function $\mathcal{V}_{a,c}$ is nonnegative and nonincreasing along the flow (where we could assume $c = 1$ by homogeneity). As we will see in Section Quadratic Lyapunov functions are common for proving convergence of gradient flows and cover for instance the Lyapunov function used to prove convergence of the gradient flow under strong convexity (2.4). Given a Lyapunov function $\mathcal{V}_{a,c}$, the idea is to find the largest nonnegative value $\tau_\star$ such that the condition

$$\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau_\star \mathcal{V}_{a,c}(X_t), \tag{2.6}$$

holds for any dimension $d \in \mathbf{N}$, any function $f \in \mathcal{F}_{\mu,\infty}$ and any trajectory X_t generated by the gradient flow. After integration, (2.6) allows to obtain a convergence guarantee of the form $\mathcal{V}_{a,c}(X_t) \leqslant e^{-\tau_\star t}\mathcal{V}_{a,c}(X_0)$. Given a certain Lyapunov function $\mathcal{V}_{a,c}$ and a time t, we get that the largest acceptable $\tau_\star$ is a solution to:

$$\begin{aligned} -\tau_\star = \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}, f \in \mathcal{F}_{\mu,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a,c}(X_t), \\ \text{subject to } & \mathcal{V}_{a,c}(X_t) = 1, \\ & \dot{X}_t = -\nabla f(X_t). \end{aligned} \tag{2.7}$$

This minimization problem is invariant with respect to t. It is established in Taylor et al. [2017c], Drori and Teboulle [2014] that these so-called performance estimation problems (PEP) can be formulated as SDPs. Because of the variable $f \in \mathcal{F}_{\mu,\infty}$, the maximization problem (2.7) is infinite-dimensional. Recall that a differentiable function $f \in \mathcal{F}_{\mu,\infty}$ verifies for all points $x, y \in \mathbf{R}^d$, $f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geqslant \frac{\mu}{2}\|x - y\|^2$. Using alternate variables f_t, $f_\star$, g_t and $g_\star$ (informally: $f_t = f(X_t)$, $f_\star = f(x_\star)$, $g_t = \nabla f(X_t)$ and $g_\star = \nabla f(x_\star) = 0$), it holds that:

$$\begin{aligned} -\tau_\star = \max_{X_t \in \mathbf{R}^d, d \in \mathbf{N}} \quad & \frac{d}{dt}\mathcal{V}_{a,c}(X_t), \\ \text{subject to } & \mathcal{V}_{a,c}(X_t) = 1, \\ & \dot{X}_t = -g_t, \\ & f_i - f_j - \langle g_j, X_i - X_j\rangle \geqslant \frac{\mu}{2}\|X_i - X_j\|^2, \ \forall i, j = t, \star. \end{aligned} \tag{2.8}$$

The fact that (2.8) produces an upper bound on $-\tau_\star$ directly follows from the fact that any sampled strongly convex function satisfy these inequalities at the sampled points (here X_t and $x_\star$). Thereby, any feasible point to (2.7) corresponds to a feasible point for (2.8) with the same objective value. In the other direction, Taylor et al. [2017c, Corollary 2] (which provides a constructive way to obtain some $f \in \mathcal{F}_{\mu,\infty}$ that interpolates the triplets $(X_i, g_i, f_i)_{i=t,\star}$) ensures that any feasible point to (2.8) can be translated to a feasible point to (2.7) with also the same objective value, thereby reaching the equivalence between formulations (2.7) and (2.8).

In a second stage, we introduce $G = \begin{pmatrix} \|X_t - x_\star\|^2 & \langle X_t - x_\star, g_t\rangle \\ \langle X_t - x_\star, g_t\rangle & \|g_t\|^2 \end{pmatrix} \succcurlyeq 0$, a Gram matrix and

a vector $F = [f_t, f_\star]$, thereby obtaining a semidefinite reformulation:

$$\begin{aligned} -\tau_\star = \max_{G \succcurlyeq 0, F \in \mathbf{R}^2} \; & b_0^\top F + \mathrm{Tr}(A_0 G), \\ \text{subject to } & b_1^\top F + \mathrm{Tr}(A_1 G) \geqslant 0, \\ & b_2^\top F + \mathrm{Tr}(A_2 G) \geqslant 0, \\ & b_3^\top F + \mathrm{Tr}(A_3 G) = 1. \end{aligned} \tag{2.9}$$

where $A_0 = \begin{pmatrix} 0 & -c \\ -c & -a \end{pmatrix}$, $A_1 = \begin{pmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} -\mu/2 & 0 \\ 0 & 0 \end{pmatrix}$, $A_3 = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$, $b_0 = [0, 0]^\top$, $b_1 = [-1,\ 1]^\top$, $b_2 = [1,\ -1]^\top$ and $b_3 = a[1,\ -1]^\top$. Consider the corresponding Lagrangian, for $F \in \mathbf{R}^2$, $G \succcurlyeq 0$, $\tau \in \mathbf{R}$ and $\lambda_1, \lambda_2 \geqslant 0$:

$$\begin{aligned} \mathcal{L}(F, G, \tau, \lambda_1, \lambda_2) = & b_0^\top F + \mathrm{Tr}(A_0 G) + \tau \cdot (b_3^\top F + \mathrm{Tr}(A_3 G) - 1) \\ & + \lambda_1 \cdot (b_1^\top F + \mathrm{Tr}(A_1 G)) + \lambda_2 \cdot (b_2^\top F + \mathrm{Tr}(A_2 G)). \end{aligned}$$

The saddle point of the Lagrangian is given by:

$$\tau_\star = \min_{\tau, \lambda_1 \geqslant 0, \lambda_2 \geqslant 0} \max_{F, G \succcurlyeq 0} \mathcal{L}(F, G, \tau, \lambda_1, \lambda_2).$$

The Lagrangian dual of the SDP (2.9) is obtained by maximizing over $F \in \mathbf{R}^2$, $G \succcurlyeq 0$:

$$\begin{aligned} -\tau_\star = \min_{\tau, \lambda_1, \lambda_2} \; & -\tau, \\ \text{subject to } & S = A_0 + \lambda_1 A_1 + \lambda_2 A_2 + \tau A_3 \preccurlyeq 0 \\ & b_0 + \lambda_1 b_1 + \lambda_2 b_2 + \tau b_3 = 0, \\ & \tau \in \mathbf{R},\ \lambda_1, \lambda_2 \geqslant 0. \end{aligned}$$

The equality with $\tau_\star$ comes from strong duality, that holds via Slater's conditions (it is relatively easy to show that there exists a Slater's point using the same construction as in Taylor et al. [2017c, Theorem 6]). Finally, since any feasible τ is a lower bound for $\tau_\star$, those developments allow arriving to the equivalence between verifying a quadratic Lyapunov function and verifying the feasibility of an LMI.

Theorem 10. *Let $a, c, \tau \geqslant 0$ and $\mu > 0$. The following assertions are equivalent:*

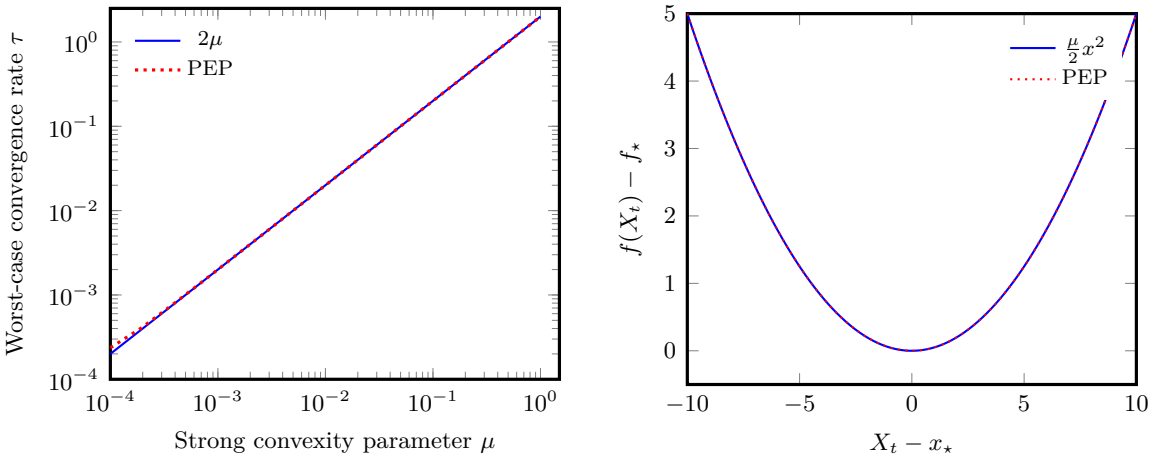
- *The inequality $\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau \mathcal{V}_{a,c}(X_t)$, is satisfied for all dimension $d \in \mathbf{N}$, for all $f \in \mathcal{F}_{\mu,\infty}$ and all trajectory X_t solutions to the gradient flow (2.3), where $\mathcal{V}_{a,c}$ is a quadratic Lyapunov function (2.5).*
- *There exist $\lambda_1, \lambda_2 \geqslant 0$ such that:*

$$S = \begin{pmatrix} \tau c - \frac{\mu}{2}(\lambda_1 + \lambda_2) & -c + \frac{\lambda_1}{2} \\ -c + \frac{\lambda_1}{2} & -a \end{pmatrix} \preccurlyeq 0,\ \tau a = \lambda_1 - \lambda_2. \tag{2.10}$$

Remark 1. *As a corollary of our result and for the class of quadratic Lyapunov functions (2.5) (and later (2.11), (2.14)), it turns out only two interpolation inequalities in $(X_t, x_\star)$ and $(x_\star, X_t)$ are involved in convergence proofs for continuous-time models (see Theorem 13 for second-order gradient flows). In other words, the framework reveals shorter proofs in continuous-time.*

A few conclusions can be drawn from the LMI equivalence from Theorem 10. For fixed value of a, c, τ, the LMI provides provides a necessary and sufficient condition for a quadratic Lyapunov function $\mathcal{V}_{a,c}$ to decrease at a specific rate $\tau \geqslant 0$ for all function in the class $\mathcal{F}_{\mu,\infty}$.

Second, we can simultaneously optimize over the class of quadratic Lyapunov functions and over the convergence rate. Indeed, given a rate τ, the LMI is jointly convex in $\lambda_1, \lambda_2, a, c$. Therefore, thanks to linearity of the feasibility problem in τ, a bisection search allows to optimize over it and to find the worst-case guarantee $\tau_\star$.



(a) Worst-case rate $\tau_\star$ for the class of quadratic Lyapunov functions (2.5)

(b) Reconstruction of a function $f \in \mathcal{F}_{\mu,\infty}$ that interpolates $x_\star$ and X_t, while matching the convergence rate $\tau_\star = 2\mu$, with $\mu = 0.1$.

Figure 2.1: Comparison between numerical values for τ obtained by solving the LMI (2.10) and the reference established in the literature [Scieur et al., 2017, Proposition 1.1], for trajectories X_t generated by gradient flow (2.3) originating from a μ-strongly convex function.

In Figure 2.1a, we obtain the fastest linear convergence rate that can be achieved using quadratic Lyapunov functions (2.5) (and even for all Lyapunov functions, since the rate is tight on a function $f(x) = \frac{\mu}{2}\|x\|_2^2$ for any time t). Together with Theorem 10 we retrieve the known linear worst-case convergence speed in $e^{-2\mu}$ from Scieur et al. [2017, Proposition 1.1] without improvement. The numerical approach allows to ensure tightness of the procedure by construction, as it guarantees the existence of a numerical function f that exactly achieves this convergence guarantee (see Figure 2.1b and the method in Taylor [2017, Chapter 3]). The next section builds on the same technique to analyze the gradient flow originating from a (possibly non strongly) convex function. In this scenario, the difficulty comes from the time-dependence of Lyapunov functions.

2.2.1.2 Minimizing convex functions

Let $f \in \mathcal{F}_{0,\infty}$ and $x_\star$ be a minimizer of f. In this case, worst-case convergence rates are often sublinear. Again, as in discrete-time, it is possible to obtain convergence guarantees using time-dependent quadratic Lyapunov functions.

The Lyapunov function $\mathcal{V}(X_t, t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ from Su et al. [2016, p. 7] verifies $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ for any dimension $d \in \mathbf{N}$, any function $f \in \mathcal{F}_{0,\infty}$, and any trajectory X_t generated by the gradient flow (2.3) (*proof*: $\frac{d}{dt}\mathcal{V}(X_t) = t\langle \nabla f(X_t), \dot{X}_t\rangle + f(X_t) - f_\star + \langle \dot{X}_t, X_t - x_\star\rangle = -t\|\nabla f(X_t)\|^2 + f(X_t) - f_\star - \langle \nabla f(X_t), X_t - x_\star\rangle \leqslant -t\|\nabla f(X_t)\|^2$ using convexity). After integration between 0 and t, we recover a convergence bound in function values from the literature [Su et al., 2016, p.7], [Fazlyab et al., 2018, Section 6.3.1]:

$$f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t}.$$

Let us adapt the techniques from Section 2.2.1.1 by considering quadratic Lyapunov functions:

$$\mathcal{V}_{a_t,c_t}(X_t, t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2, \tag{2.11}$$

where $a_t, c_t \geqslant 0$ are functions differentiable with respect to time such that the function $\mathcal{V}_{a_t,c_t}$ is nonnegative and nonincreasing along the flow. When a quadratic Lyapunov function decreases along the trajectory X_t, that is $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t) \leqslant 0$, a convergence guarantee in function values is given by

$$f(X_t) - f_\star \leqslant \frac{\mathcal{V}_{a_0,c_0}(x_0,0)}{a_t} = \frac{a_0(f(x_0)-f_\star) + c_0\|x_0 - x_\star\|^2}{a_t}.$$

Verifying a quadratic Lyapunov function can be cast as verifying that the following maximization problem is nonpositive:

$$\begin{aligned} 0 \geqslant \max_{X_t \in \mathbf{R}^d, d\in\mathbf{N}, f\in\mathcal{F}_{0,\infty}} \quad & \frac{d}{dt}\mathcal{V}_{a_t,c_t} \\ \text{subject to} \quad & \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

Remark 2. *The strongly convex case as defined above is a particular case of the convex one, using a specific Lyapunov function $\Phi(\cdot)$, such that $\mathcal{V}(X_t,t) = e^{\tau t}\Phi(X_t)$ where $\Phi(X_t) = a\cdot(f(X_t)-f_\star) + c\cdot\|X_t - x_\star\|^2$. Then, $\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0$ is equivalent to $\frac{d}{dt}\Phi(X_t) \leqslant -\tau\Phi(X_t)$.*

Theorem 11. *Let $a_t, c_t \geqslant 0$ be continuously differentiable with respect to time. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t) \leqslant 0$, is satisfied for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{0,\infty}$ and all trajectory X_t generated by the gradient flow (2.3), where $\mathcal{V}_{a_t,c_t}$ is a quadratic Lyapunov function defined in (2.11).*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that:*

$$S = \begin{pmatrix} \dot{c}_t & -c_t + \frac{\lambda_t^{(1)}}{2} \\ -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preccurlyeq 0, \quad \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

Proof. The LMI is obtained following the previous methodology (Appendix 2.A.1). □

Choosing $\lambda_t^{(1)} = 1$, $\lambda_t^{(2)} = 0$, together with $c_t = \frac{1}{2}$ and $a_t = t$, the conditions in the LMI from Theorem 11 are satisfied. We retrieve the Lyapunov function $\mathcal{V}(x,t) = t(f(x)-f_\star) + \frac{1}{2}\|x - x_\star\|^2$ from Su et al. [2016, p. 7].

Similar to Theorem 10, the LMI from Theorem 11, and hence the problem of looking for a Lyapunov function, is jointly convex in $\lambda_t^{(1)}$, $\lambda_t^{(2)}$, c_t, a_t, $\dot{a}_t$, $\dot{c}_t$. This Lyapunov analysis can also be validated numerically, as for the gradient flow originating from strongly convex functions.

2.2.2 Accelerated gradient flows

A major improvement to gradient descent dates back to Nesterov [1983], who introduced an accelerated gradient method (AGM):

$$\begin{aligned} x_{k+1} &= y_k - \gamma\nabla f(y_k), \\ y_{k+1} &= x_{k+1} + \alpha_k(x_{k+1} - x_k), \end{aligned} \tag{2.12}$$

where $\gamma, \alpha_k \geqslant 0$ depend on the class of functions to minimize. The combination of past iterates allows more control over the accumulated error. The idea of incorporating momentum was first introduced by Polyak [1964] with the heavy-ball method, starting from $x_0, x_1 \in \mathbf{R}^d$, and for a momentum $\alpha_k > 0$:

$$x_{k+2} = x_{k+1} + \alpha_k(x_{k+1} - x_k) - \gamma\nabla f(x_{k+1}). \tag{2.13}$$

Yet, compared to Nesterov's accelerated gradient method, the heavy-ball method lacks global acceleration beyond quadratics.

When the step size γ goes to zero, these schemes happen to be closely related to second-order differential equations, where $\beta_t \geqslant 0$ is a continuous function depending on α_k:

$$\ddot{X}_t + \beta_t \dot{X} + \nabla f(X_t) = 0.$$

Recently, accelerated gradient methods have been analyzed using second-order differential equations [Shi et al., 2022, Sanz Serna and Zygalakis, 2021, Wilson et al., 2021, Hu and Lessard, 2017]. Reversely, the accelerated gradient method and the heavy-ball method may be seen as discretization schemes of these second-order ODEs, as many other schemes. Discretization techniques are, among others, discussed by Xu et al. [2018a], Shi et al. [2020], Wilson et al. [2021]. Taking integration theory's point of view, Scieur et al. [2017] proved that these multi-step methods may even be seen as discretization schemes of the gradient flow (for quadratics).

Again, ODEs and multi-step first-order methods as defined above are often handled using quadratic Lyapunov functions. We extend the systematic Lyapunov approach developed previously to accelerated gradient flows. Let $\mathcal{V}_{a_t,P_t}$ be a quadratic Lyapunov function for second-order gradient flows, taken in the class of quadratic functions:

$$\mathcal{V}_{a_t,P_t}(X_t,t) = a_t(f(X_t) - f_\star) + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top (P_t \otimes I_d) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}, \tag{2.14}$$

where $P_t = \begin{pmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{pmatrix}$, $a_t \geqslant 0$ are continuously differentiable with respect to time and such that the Lyapunov function is nonnegative and nonincreasing along the flow. After integration of a quadratic Lyapunov function $\mathcal{V}_{a_t,c_t}$ between 0 and t, this approach leads to convergence bounds for instance in function values $f(X_t) - f_\star \leqslant \frac{\mathcal{V}(x_0)}{a_t}$.

2.2.2.1 Minimizing strongly convex functions

Let $f \in \mathcal{F}_{\mu,L}$, with strong convexity parameter $\mu > 0$, and let Nesterov's accelerated gradient methods' parameters by defined by $\gamma = \frac{1}{L}$ and $\alpha = \frac{1-\sqrt{\mu\gamma}}{1+\sqrt{\mu\gamma}}$ in (2.12). When the step size γ goes to zero, the continuous-time limit of y_k in (2.12) is exactly the Polyak damped oscillator [Polyak, 1964]:

$$\ddot{X}_t + 2\sqrt{\mu}\dot{X}_t + \nabla f(X_t) = 0, \tag{2.15}$$

as it has already been highlighted in previous works [Fazlyab et al., 2018, Sanz Serna and Zygalakis, 2021, Shi et al., 2022]. The heavy-ball method (2.13) reaches the same continuous-time limit when reducing the step size. Shi et al. [2022] derived a convergence guarantee in $f(X_t) - f_\star = \mathcal{O}(e^{-\frac{\sqrt{\mu}t}{4}})$ using a Lyapunov-based approach. This bound was improved to $f(X_t) - f_\star = \mathcal{O}(e^{-\sqrt{\mu}t})$, by Wilson et al. [2021, Appendix B] using the Bregman-Lagrangian approach, and by Sanz Serna and Zygalakis [2021], Fazlyab et al. [2018] using the IQC framework. Using the methodology from Theorem 10, we show that verifying linear convergence guarantees using quadratic Lyapunov functions with constant parameters (2.14) can be cast as an LMI.

Theorem 12. *Let $\mu > 0$ and $\tau \geqslant 0$. Let $a \geqslant 0$, and P be a symmetric matrix. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a,P}(X_t) \leqslant -\tau \mathcal{V}_{a,P}(X_t)$, is satisfied for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{\mu,\infty}$ and all trajectory X_t generated by the Polyak damped oscillator (2.15), where $\mathcal{V}_{a,P}$ is a quadratic Lyapunov function (2.14).*

- *There exist* $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$, *such that:*

$$\begin{aligned}
&\begin{pmatrix} -\frac{\mu}{2}(\lambda_1+\lambda_2)+\tau p_{11} & p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & -p_{12}+\frac{\lambda_1}{2} \\ p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & 2(p_{12}-2\sqrt{\mu}p_{22})+\tau p_{22} & -p_{22}+\frac{a}{2} \\ -p_{12}+\frac{\lambda_1}{2} & -p_{22}+\frac{a}{2} & 0 \end{pmatrix} \preccurlyeq 0, \\
&\tau a = \lambda_1 - \lambda_2, \\
&\begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{\mu}{2}(\nu_1+\nu_2) & 0 & \frac{-\nu_1}{2} \\ 0 & 0 & 0 \\ \frac{-\nu_1}{2} & 0 & 0 \end{pmatrix} \succcurlyeq 0, \\
&a = \nu_2 - \nu_1.
\end{aligned} \tag{2.16}$$

Proof. The equivalence is obtained using the methodology developed in Section 2.2.1.1 and introducing the Gram matrix $G = P^\top P$, where $P = (\dot{X}_t, X_t - x_\star, g_t)$, where g_t holds for $\nabla f(X_t)$. The first LMI refers to the non-increasing condition for the Lyapunov function. The second one refers to the positivity constraint $\mathcal{V}_{a,P}(X_t) \geqslant 0$, for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{0,\infty}$, and all trajectory X_t generated by Polyak damped oscillator (2.15), as done for discrete-time methods in Taylor et al. [2018a, Th. 7]. □

As for the gradient flow, the LMI is jointly convex in $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$, in the Lyapunov parameters a, P, and is linear in τ. Hence, we can perform a bisection search over τ to find the fastest linear convergence rate that can be verified using quadratic Lyapunov functions, as done in Figure 2.2a. The framework provides in addition a numerical tool for choosing Lyapunov parameters for which the worst-case linear convergence rate is achieved. Figure 2.2b helped providing an intuition for parameters in Corollary 1.

Corollary 1. *Let* $\mu \geqslant 0$. *The function*

$$\mathcal{V}(X_t) = f(X_t) - f_\star + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \left(\begin{pmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{pmatrix} \otimes I_d \right) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix},$$

verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -4/3\sqrt{\mu}\mathcal{V}(X_t)$ *for all dimension* $d \in \mathbf{N}$, *all function* $f \in \mathcal{F}_{\mu,\infty}$, *and all trajectory* X_t *generated by the Polyak damped oscillator* (2.15). *Tight rate is achieved for* $f(x) = \frac{1}{2}\mu x^2$.

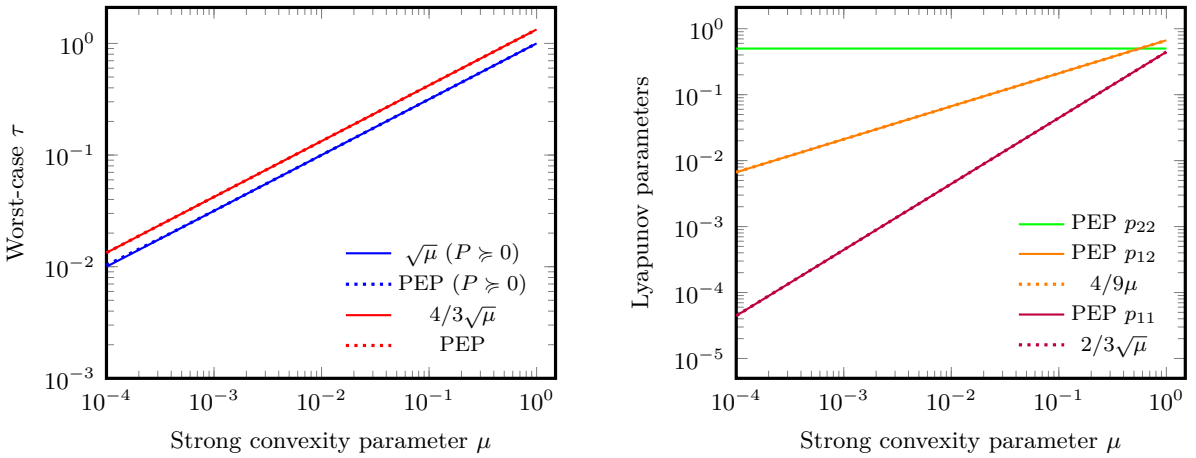
Proof. Taking $\lambda_1 = 4/3\sqrt{\mu}$, $\lambda_2 = 0$, $\nu_1 = 0$ and $\nu_2 = 1$, we verify the LMI for this Lyapunov function $\mathcal{V}$, with $\tau = 4/3\sqrt{\mu}$. □

This class of quadratic Lyapunov functions is inspired by Taylor et al. [2018a] in discrete-time, where a stricter positivity condition on $P \succcurlyeq 0$ hindered proving tight convergence of Nesterov's accelerated gradient. Similarly in our context, the Lyapunov function from Corollary 2.15 is defined by $P = \begin{pmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{pmatrix}$, which is not positive semidefinite. Usually in the continuous-time models' literature, we only consider matrices P that are positive semidefinite, such as in the Lyapunov function from Sanz Serna and Zygalakis [2021], Shi et al. [2022, Theorem 4.3]:

$$\mathcal{V}(X_t) = f(X_t) - f_\star + \frac{1}{2}\begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \left(\begin{pmatrix} \mu & \sqrt{\mu} \\ \sqrt{\mu} & 1 \end{pmatrix} \otimes I_d \right) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}$$

that verifies $\dfrac{d}{dt}\mathcal{V}(X_t) \leqslant -\sqrt{\mu}\mathcal{V}(X_t)$ for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{\mu,\infty}$ and all trajectory X_t generated by Polyak damped oscillator. This Lyapunov is a feasible point of the LMI (2.16) from Theorem 12, with $\tau = \sqrt{\mu}$, $\lambda_1 = \sqrt{\mu}$, $\lambda_2 = 0$, $\nu_1 = 0$, $\nu_2 = a = 1$. By relaxing

the condition $P \succcurlyeq 0$, we thus improve results from Sanz Serna and Zygalakis [2021] and Wilson et al. [2021] by a factor 4/3 in Corollary 1.



(a) Best guarantees found within the class of quadratic Lyapunov functions (2.14).

(b) Lyapunov parameters P in (2.14) for $\tau = 4/3\sqrt{\mu}$ and $a = 1$, as a function of the condition number μ.

Figure 2.2: Comparison between the worst-case guarantee obtained numerically with PEP, and its references, for the Polyak damped oscillator (2.15) originating from μ-strongly convex functions, and for quadratic Lyapunov functions (2.14).

Figure 2.2a together with Corollary 1 allows to conclude that this bound cannot be improved when changing the Lyapunov function among the class of quadratic functions (2.14).

2.2.2.2 Minimizing convex functions

As for the gradient flow, rates are sublinear when the accelerated gradient flow originates from convex functions. Let $f \in \mathcal{F}_{0,L}$, $\gamma \leqslant \frac{1}{L}$ be the step size, and $\alpha_k = \frac{k-1}{k+2}$ be the scheme parameter in Nesterov's accelerated method. Su et al. [2016, Section 2] proved the connection between the first-order scheme and a second-order ODE known as the accelerated gradient flow (AGF):

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0. \tag{2.17}$$

Su et al. [2016, Theorem 3] proved that the following inequality is verified for all function $f \in \mathcal{F}_{0,\infty}$ and all trajectory X_t generated by the accelerated gradient flow (2.17):

$$f(X_t) - f_\star \leq 2\frac{\|x_0 - x_\star\|^2}{t^2}.$$

Their proof exhibits a Lyapunov function $\mathcal{V}(X_t, t) = t^2(f(X_t) - f_\star) + 2\|(X_t - x_\star) + \frac{t}{2}\dot{X}_t\|^2$, which is decreasing along trajectories X_t (*proof:* $\frac{d}{dt}\mathcal{V}(X_t,t) = 2t(f(X_t)-f_\star) + t^2\langle \dot{X}_t, \nabla f(X_t)\rangle + 2\langle X_t - x_\star + \frac{t}{2}\dot{X}_t, 3\dot{X}_t + t\ddot{X}_t\rangle = 2t(f(X_t)-f_\star) - \langle \nabla f(X_t), X_t - x_\star\rangle \leqslant 0$ by convexity of f). In this context, we obtain again an LMI equivalence for verifying a quadratic Lyapunov function (2.14).

Theorem 13. *Let $a_t \geqslant 0$, $P_t \succcurlyeq 0$ be functions continuously differentiable with respect to time. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a_t,P_t}(X_t,t) \leqslant 0$, is satisfied for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{0,\infty}$ and all trajectory X_t generated by the accelerated gradient flow (2.17), where $\mathcal{V}_{a_t,P_t}$ is a quadratic Lyapunov function (2.14).*

- *There exist* $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ *such that:*

$$S = \begin{pmatrix} \dot{p}_t^{(11)} & p_t^{(11)} - \frac{3}{t}p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} \\ p_t^{(11)} - \frac{3}{t}p_t^{(12)} + \dot{p}_t^{(12)} & 2(p_t^{(12)} - \frac{3}{t}p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} & -p_t^{(22)} + \frac{a_t}{2} & 0 \end{pmatrix} \preccurlyeq 0,$$
$$\dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

Proof. The proof follows those from Theorem 12 and 10. □

The Lyapunov function $\mathcal{V}(X_t, t) = t^2(f(X_t) - f_\star) + 2\|(X_t - x_\star) + \frac{t}{2}\dot{X}\|^2$ exhibited by Su et al. [2016, Theorem 3] is a feasible point of the LMI, with $a_t = t^2$ and $P_t = 2\begin{pmatrix} 1 & t/2 \\ t/2 & t^2/4 \end{pmatrix}$ for the Lyapunov parameters, and $\lambda_t^{(1)} = t$, $\lambda_t^{(2)} = 0$. It is possible to retrieve these results numerically, as it was done for the accelerated gradient flow originating from strongly convex functions.

2.2.3 Higher-order convergence and time dilation

In this section, we analyze convergence of the first and second-order non-autonomous gradient flows, and provide convergence guarantees depending on their parametrization. It appears that higher-order convergence of non-autonomous gradient flows is highly connected to time dilation.

2.2.3.1 A non-autonomous first-order gradient flow

Let $f \in \mathcal{F}_{0,\infty}$. Let the non-autonomous first-order gradient flow be defined by:

$$\dot{X}_t = -\alpha_t \nabla f(X_t), \;\; X_0 = x_0 \in \mathbf{R}^d, \tag{2.18}$$

where $\alpha_t \geqslant 0$ is a continuous function (such that the flow is converging). It is natural to wonder if it is possible to accelerate such gradient flows by changing α_t. A change of variable connects this ODE to the gradient flow (2.3), for which $\alpha_t = 1$. Let Y_t be the solution to the gradient flow, and $\tau_t = \int_0^t \alpha_s ds$ be a time rescaling. Then, the variable $X_t = Y_{\tau_t}$ verifies $\dot{X}_t = \frac{d}{dt}Y_{\tau_t} = \alpha_t \dot{Y}_{\tau_t} = -\alpha_t \nabla f(Y_{\tau_t}) = -\alpha_t \nabla f(X_t)$, which is exactly the non-autonomous gradient flow. The following corollaries can be obtained by performing this change of variable in Theorem 11.

Corollary 2. *Let* $\mu \geqslant 0$.

- *If* $\mu > 0$, *the function* $\mathcal{V}(X_t, t) = e^{2\mu \int_0^t \alpha_s ds}(f(X_t) - f_\star)$ *verifies* $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant -2\mu\alpha_t \mathcal{V}(X_t, t)$ *for all dimension* $d \in \mathbf{N}$, *all function* $f \in \mathcal{F}_{\mu,\infty}$ *and all trajectory* X_t *generated by the non-autonomous gradient flow* (2.18).

  A convergence guarantee is given by $f(X_t) - f_\star \leqslant e^{-2\mu \int_0^t \alpha_s ds}(f(x_0) - f_\star)$.

- *If* $\mu = 0$, *the function* $\mathcal{V}(X_t, t) = \left(\int_0^t \alpha_s ds\right)(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ *verifies* $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ *for all dimension* $d \in \mathbf{N}$, *all function* $f \in \mathcal{F}_{0,\infty}$ *and all trajectory* X_t *generated by the non-autonomous gradient flow* (2.18).

  A convergence guarantee is given by $f(X_t) - f_\star \leqslant \dfrac{1}{2\int_0^t \alpha_s ds}\|x_0 - x_\star\|^2$.

Remark 3. *When* $\alpha_t = 1$ *above, that is* $\tau_t = t$, *we recover exactly the Lyapunov functions* $\mathcal{V}(X_t, t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2$ *from Theorem 11 for convex functions, and* $\mathcal{V}(X_t) = e^{2\mu t}(f(X_t) - f_\star)$ *from Theorem 10 for strongly convex functions.*

As mentioned by Orvieto and Lucchi [2020] in the stochastic setting, and for accelerated methods by Wibisono et al. [2016], one can thus either work with X_t generated by the non-autonomous gradient flow (2.18), or with Y_t generated by the gradient flow. However, acceleration on X_t is not preserved after discretizing the flow. For example, applying explicit Euler scheme to the non-autonomous gradient flow (2.18) with $\alpha_s = \alpha > 0$ and originating from a function $f \in \mathcal{F}_{0,L}$, a condition on step sizes h arises $0 \leqslant h \leqslant \frac{2}{L\alpha}$.

When focusing on continuous-time models for analyzing explicit first-order methods, we usually prefer working with the gradient flow (2.3). However, the non-autonomous gradient flow (2.18) may be useful for analyzing other methods such as proximal methods. More generally, in the next section, we analyze a family of second-order gradient flows without adjusting the time scale (taking $\alpha_t = 1$).

2.2.3.2 A non-autonomous second-order gradient flow

Nesterov's accelerated gradient flow reaches an $\mathcal{O}(\frac{1}{t^2})$ convergence in function values (see Theorem 13). We study convergence properties of a non-autonomous second-order gradient flow and compare them with those of the accelerated gradient flow, for the family of quadratic Lyapunov functions (2.14). Let $\beta_t \geqslant 0$ be a continuous function, and a second-order non-autonomous gradient flow:

$$\ddot{X}_t + \beta_t \dot{X}_t + \nabla f(X_t) = 0. \tag{2.19}$$

Remark 4. *Wibisono et al. [2016, Theorem 2.2] proved that this ODE is related to the family of ODEs defined by $\dot{Y}_t + \tilde{\beta}_t Y_t + \tilde{\alpha}_t \nabla f(Y_t) = 0$. Let $\alpha_t > 0$ be a continuously differentiable function with respect to time, $\tau_t = \int_0^t \alpha_s ds$ a time rescaling and X_t a solution to (2.19). The trajectory $Y_t = X_{\tau_t}$ is solution to $\ddot{Y}_t + (\beta_{\tau_t}\sqrt{\alpha_t} - \frac{\dot{\alpha}_t}{2\alpha_t})\dot{Y}_t + \alpha_t \nabla f(Y_t) = 0$. In contrast with α_t in (2.18), note that changing β_t in (2.19) does not correspond to a time rescaling.*

Theorem 14 provides a systematic condition for the function $\mathcal{V}$ taken among the class of quadratic Lyapunov functions (2.14) to be a Lyapunov function for the second-order gradient flow (2.19).

Theorem 14. *Let $a_t \geqslant 0$ and $P_t \succcurlyeq 0$ be continuously differentiable with respect to time, and $\mu \geqslant 0$. The following assertions are equivalent:*

- *The inequality $\dfrac{d}{dt}\mathcal{V}_{a_t,P_t}(X_t,t) \leqslant 0$ is satisfied for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{\mu,\infty}$ and all trajectory X_t generated by the non-autonomous second-order gradient flow (2.19), where $\mathcal{V}_{a_t,P_t}$ is a quadratic Lyapunov function of the form (2.14).*
- *There exist $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that:*

$$\begin{pmatrix} -\frac{\mu}{2}(\lambda_t^{(1)} + \lambda_t^{(2)}) + \dot{p}_t^{(11)} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} \\ p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} & -p_t^{(22)} + \frac{a_t}{2} & 0 \end{pmatrix} \preccurlyeq 0, \tag{2.20}$$

$$\dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

The LMI (2.20) from Theorem 14 is parametrized by β_t. When $\beta_t = \frac{3}{t}$ and $\mu = 0$, we recover Theorem 13 for Nesterov's accelerated gradient flow.

Corollary 3. *Let $\mu \geqslant 0$. The function*

$$\mathcal{V}(X_t,t) = a_t(f(X_t) - f_\star) + \frac{1}{2a_t}\|a_t\dot{X}_t + \dot{a}_t(X_t - x_\star)\|^2,$$

with a_t defined by:

- *If* $\mu > 0$, $a_t = e^{\tau t}$, *with* $\tau = \min(\sqrt{\mu}, \frac{2}{3}\beta_t)$,
- *If* $\mu = 0$, $a_t = \min\left((\sqrt{a_0} + (\sqrt{p_0^{(11)}}/2)t)^2, \lim_{\epsilon \to 0, \epsilon > 0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}\right)$,

verifies $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$ *for all dimension* $d \in \mathbf{N}$, *all function* $f \in \mathcal{F}_{\mu,\infty}$ *and all trajectory* X_t *generated the second-order gradient flow* (2.19).

Proof. The proof follows from Theorem 14 and is detailed in Appendix 2.A.2. □

Given a convex function $f \in \mathcal{F}_{0,\infty}$ and a quadratic Lyapunov function, Corollary 3 allows to conclude the following about the convergence of the second-order gradient flow (2.19): given the class of quadratic Lyapunov functions (2.14), it cannot converge faster in function values than Nesterov's accelerated gradient flow, i.e., not faster than $\mathcal{O}(\frac{1}{t^2})$.

To analyze Nesterov's accelerated gradient methods using ODEs, Su et al. [2016] introduced a parametrized second-order gradient flow, that fits the model (2.19):

$$\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0, \tag{2.21}$$

where $r \geqslant 0$. When $r \geqslant 3$, the guarantee $f(X_t) - f_\star \leqslant \frac{(r-1)^2\|x_0 - x_\star\|^2}{2t^2}$ holds for any function $f \in \mathcal{F}_{0,\infty}$ and any trajectory X_t generated by the accelerated gradient flow (2.21) [Su et al., 2016, Theorem 5]. When $r < 3$, Attouch et al. [2019, Theorem 2.1] proved a convergence bound in function values $f(X_t) - f_\star = \mathcal{O}(\frac{1}{t^{2r/3}})$. Using Corollary 3, we retrieve a similar bound in function values $f(X_t) - f_\star \leqslant \frac{2\|x_0 - x_\star\|^2 r^2}{9t^{\frac{2r}{3}}}$.

Remark 5. *Polynomial convergence can be achieved up to a time rescaling, as it was shown by Wibisono et al. [2016]. Given* $\tau_t = t^{p/2}$ ($\alpha_t = \frac{p}{2}t^{p/2-1}$), *Nesterov's accelerated gradient flow* ($r = 3$) *transforms into* $\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + \frac{p^2}{4}t^{p-2}\nabla f(X_t) = 0$ *for* $p \geqslant 2$. *Corollary 3 ensures convergence in function values* $f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t^p}$, *for all dimension* $d \in \mathbf{N}$, *all function* $f \in \mathcal{F}_{0,\infty}$ *and all trajectory* X_t *generated by the second-order gradient flow* (2.19).

We have extended the performance estimation approach to continuous-time models using Lyapunov functions. Given a (possibly accelerated) gradient flow and a class of functions, we presented a semidefinite formulation equivalent with the existence of a certain type of quadratic Lyapunov function. The next section is devoted to the analysis of continuous-time models approximating SGD.

2.3 SDEs for modeling SGD

Convergence results for stochastic convex optimization often require additional assumptions on problem classes, refined choices of step sizes and averaged iterates. Their analyses raise more complex proofs in contrast with deterministic methods. Prior works have been concerned with connections between stochastic methods and stochastic differential equations (SDEs) [Li et al., 2017, Orvieto and Lucchi, 2020, Xu et al., 2018a, Shi et al., 2019, 2022]. This section is devoted to convergence analyses of SDEs approximating stochastic methods using a systematic Lyapunov approach. Verifying a small-sized LMI will be sufficient (but not necessary) for verifying a quadratic Lyapunov function.

Stochastic gradient descent (SGD) is given by:

$$x_{k+1} = x_k - \gamma \nabla \tilde{f}(x_k, \xi_{i_k}), \tag{2.22}$$

where $\gamma > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, ..., \xi_n)$, and where $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$. Li et al. [2017] introduced stochastic modified equations (SMEs) to model SGD, rewriting it as

$$x_{k+1} = x_k - \gamma \nabla \tilde{f}(x_k, \xi_{i_k}) + \sqrt{\gamma} V_k(x_k),$$

where $V_k(x) = \sqrt{\gamma}\left(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k})\right)$ has zero mean and a covariance matrix equal to $\gamma\Sigma(x_k) = \gamma(\sum_{i=1}^n (\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k}))(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{i_k}))^\top)$.

The corresponding SDE is given by:

$$dX_t = -\nabla f(X_t)dt + (\gamma \Sigma(X_t))^{1/2} dB_t, \tag{2.23}$$

where B_t is a standard Brownian motion. The SDE (2.23) is an (order-1 weak) approximation of SGD ([Li et al., 2017, Theorem 1], [Li et al., 2019]), which allows to take into account the role of constant step size in the dynamics of SGD (while keeping them small). Under mild assumptions on f, Li et al. [2019] proved the weak approximation of SGD by this SDE on a finite interval $[0, T]$: there exists $C > 0$ such that $\|\mathbf{E}[x_k] - \mathbf{E}[X(k\gamma)]\| \leqslant C\gamma$ for $k \in [0, \frac{T}{\gamma}]$. However, the approximation point of view from this approach is relatively limited since C depends exponentially on T.

Remark 6. *The SDE (2.23) is an approximation of SGD for small step sizes $\gamma \geqslant 0$. When the step size goes to zero, the noise term actually disappears, and the limiting ODE of SGD is exactly the gradient flow (2.3). In contrasts, the stochastic Langevin dynamics [Langevin, 1908] $x_{k+1} = x_k - \gamma \nabla f(x_k) - \sqrt{\gamma}\xi_k$ has the limiting ODE $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$, where the step size is not taken into account.*

Compared to gradient descent, SGD does not converge to a stationary point under constant step sizes [Bach and Moulines, 2011, Taylor and Bach, 2019]. Convergence to a stationary point requires diminishing step sizes such as $\gamma_k = \frac{1}{\sqrt{k}}$. Li et al. [2017, Section 4.1] (and later Orvieto and Lucchi [2020, Section 2.1]) proposed to include this varying learning rate in the dynamics:

$$x_{k+1} = x_k - \gamma h_k \nabla f(x_k),$$

where γ is the maximum allowed learning rate and $h_k \in [0, 1]$ is the time-varying part. For $h_t \geqslant 0$ a continuous function playing the role of step sizes h_k, the SDE is:

$$dX_t = -h_t \nabla f(X_t)dt + h_t(\gamma \Sigma(X_t))^{1/2} dB_t. \tag{2.24}$$

We treat the covariance matrix $\Sigma(X_t)$ as symmetric, already implied by the notation $\Sigma(X_t)^{1/2}$, but unstructured with bounded variance $\Sigma(X_t) \preccurlyeq \Sigma$ along any trajectory X_t generated by the approximating SDE (2.24). Compared to ODEs, functions $f \in \mathcal{F}_{0,L}$ to be optimized using SDEs are in addition assumed to be possibly L-smooth with $L \in (0, \infty]$ and to be twice continuously differentiable.

In this section, we analyze approximating SDEs with averaging techniques, and include time-varying step sizes later. Verifying Lyapunov functions thanks to small-sized LMIs, we retrieve convergence results from discrete-time optimization methods, using appropriate choices of step sizes.

2.3.1 Lyapunov functions do not always extend to the stochastic setting

The analysis of the gradient flow in the deterministic case provides Lyapunov functions that are decreasing along any trajectory generated by (2.3) (see Section 2.2). The direct transfer of these Lyapunov functions to the stochastic setting is not always suited to the variance term, as

detailed below. Under constant step sizes $\gamma > 0$ (and $h_t = 1$), an approximating SDE of SGD is:

$$dX_t = -\nabla f(X_t)dt + (\gamma\Sigma(X_t))^{1/2}dB_t.$$

In SDE theory, a differential with respect to time of a function of a solution to a stochastic process is given by Ito's Lemma [Särkkä and Solin, 2019, Theorem 4.2].

Lemma 3 (Ito's Lemma)**.** *Let g be a twice continuously differentiable function, and X_t be a stochastic process solution to the SDE (2.23), then*

$$dg(X_t, t) = \frac{\partial}{\partial t}g(X_t, t)dt + \frac{\partial}{\partial x}g(X_t, t)dX_t + \frac{1}{2}\gamma\mathrm{Tr}(\frac{\partial^2}{\partial x^2}g(X_t, t)\Sigma(X_t))dt.$$

2.3.1.1 Minimizing strongly convex functions

When the SDE originates from (possibly non-smooth) strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$, the Lyapunov function from the deterministic setting extends well to SDEs. In the deterministic setting, we have shown in Theorem 10 that the function $\mathcal{V}(x, t) = e^{2\mu t}(f(x) - f_\star)$ is a Lyapunov function along the gradient flow. Given X_t a solution to the SDE (2.23), $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{2}e^{2\mu t}\gamma\mathbf{E}\mathrm{Tr}(\nabla^2_{xx}f(X_t)\Sigma(X_t))$ follows from Ito's formula. After integration between 0 and t, we have:

$$\mathbf{E}(f(X_t) - f_\star) \leqslant e^{-2\mu t}\left(f(x_0) - f_\star + \frac{1}{2}\gamma\int_0^t e^{2\mu s}\mathbf{E}\mathrm{Tr}(\nabla^2_{xx}f(X_s)\Sigma(X_s))ds\right).$$

We cannot conclude about the convergence of the trajectory X_t without additional requirements on the problem class. For example, let the smoothness parameter of f be finite $L < \infty$ and recalling the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term is bounded by $\frac{1}{2}L\gamma\mathrm{Tr}(\Sigma)$. Therefore, under constant step sizes, the SDE approximating SGD converges to a diffusion. In addition, it cannot get to a stationary point $x_\star$, because of the extra term, which depends linearly in the step size γ. As in the deterministic case, the forgetting of initial conditions remains of order $\mathcal{O}(e^{-2\mu t})$.

2.3.1.2 Minimizing convex functions

When $f \in \mathcal{F}_{0,\infty}$, Lyapunov functions induce convergence bounds with a possibly diverging variance term. When considering the deterministic gradient flow (2.3) originating from functions $f \in \mathcal{F}_{0,\infty}$, the Lyapunov function $\mathcal{V}(x, t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$ followed from Theorem 11. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, X_t be any trajectory solution to the SDE (2.23). Thanks to Ito's formula, we compute the derivative of $\mathcal{V}(\cdot)$ with respect to time as follows: $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant -t\mathbf{E}\|\nabla f(X_t)\|^2 + \mathbf{E}\frac{1}{2}\mathrm{Tr}((t\nabla^2_x f(X_t) + I)\Sigma(X_t))$. Using the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$ and assuming in addition f to be L-smooth with $L < +\infty$, a convergence bound in function values is given by:

$$\mathbf{E}(f(X_t) - f_\star) \leqslant \frac{\|x_0 - x_\star\|^2}{t} + \frac{1}{2}(L\frac{t}{2} + 1)\mathrm{Tr}(\Sigma).$$

Such an inequality does not allows to conclude about convergence of the SDE (2.24) without further assumptions.

In discrete-time, Taylor and Bach proved a comparable convergence bound for SGD [Taylor and Bach, 2019, Theorem 5], applying the Lyapunov performance estimation approach under similar assumptions (bounded variance, smoothness of f). Optimization methods and alternative techniques have been developed to ensure the global convergence of SGD to the optimum, among them averaging [Ruppert, 1988] and diminishing step sizes.

2.3.2 Diminishing the step size is a key to success

We study convergence of SDEs with time-varying step sizes (2.24). In contrast to the deterministic setting, in which time-varying step sizes correspond to a time rescaling whose benefit disappears after discretization, such a time rescaling plays a direct role in the variance term (explicit formula by Orvieto and Lucchi [2020, Theorem 5]).

Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, X_t be generated by (2.24), and $\mathcal{V}$ be a function. Our goal is to control the maximization problem

$$\max_{X_t \in \mathbf{R}^d, d \in \mathbf{N},\ f \in \mathcal{F}_{0,\infty}} \frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t),$$
$$\text{subject to } \ dX_t = -h_t \nabla f(X_t)dt + h_t(\gamma\Sigma(X_t))^{1/2}dB_t,$$

where $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) = \mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(X_t,t) + \frac{\partial}{\partial x}\mathcal{V}(X_t,t)\frac{dX_t}{dt}] + \frac{\gamma}{2}\mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial x^2}\mathcal{V}(X_t,t)\Sigma(X_t))$ is computed using Ito's formula. The first two terms correspond exactly to taking the derivative in trajectories generated by ODEs (2.18) (or the SDEs (2.24) with $\gamma = 0$), and the last term corresponds to a variance term. Because of the trace in a second-order derivative of f and in the covariance matrix $\Sigma(X_t)$, we do not take this term into account in an LMI formulation. Instead, we propose to first derive a family of Lyapunov functions for associated ODEs using LMIs (deterministic setting), and then to optimize their parameters so that the variance term converges conveniently. In other words, the following functions $\mathcal{V}$ are Lyapunov functions for induced ODEs ($\gamma = 0$), but are not Lyapunov functions for SDEs under consideration. This approach allows a systematic computation of such quadratic functions and leads to convergence guarantees.

Corollary 4. *Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, and $X_t \in \mathbf{R}^d$ be generated by the SDE (2.24). The quadratic function*

$$\mathcal{V}(X_t,t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2,$$

with $\dot{a}_t^{(1)} = 2h_t$ verifies $\frac{d}{dt}\mathbf{E}(\mathcal{V}(X_t,t) \leqslant h_t^2\mathbf{E}\mathrm{Tr}((\nabla^2_{xx}f(X_t)a_t^{(1)} + \frac{1}{2}I_d)\Sigma(X_t))$.

Furthermore, it holds that:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{\gamma}{2a_t^{(1)}}\int_0^t h_s^2\mathbf{E}\mathrm{Tr}((\nabla^2_{xx}f(X_s)a_s^{(1)} + \frac{1}{2}I_d)\Sigma(X_s))ds.$$

Proof. The function $\mathcal{V}$ is obtained using Corollary 2, and is a Lyapunov function for a non-autonomous first-order gradient flow. The bound for $\mathbf{E}[f(X_t) - f_\star]$ is derived using Ito's formula on $\mathcal{V}$ along trajectories X_t generated by the SDE (2.24). □

The convergence bound from Corollary 4 is divided in two terms: a term that forgets the initial conditions and a variance term due to noise. Convergence is mostly controlled by the step size ($\dot{a}_t^{(1)} = 2h_t$). Bach and Moulines [2011, Theorem 5] provided a comparable but much more complex analysis for stochastic gradient descent, for a specific family of step sizes. To compare to our results, let us consider step sizes defined by $h_t = \frac{1}{(t+1)^\alpha}$, where $\alpha \geqslant 0$. A possible convergence guarantee arises from Corollary 4 with $a_t^{(1)} = (t+1)^{1-\alpha}$. The forgetting of the initial condition is thus bounded by $\frac{\|x_0 - x_\star\|^2}{(t+1)^{1-\alpha}}$. Provided $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term is bounded by:

$$\begin{cases} \frac{\gamma\mathrm{Tr}(\Sigma)}{(t+1)^{1-\alpha}}\left(L\frac{(t+1)^{2-3\alpha}-1}{2-3\alpha} + \frac{1}{2}\frac{(t+1)^{1-2\alpha}-1}{1-2\alpha}\right) & \text{if } 0 \leqslant \alpha < 1, \\ \frac{\gamma\mathrm{Tr}(\Sigma)}{\log(t)}\left(L\int_0^t \frac{\log(s+1)}{(s+1)^2}ds + \frac{1}{2}(1 - \frac{1}{t+1})\right) & \text{if } \alpha = 1. \end{cases}$$

The variance term converges to zero if and only if $\alpha \geqslant \frac{1}{2}$. In other words, convergence is not guaranteed for constant step sizes ($\alpha = 0$). For $\alpha \in (1/2, 2/3)$, the convergence in function value is bounded by $\mathcal{O}(\frac{1}{t^{2\alpha-1}})$. For $\alpha \in (2/3, 1)$, the convergence in function value is bounded by $\mathcal{O}(\frac{1}{t^{1-\alpha}})$. As for SGD [Bach and Moulines, 2011, Theorem 5], the convergence regime changes at $\alpha = \frac{2}{3}$ with a global convergence rate in $\frac{1}{t^{1/3}}$, for which the variance term and the term that forgets the initial conditions converge at the same rate (up to $\log(t)$). It is therefore possible to reach convergence with diminishing step sizes. Other techniques, such as averaging, have been developed to improve the trade-off between faster convergence and larger step sizes.

2.3.3 Averaging for larger step sizes

Polyak-Ruppert averaging [Ruppert, 1988, Polyak and Juditsky, 1992] is a standard way to improve convergence of SGD. In the discrete-time setting, convergence guarantees are considered at an averaged sequence, where i_k is drawn uniformly in $[1, ..., n]$ and n the sample size:

$$\begin{aligned} x_{k+1} &= x_k - \gamma h_k \nabla f_{i_k}(x_k), \\ \bar{x}_k &= \frac{1}{k} \sum_{i=1}^{k} x_i. \end{aligned} \tag{2.25}$$

Other averaging techniques were developed later, such as primal averaging [Tao et al., 2018] and averaging with respect to some nonnegative function [Le Roux, 2019].

2.3.3.1 Polyak-Ruppert averaging

In this section, we analyze convergence properties of an SDE (2.24) approximating SGD with time-varying step sizes under Polyak-Ruppert averaging. Taylor and Bach [2019, Theorem 6] provided a systematic design of Lyapunov functions for (2.25), and a condition on the step size for convergence to the optimum. An approximating SDE for Polyak-Ruppert averaging, defined by $\bar{X}_t = \frac{1}{t}\int_0^t X_s ds$, is given by:

$$\begin{aligned} dX_t &= -h_t \nabla f(X_t) dt + h_t (\gamma \Sigma(X_t))^{1/2} dB_t, \\ d\bar{X}_t &= \frac{X_t - \bar{X}_t}{t} dt, \end{aligned} \tag{2.26}$$

with step size $\gamma > 0$ that is taken close to zero, and a variable term $h_t \in [0, 1]$. We introduce the family of quadratic functions taking the averaged sequence into account:

$$\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top (P_t \otimes I_d) \begin{pmatrix} X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}, \tag{2.27}$$

where $a_t^{(1)}, a_t^{(2)} \geqslant 0$ and $P_t = \begin{pmatrix} p_t^{(11)} & p_t^{(12)} \\ p_t^{(12)} & p_t^{(22)} \end{pmatrix} \succcurlyeq 0$ are continuously differentiable with respect to time. Given a quadratic function $\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}$ and an SDE (2.26) under Polyak-Ruppert averaging, we present in Theorem 15 a way to control the quantity:

$$\begin{aligned} \max_{f \in \mathcal{F}_{0,\infty}, d \in \mathbf{N}, X_t \in \mathbf{R}^d} \quad & \frac{d}{dt} \mathbf{E} \mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t), \\ \text{subject to } & dX_t = -h_t \nabla f(X_t) dt + h_t (\gamma \Sigma(X_t))^{1/2} dB_t, \\ & d\bar{X}_t = \frac{X_t - \bar{X}_t}{t} dt. \end{aligned}$$

Theorem 15. *Let $h_t \geqslant 0$, and $a_t \geqslant 0$, $P_t \succcurlyeq 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t) \in \mathbf{R}^d \times \mathbf{R}^d$ be a trajectory generated by the SDE under Polyak-Ruppert averaging (2.26). Let $\mathcal{V}$ be a quadratic function (2.27), such that there exist $\lambda_t^{(1)}, ..., \lambda_t^{(6)} \geqslant 0$ verifying:*

$$\begin{pmatrix} \dot{p}_t^{(11)} + \frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)} + \frac{p_t^{(22)} - p_t^{(12)}}{t} & \frac{\lambda_t^{(6)} + \lambda_t^{(4)} - 2h_t p_t^{(11)}}{2} & \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} \\ \dot{p}_t^{(12)} + \frac{p_t^{(22)} - p_t^{(12)}}{t} & \dot{p}_t^{(22)} - \frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)} - 2h_t p_t^{(12)}}{2} & -\frac{a_t^{(2)}}{2t} + \frac{\lambda_t^{(5)} + \lambda_t^{(3)}}{2} \\ \frac{\lambda_t^{(6)} + \lambda_t^{(4)} - 2h_t p_t^{(11)}}{2} & -\frac{\lambda_t^{(6)} - 2h_t p_t^{(12)}}{2} & -a_t^{(1)} & 0 \\ \frac{a_t^{(2)}}{2t} - \frac{\lambda_t^{(5)}}{2} & -\frac{a_t^{(2)}}{2t} + \frac{\lambda_t^{(5)} + \lambda_t^{(3)}}{2} & 0 & 0 \end{pmatrix} \preccurlyeq 0,$$

$$\dot{a}_t^{(1)} + \lambda_t^{(1)} + \lambda_t^{(5)} = \lambda_t^{(4)} + \lambda_t^{(6)},$$
$$\dot{a}_t^{(2)} + \lambda_t^{(2)} + \lambda_t^{(6)} = \lambda_t^{(3)} + \lambda_t^{(5)} + \dot{a}_t^{(1)}.$$

Then, the following inequality is satisfied:

$$\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t, t) \leqslant \frac{1}{2}\mathbf{E}\mathrm{Tr}\left((a_t^{(1)}\nabla_{xx} f(X_t) + 2p_t^{(11)} I_d)\Sigma(X_t)\right) h_t^2 \gamma.$$

Proof. The proof follows the method from Section 2.2.1.1 (see Appendix 2.B.1). □

The variance term $\frac{1}{2}\mathbf{E}\mathrm{Tr}((a_t^{(1)}\nabla_{xx} f(X_t) + 2p_t^{(11)} I_d)\Sigma(X_t))h_t^2\gamma$ from Theorem 15 increases with $a_t^{(1)}$, and its convergence requires additional assumptions on f, for example smoothness. For this reason, we propose to analyze convergence guarantees based on functions $\mathcal{V}_{0,a_t^{(2)},P_t}$, on the averaged sequence only.

Corollary 5 (Averaging and diminishing step sizes)**.** *Let $h_t \geqslant 0$ and $a_t^{(2)} \geqslant 0$ be continuously differentiable. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t)$ be a trajectory generated by the SDE under Polyak-Ruppert averaging (2.26). Assuming that $\dot{a}_t^{(2)} \leqslant \frac{a_t^{(2)}}{t}$ and $t \to \frac{a_t^{(2)}}{th_t}$ is a non-increasing function, the function*

$$\mathcal{V}(X_t, \bar{X}_t, t) = a_t^{(2)}(f(\bar{X}_t) - f_\star) + \frac{a_t^{(2)}}{2th_t}\|X_t - X_\star\|^2$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t, t) \leqslant \frac{a_t^{(2)}}{t} h_t \mathbf{E}\mathrm{Tr}(\Sigma(X_t))$. Furthermore, it holds that $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{2a_t^{(2)}} + \frac{\gamma}{2a_t^{(2)}} \int_0^t \frac{a_s^{(2)}}{s} h_s \mathbf{E}\mathrm{Tr}(\Sigma(X_s))ds$.

Proof. The proof follows from the LMI in Theorem 15 with the choices $\lambda_t^{(3)} = \lambda_t^{(6)} = \lambda_t^{(2)} = 0$, $\lambda_t^{(5)} = \frac{a_t^{(2)}}{t}$, $\lambda_t^{(4)} = \lambda_t^{(1)} = h_t p_t^{(11)}$, $\dot{a}_t^{(2)} = \frac{a_t^{(2)}}{t}$, $\dot{p}_t^{(11)} = (\frac{a_t^{(2)}}{th_t})^{\cdot}$, $\dot{p}_t^{(12)} = \dot{p}_t^{(22)} = 0$. □

When $a_t^{(2)} = t$ (its maximal possible value), the step size verifies $\dot{h}_t \leqslant 0$. The variance term does not diverge if and only if h_t is constant. Recalling the unbounded covariance $\Sigma_t \preccurlyeq \Sigma$, a convergence bound is given by $\mathbf{E}[f(\bar{X}_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{2t} + \frac{1}{2}\mathrm{Tr}(\Sigma)\gamma h$. The decreasing condition on $t \to \frac{a_t^{(2)}}{th_t}$ suggests a trade-off between convergence and diminishing step size, as obtained without averaging.

Under the assumptions of Corollary 5, recalling that $\Sigma(X_t) \preccurlyeq \Sigma$ is a bounded covariance matrix, let $h_t = \frac{1}{(t+1)^\alpha}$ be the step size and $a_t^{(2)} = t^\beta$, where $\alpha \geqslant 0$ and $0 \leqslant \beta \leqslant 1$. The decreasing condition imposes $\alpha + \beta \leqslant 1$. In comparison with step sizes requirements drawn from Corollary 4, where convergence required $\alpha \geqslant \frac{1}{2}$, averaging allows larger step sizes.

A different behavior is expected from α and β: on the one hand, an ideal step size should be large (α small), and on the other hand, we aim at converging as fast as possible (β large). The term that forgets the initial conditions behaves as $\mathcal{O}(\frac{1}{t^\beta})$, and the variance term as $\mathcal{O}(\frac{1}{t^\alpha})$ if $\beta \neq \alpha$. When $\alpha = \beta$, the variance term behaves as $\mathcal{O}(\frac{\log(t)}{2t^\beta})$. Hence, a natural choice is $\alpha = \beta = \frac{1}{2}$, retrieving results from Bach and Moulines [2011, Theorem 4] and Taylor and Bach [2019, Table 2] in discrete-time optimization.

2.3.3.2 Weighted averaging

Polyak-Ruppert averaging performs uniform averaging of any trajectory X_t over the time step. We introduce weighted averaging to analyze SGD, that is defined with respect to a function $u_t \geqslant 0$ [Le Roux, 2019]:

$$\bar{x}_t^u = \frac{1}{\int_0^t u_s ds} \int_0^t u_s x_s ds.$$

Under weighted averaging, and introducing $C_t^u = \frac{u_t}{\int_0^t u_s ds}$, the SDE is given by:

$$\begin{aligned} dX_t &= -h_t \nabla f(X_t) dt + h_t (\gamma \Sigma(X_t))^{1/2} dB_t, \\ d\bar{X}_t^u &= (X_t - \bar{X}_t^u) C_t^u dt. \end{aligned} \tag{2.28}$$

We study convergence of this generalized version of Polyak-Ruppert averaging, and compare it to traditional averaging techniques.

Theorem 16. *Let $\mu \geqslant 0$. Let $d \in \mathbf{N}$ be the dimension, $f \in \mathcal{F}_{\mu,\infty}$ be a twice continuously differentiable (possibly strongly convex) function and $(X_t, \bar{X}_t)$ be trajectories generated by an SDE under generalized averaging (2.28). Assuming $\left(\frac{\int_0^t u_s ds}{2h_t}\right)^{\cdot} 2\mu u_t$, the function*

$$\mathcal{V}(X_t, \bar{X}_t^u, t) = \frac{u_t}{C_t^u}(f(\bar{X}_t^u) - f_\star) + \frac{u_t}{2h_t}\|X_t - x_\star\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t^u, t) \leqslant \frac{1}{2} u_t h_t \gamma \mathbf{E}\mathrm{Tr}(\Sigma(X_t))$.

Then, it holds that $\mathbf{E}[f(\bar{X}_t^u) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2 u_0}{2h_0 \int_0^t u_s ds} + \frac{\gamma}{4\int_0^t u_s ds} \int_0^t u_s h_s \mathbf{E}\mathrm{Tr}(\Sigma(X_s)) ds.$

Proof. The proof follows those from Theorem 15 and Corollary 5 with $\frac{1}{t} \to C_t^u$. □

Under the convexity assumption ($\mu = 0$), u_t verifies $\left(\frac{\int_0^t u_s ds}{2h_t}\right)^{\cdot} \leqslant 0$. Assuming step sizes of the form $h_t = \frac{1}{(t+1)^\alpha}$ and an averaging function $u_t = \frac{1}{(t+1)^\beta}$, with $\alpha, \beta \geqslant 0$, it follows that $\beta \geqslant \alpha$. Recall that the covariance matrix $\Sigma(X_t) \preccurlyeq \Sigma$ is bounded. From Theorem 16, the terms that forgets the initial conditions behaves as $\mathcal{O}(\frac{1}{(t+1)^{1-\beta}})$, and the variance term as $\mathcal{O}(\frac{1}{(t+1)^\alpha})$. Both terms converge at same rate for $\alpha = \beta = \frac{1}{2}$ (up to $\log(t+1)$). In discrete time, similar convergence results have been derived for SGD under Polyak-Ruppert averaging [Bach and Moulines, 2011, Theorem 6].

Under the strong convexity assumption ($\mu > 0$), polynomial convergence can be reached for the term that contains the initial conditions. However, the variance term cannot converge faster than the step size. Recalling that the covariance matrix $\Sigma(X_t) \preccurlyeq \Sigma$ is bounded, if $u_t = (t+1)^\beta$ and $h_t = (t+1)^{-\alpha}$ with $\alpha, \beta \geqslant 0$, the condition $\left(\frac{\int_0^t u_s ds}{2h_t}\right)^{\cdot} \leqslant 2\mu u_t$ implies that $\alpha = 0$. The variance term is then exactly equal to the step size $h_t = h_0$. Hence, weighted averaging allows

a better convergence for the terms containing initial conditions, but does not play a role in the variance term. To conclude, weighted averaging does not improve convergence results obtained under Polyak-Ruppert averaging. The trade-off between the forgetting of the initial conditions and the noise term mostly relies on step sizes.

We have analyzed convergence of SGD together with averaging techniques using approximating SDEs (2.24). Continuous-time analyses lead to similar convergence results, while benefiting from simpler formulations and fewer assumptions especially on step sizes. Using this approach, we analyzed the trade-off between nonuniform averaging and step sizes, paving the way to a better understanding of averaging techniques. In the next session, we explore new convergence analyses for stochastic accelerated methods.

2.4 Accelerating the gradient flow

For both stochastic and deterministic models, we have retrieved known convergence results for continuous-time models approximating optimization methods. In this section, we provide convergence guarantees for a family of second-order gradient flows, including in particular AGF (2.17).

In the deterministic setting, convergence of gradient descent was improved using a momentum. In this section, let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $\gamma > 0$ be constant step sizes. An approximating SDE (for order-1 weak approximations) for Nesterov's accelerated gradient is given by Li et al. [2019, Theorem 16, Section 4.4]:

$$d^2X_t + \frac{3}{t}dX_t + \nabla f(X_t)dt + \sqrt{\gamma\Sigma(X_t)}dB_t = 0. \tag{2.29}$$

As for SGD, the function $\mathcal{V}(x, \dot{x}, t) = t^2(f(x) - f_\star) + 2\|(x - x_\star) + \frac{t}{2}\dot{x}\|^2$ obtained from Theorem 13 does not allow to conclude about convergence to a stationary point of trajectories X_t generated by the stochastic accelerated gradient flows (2.29). Using the bounded covariance assumption $\Sigma(X_t) \preccurlyeq \Sigma$ and applying Ito's formula to $\mathcal{V}(\cdot)$ along X_t:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{t^2} + \gamma \mathrm{Tr}(\Sigma) 3t.$$

In the following, we explore Polyak-Ruppert averaging together with diminishing step sizes, to analyze convergence of second-order SDEs.

2.4.1 Averaging does not preserve convergence rates

Averaging was a key to success for improving convergence of SGD (see Section 2.3). It is natural to wonder if averaging preserves the acceleration of Nesterov's gradient flow [Su et al., 2016]. Let us define the stochastic accelerated gradient flow under Polyak-Ruppert averaging defined by $\bar{X}_t = \frac{1}{t}\int_0^t X_s ds$:

$$\begin{aligned} d^2X_t + \frac{3}{t}dX_t + \nabla f(X_t)dt + \sqrt{\gamma\Sigma}dB_t = 0, \\ d\bar{X}_t = \frac{X_t - \bar{X}_t}{t}dt. \end{aligned} \tag{2.30}$$

The family of quadratic functions into consideration is given by:

$$\mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(X_t, \bar{X}_t, t) = \begin{pmatrix} a_t^{(1)} \\ a_t^{(2)} \end{pmatrix}^\top \begin{pmatrix} f(X_t) - f_\star \\ f(\bar{X}_t) - f_\star \end{pmatrix} + \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}^\top (P_t \otimes I_d) \begin{pmatrix} \dot{X}_t \\ X_t - X_\star \\ \bar{X}_t - X_\star \end{pmatrix}, \tag{2.31}$$

where $P_t \succcurlyeq 0$ and $a_t^{(1)}, a_t^{(2)} \geqslant 0$ are continuously differentiable functions.

Theorem 17. *Let $a_t^{(1)}, a_t^{(2)} \geqslant 0$ and $P_t \succcurlyeq 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and $(X_t, \bar{X}_t)$ be a trajectory generated by the stochastic accelerated gradient flow under Polyak-Ruppert averaging (2.30) with constant step sizes ($h_t = 1$). Let in addition $\mathcal{V} = \mathcal{V}_{a_t^{(1)}, a_t^{(2)}, P_t}(\cdot)$ be a quadratic function (2.31). Then it holds that:*

- *When $a_t^{(1)} = 0$, if the function $\mathcal{V}$ verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, \bar{X}_t, t) \leqslant \mathrm{Tr}(2p_t^{(11)}\gamma\Sigma(X_t))$, then $\mathcal{V} = 0$.*
- *When $a_t^{(2)} = 0$, the function*

$$\mathcal{V}(X_t, t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2a_t^{(1)}}\|a_t^{(1)}\dot{X}_t + \dot{a}_t^{(1)}(X_t - x_\star)\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \mathbf{E}\mathrm{Tr}(2a_t^{(1)}\gamma\Sigma(X_t))$, with $a_t^{(1)} \leqslant t^2$. Furthermore, it holds that $\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{2\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{1}{2a_t^{(1)}}\gamma\int_0^t a_s^{(1)}\mathbf{E}\mathrm{Tr}(\Sigma(X_s))ds$.

Proof. The proof follows from Theorem 14 (see Appendix 2.B.2). The first statement holds when considering a function $\beta_t \geqslant 0$ instead of $\frac{3}{t}$. □

Without further assumptions on the covariance Σ_t, the accelerated gradient flow under Polyak-Ruppert averaging (2.30) with constant step sizes admits no quadratic function that allows to forget the initial conditions while reducing the variance term. Therefore, Polyak-Ruppert averaging plays a different role in the stochastic accelerated gradient flow compared to the stochastic gradient flow approximating SGD (2.26).

2.4.2 Diminishing step sizes do not help preserve acceleration

Averaging under constant step sizes was not conclusive for finding a convergence guarantee for Nesterov's accelerated gradient flow with a diffusion term. We consider a second-order non-autonomous stochastic gradient flow with time-varying step sizes $h_t \geqslant 0$,

$$d^2X_t + \beta_t dX_t + h_t\nabla f(X_t)dt + h_t\sqrt{\gamma\Sigma(X_t)}dB_t = 0. \tag{2.32}$$

To study the convergence of (2.32), we generate a quadratic function within the class (2.14).

Theorem 18. *Let $h_t \geqslant 0$, $\gamma > 0$ and $a_t \geqslant 0$ be continuously differentiable with respect to time. Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function and X_t be a trajectory generated by the stochastic second-order gradient flow (2.32). Assuming $\dot{a}_t \leqslant a_t\frac{2}{3}(\beta_t + \frac{1}{2}\frac{\dot{h}_t}{h_t})$ and $t \to \frac{(\dot{a}_t)^2}{2h_t a_t}$ a decreasing function, the function*

$$\mathcal{V}(X_t) = a_t(f(X_t) - f_\star) + \frac{1}{2h_t a_t}\|a_t\dot{X}_t + \dot{a}_t(X_t - x_\star)\|^2,$$

verifies $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t, t) \leqslant \frac{1}{4}\mathbf{E}\mathrm{Tr}(a_t h_t\Sigma(X_t))\gamma$.

Proof. This result is obtained by extending the LMI for ODEs from Theorem 14 and Corollary 3 to time-varying step sizes. □

To compare to previous results, we consider parametrized step sizes $h_t = \frac{1}{(t+1)^\alpha}$ and SDEs with $\beta_t = \frac{b}{t}$, where $\alpha, b > 0$. We derive a convergence bound using Theorem 18 together with Theorem 14, that $\dot{a}_t \leqslant \beta\frac{a_t}{t}$, where $\beta \leqslant \min(\frac{2b-\alpha}{3}, 2 - \alpha)$:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\beta^2}{t^\beta}\|x_0 - x_\star\|^2 + \frac{\gamma}{4t^\beta}\int_0^t \frac{s^\beta}{(s+1)^\alpha}\mathbf{E}\mathrm{Tr}(\Sigma(X_s))ds.$$

On the one hand, the smaller the step sizes, the better the convergence for the term that contains the initial conditions. On the other hand, under bounded covariance $\Sigma(X_t) \preccurlyeq \Sigma$, the variance term behaves as $\mathcal{O}(\frac{1}{t^\beta})$ if $\beta \leqslant \alpha - 1$, and as $\mathcal{O}(\frac{1}{t^{\alpha-1}})$ otherwise (convergence requiring then $\alpha \leqslant 1$ and $\beta \leqslant 1$). For Nesterov's accelerated gradient flow with $b = 3$, we have $\beta = 2 - \alpha \leqslant \alpha - 1$, and therefore $\alpha \geqslant \frac{3}{2}$. Taking $\alpha = \frac{3}{2}$, a convergence bound is given by:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{9}{4\sqrt{t}}\|x_0 - x_\star\|^2 + \frac{\log t}{\sqrt{t}}\gamma \mathrm{Tr}(\Sigma).$$

We retrieve the result from Corollary 5 for the SDE approximating SGD under Polyak-Ruppert averaging. Yet this result requires smaller step sizes ($\alpha \geqslant \frac{3}{2}$). It does not seem possible to accelerate SGD with diminishing step sizes. Ghadimi and Lan [2015, Corollary 3] proved a convergence bound for a stochastic accelerated gradient method with $\beta = 2$ and $\alpha = \frac{1}{2}$, that we do not retrieve. Yet, in their approach, functions are minimized over a compact convex domain, whereas our approach focuses on an unbounded domain.

2.5 Conclusion and future work

We have developed a systematic approach for finding quadratic Lyapunov functions for families of ODEs and SDEs approximating SGD. Verifying such a Lyapunov function is cast as verifying the feasibility of a small-sized LMI. From this formulation, it is possible to efficiently search for quadratic Lyapunov functions for arriving to convergence bounds. While we retrieve convergence guarantees similar to those of discrete-time systems, continuous-time models require less assumptions on the problem classes and can be analyzed through shorter proofs.

While obtaining guarantees for stochastic optimization methods might be tedious, the SDE approach allows for simpler analyses of the trade-off between the variance term and the term that forgets the initial conditions. A shortcoming of this approach is that this analysis does not include approximation guarantees between optimization methods and their continuous-time counterparts. In the deterministic setting, stability techniques are often developed to quantify this approximation efficiently [Gautschi, 2012]. In stochastic analysis, stochastic modified equation have been introduced by Li et al. [2017, Theorem 1] to better approximate SGD, and stochastic methods with momentum. For non-convex functions, some analyses have also been done by Shi et al. [2020]. However, these approximation theorems often require many assumptions on the class of functions, which we believe could be further simplified using computer-assisted proofs.

Concerning possible extensions, this work relies on a specific family of quadratic Lyapunov functions provided by (2.14) and (2.11). Whereas it was sufficient for our purposes, it turns out that it is possible to extend it by taking into account terms of the forms $\int_0^t f(X_s)ds$ or integrals of the quadratic terms that are used in our Lyapunov functions. While we did not consider those terms here, they could be useful for studying other methods or in other settings. The attentive reader might also have realized that the PEP technique for continuous-time systems can be applied without difficulty to differential [Bolte et al., 2010, Section 3.2] and monotone inclusions [Boţ and Csetnek, 2016, 2017] problems. For monotone inclusions, interpolation results for casting the PEPs as semidefinite programs can be found in Ryu et al. [2020]. Finally, we note that it is still not clear how to use PEP-related techniques for directly studying higher-order methods and assumptions (already appearing in the variance term in the stochastic setting), which are also common for continuous-time systems [Attouch et al., 2016], or equivalents in the monotone inclusion setting [Chavdarova et al., 2021, Bot et al., 2023]. The problem here is the lack of a clean performance estimation reformulation, beyond the somewhat indirect approach by De Klerk et al. [2020]. We leave those investigations for future work.

Appendices

2.A Proof for ODEs

2.A.1 Proofs for Theorem 11

Proof. Following the same procedure as for Theorem 10 under strong-convexity assumption, and given a quadratic Lyapunov function $\mathcal{V}_{a_t,c_t}(X_t,t) = a_t\left(f(x_t)-f_\star\right)+c_t\|X_t-x_\star\|_2^2$ with $a_t, c_t \geqslant 0$, the maximization problem can be formulated into a semidefinite program

$$\begin{aligned} 0 \geqslant \max_{G\succcurlyeq 0, F\in\mathbf{R}^2} & \; b_0^\top F + \operatorname{Tr}(A_0 G), \\ \text{subject to } & \; b_i^\top F + \operatorname{Tr}(A_i G) \geqslant 0,\ i\in\{1,2\} \end{aligned}$$

where $A_0 = \begin{pmatrix} \dot{c}_t & -c_t \\ -c_t & -a_t \end{pmatrix}$, $A_1 = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, $b_0 = \dot{a}_t[1,\ -1]^\top$ $b_1 = [-1,\ 1]^\top$ and $b_2 = [1,\ -1]^\top$, whose Lagrangian dual is given by the feasibility problem,

$$\min_{\lambda_t^{(1)},\lambda_t^{(2)}\geqslant 0} \ 0 \ \text{ s. t. } S = \begin{pmatrix} \dot{c}_t & -c_t+\frac{\lambda_t^{(1)}}{2} \\ -c_t+\frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preccurlyeq 0,\ \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

This formulation is exactly the LMI feasibility problem from the statement of Theorem 11.

□

2.A.2 Proof for Corollary 3

Proof. We prove the claim in the convex case; the strongly convex setting is deduced by assuming an exponential form for the parameters $a_t = ae^{t\tau}$, and $P_t = Pe^{t\tau}$, where $\tau > 0$ is a (linear) convergence rate to be determined.

From Theorem 14, verifying the Lyapunov function can be framed as verifying the feasibility of an LMI. That is, the desired Lyapunov function can be deduced from the existence of $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ such that:

$$\begin{aligned} S = \begin{pmatrix} -\frac{\mu}{2}(\lambda_t^{(1)}+\lambda_t^{(2)}) + \dot{p}_t^{(11)} & p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} \\ p_t^{(11)} - \beta_t p_t^{(12)} + \dot{p}_t^{(12)} & 2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)} & -p_t^{(22)} + \frac{a_t}{2} \\ -p_t^{(12)} + \frac{\lambda_t^{(1)}}{2} & -p_t^{(22)} + \frac{a_t}{2} & 0 \end{pmatrix} \preccurlyeq 0, \\ \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}. \end{aligned}$$

Because of the zero diagonal term in $S = (s_{ij})_{1\geqslant ij\geqslant 3} \preccurlyeq 0$, it follows that $p_t^{(22)} = \frac{a_t}{2}$ and $p_t^{(12)} = \frac{\dot{a}_t}{2}$ (otherwise, the submatrix $\begin{pmatrix} s_{22} & s_{23} \\ s_{23} & 0 \end{pmatrix} \preccurlyeq 0$ has a strictly nonpositive determinant, which is a contradiction).

Let us now assume that $\lambda_t^{(2)} = 0$, which leads to $\dot{a}_t = \lambda_t^{(1)}$. Because of the null entries, the matrix S can be reduced to a 2×2 semi-definite negative matrix. Such a matrix has nonpositive diagonal terms (since $s_{11}s_{22} \geqslant (s_{12})^2$ and $s_{11}+s_{22}\geqslant 0$), that are $\dot{p}_t^{(11)} \leqslant 0$ and $2(p_t^{(12)} - \beta_t p_t^{(22)}) + \dot{p}_t^{(22)}$. which simplifies into $\dot{a}_t \leqslant \frac{2}{3}\beta_t a_t$.

For all $\epsilon > 0$, after integration of $\dot{a}_t \leqslant \frac{2}{3}\beta_t a_t$ between ϵ and t, that $a_t \leqslant a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}$. Therefore, $a_t \leqslant \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}$. Recalling that P_t is positive semidefinite, its determinant is nonnegative

$p_t^{(11)}p_t^{(22)} - (p_t^{(12)})^2 \geqslant 0$, that is $p_t^{(11)} \geqslant \frac{(p_t^{(12)})^2}{p_t^{(22)}} = \frac{(\dot{a}_t)^2}{2a_t}$. After integration between 0 and t, we obtain the following bound on a_t: $\sqrt{a_t} \leqslant \sqrt{a_0} + \frac{\sqrt{p_0^{(11)}}}{2}t$.

In other words, $a_t \leqslant \min\left((\sqrt{a_0} + \frac{\sqrt{p_0^{(11)}}}{2}t)^2, \lim_{\epsilon\to 0} a_\epsilon e^{\int_\epsilon^t \frac{2}{3}\beta_s ds}\right)$. □

2.B Proofs for SDEs

2.B.1 Proof for Theorem 15

Proof. We rewrite the SDE into

$$\begin{pmatrix} dX_t \\ d\bar{X}_t \end{pmatrix} = \left(\begin{pmatrix} 0 & 0 \\ \frac{1}{t} & -\frac{1}{t} \end{pmatrix} \otimes I_d\right)\begin{pmatrix} X_t \\ \bar{X}_t \end{pmatrix} dt + \left(\begin{pmatrix} -h_t & 0 \\ 0 & 0 \end{pmatrix} \otimes I_d\right)\begin{pmatrix} \nabla f(X_t) \\ \nabla f(\bar{X}_t) \end{pmatrix} dt + \left(\begin{pmatrix} h_t(\gamma\Sigma(X_t))^{1/2} \\ 0 \end{pmatrix} \otimes I_d\right) dB_t$$

and denote $Y_t = \begin{pmatrix} X_t \\ \bar{X}_t \end{pmatrix}$.

We consider the quadratic function (2.27) $\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t, \bar{X}_t, t) = \mathcal{V}(Y_t, t)$. Applying Ito's formula, its derivative with respect to time is $\frac{d}{dt}\mathcal{V}(Y_t,t) = \frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt} + \frac{1}{2}\gamma h_t^2 \mathrm{Tr}[\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t,t)^\top \begin{pmatrix} \Sigma(X_t) \\ 0 \end{pmatrix}]$. By taking the expectation of Ito's formula, we have $\frac{d}{dt}\mathbf{E}\mathcal{V}(Y_t,t) = \mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}] + \frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t,t)^\top \begin{pmatrix} \Sigma(X_t) \\ 0 \end{pmatrix})$.

The variance term $\frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial Y_t^2}\mathcal{V}(Y_t,t)^\top \begin{pmatrix} \Sigma(X_t) \\ 0 \end{pmatrix})$ depends on the second derivative of f in the space variable (X_t), and on the covariance Σ_t. We do not take this term into account in the performance estimation frawework.

We formulate the problem of verifying a quadratic function as verifying the inequality $\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,\bar{X}_t,t) \leqslant \frac{1}{2}\gamma h_t^2 \mathbf{E}\mathrm{Tr}(\frac{\partial^2}{\partial X_t^2}\mathcal{V}(X_t,\bar{X}_t,t)^\top \Sigma(X_t))$ (that is $\mathbf{E}[\frac{\partial}{\partial t}\mathcal{V}(Y_t,t) + \frac{\partial}{\partial y}\mathcal{V}\frac{dY_t}{dt}] \leqslant 0$) holds for any twice continuously differentiable function $f \in \mathcal{F}_{0,\infty}$, and any trajectory $(X_t, \bar{X}_t)$ generated by the SDE under Polyak-Ruppert averaging (2.26).

This problem is equivalent with verifying that $\frac{d}{dt}\mathcal{V}(\tilde{Y}_t,t) \leqslant 0$ holds for any function twice continuously differentiable $f \in \mathcal{F}_{0,\infty}$ and any trajectory $\tilde{Y}_t = \begin{pmatrix} \tilde{X}_t & \tilde{\bar{X}}_t \end{pmatrix}^\top$ generated by deterministic gradient flow from the SDE (2.26) with $\gamma = 0$.

We follow the methodology developed in Section 2.2 for ODEs. We formulate the verification of such a Lyapunov function as the maximization problem

$$0 \geqslant \max_{X_t\in\mathbf{R}^d, d\in\mathbf{N},\ f\in\mathcal{F}_{0,\infty}} \frac{d}{dt}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t)$$
$$\text{subject to } \dot{X}_t = -h_t\nabla f(\bar{X}_t),\ \dot{\bar{X}}_t = \frac{X_t - \bar{X}_t}{t}dt.$$

This maximization problem can be formulated as the following SDP program

$$\min_{\lambda_t^{(i)}\geqslant 0,\ i\in\{1,\dots,6\}} \quad 0$$
$$\begin{pmatrix} \dot{p}_t^{(11)}+\frac{2p_t^{(12)}}{t} & \dot{p}_t^{(12)}+\frac{p_t^{(22)}-p_t^{(12)}}{t} & \frac{\lambda_t^{(6)}+\lambda_t^{(4)}-2h_tp_t^{(11)}}{2} & \frac{a_t^{(2)}}{2t}-\frac{\lambda_t^{(5)}}{2} \\ \dot{p}_t^{(12)}+\frac{p_t^{(22)}-p_t^{(12)}}{t} & \dot{p}_t^{(22)}-\frac{2p_t^{(22)}}{t} & -\frac{\lambda_t^{(6)}-2h_tp_t^{(12)}}{2} & -\frac{a_t^{(2)}}{2t}+\frac{\lambda_t^{(5)}+\lambda_t^{(3)}}{2} \\ \frac{\lambda_t^{(6)}+\lambda_t^{(4)}-2h_tp_t^{(11)}}{2} & -\frac{\lambda_t^{(6)}-2h_tp_t^{(12)}}{2} & -a_t^{(1)} & 0 \\ \frac{a_t^{(2)}}{2t}-\frac{\lambda_t^{(5)}}{2} & -\frac{a_t^{(2)}}{2t}+\frac{\lambda_t^{(5)}+\lambda_t^{(3)}}{2} & 0 & 0 \end{pmatrix} \preccurlyeq 0,$$
$$\dot{a}_t^{(1)}+\lambda_t^{(1)}+\lambda_t^{(5)}=\lambda_t^{(4)}+\lambda_t^{(6)},$$
$$\dot{a}_t^{(2)}+\lambda_t^{(2)}+\lambda_t^{(6)}=\lambda_t^{(3)}+\lambda_t^{(5)}+\dot{a}_t^{(1)}.$$

Note that the Gram matrix G and function value vector F to be introduced are different from the setting explored in Section 2.2, more precisely $G=Q^\top Q\succcurlyeq 0,\ Q=[X_t-x_\star,\bar{X}_t-x_\star,g_t,\bar{g}_t]$ and $F=(f(X_t),f(\bar{X}_t),f_\star)$.

The final inequality follows from Ito's formula. □

2.B.2 Proof for Theorem 17

Proof. From the reasoning from Theorem 11 and Theorem 15, worst-case guarantee can be formulated into a maximization problem:

$$\begin{aligned} \max_{f\in\mathcal{F}_{0,\infty},d\in\mathbf{N},X_t\in\mathrm{R}^d} \quad & \frac{d}{dt}\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}(X_t,\bar{X}_t,t), \\ \text{s.t.} \quad & d^2X_t+\beta_t dX_t+\nabla f(X_t)dt+\sqrt{\gamma\Sigma}dB_t=0, \\ & d\bar{X}_t=\frac{X_t-\bar{X}_t}{t}dt. \end{aligned}$$

A control of this quantity can be achivied with the following LMI, for $\lambda_t^{(i)}\geqslant 0,\ i\in\{1,\dots,6\}$ and where $*$ corresponds to the symmetric entries of the matrix, and where $\beta_t=\frac{3}{t}$,

$$\begin{pmatrix} \dot{p}_t^{(11)} & \dot{p}_t^{(12)}-\beta_tp_t^{(12)} & \dot{p}_t^{(13)}-\frac{p_t^{(13)}}{t} & -p_t^{(11)} & 0 \\ -2\beta_tp_t^{(11)} & +\frac{p_t^{(13)}}{t}+p_t^{(22)} & -p_t^{(12)}\beta_t+p_t^{(23)} & & \\ * & \dot{p}_t^{(22)}+2\frac{p_t^{(22)}}{t} & p_t^{(23)}-2\frac{2p_t^{(22)}}{t} & a_t^{(1)}-p_t^{(12)} & \frac{a_t^{(2)}}{2t}-\frac{\lambda_t^{(5)}}{2} \\ & & +\frac{p_t^{(33)}}{t} & +\frac{\lambda_t^{(4)}+\lambda_t^{(6)}}{2} & \\ * & * & \dot{p}_t^{(33)}-\frac{2p_t^{(33)}}{t} & -\frac{\lambda_t^{(6)}}{2}-p_t^{(13)} & \frac{\lambda_t^{(5)}+\lambda_t^{(3)}-a_t^{(2)}/t}{2} \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \end{pmatrix} \preccurlyeq 0,$$
$$\lambda_t^{(6)}+\lambda_t^{(4)}=\lambda_t^{(5)}+\lambda_t^{(1)}+\dot{a}_t^{(1)},\ \ \lambda_t^{(2)}+\lambda_t^{(6)}+\dot{a}_t^{(2)}=\lambda_t^{(3)}+\lambda_t^{(5)}.$$

Thus, $p_t^{(11)}=0$, and because P_t is positive semidefinite, $p_t^{(12)}=p_t^{(13)}=0$. If in addition $a_t^{(1)}=0$, the unique feasible Lyapunov function is $\mathcal{V}_{a_t^{(1)},a_t^{(2)},P_t}=0$. Otherwise, the LMI can be simplified to the LMI of Theorem 14 and the Lyapunov function follows from Corollary 3. □

2.C Strongly convex regularization

When minimizing convex functions over convex domains, regularization techniques can improve convergence of gradient descent, as stated in Lessard et al. [2016, Section 5.4] and Nesterov

[2012a, discussion column]. The goal if this appendix if to analyze strongly convex regularization in continuous-time using the Lyapunov framework. This technique allows in particular to use methods (resp. ODEs) adapted to strongly convex functions leading to an optimal complexity bound.

Let $f \in \mathcal{F}_{0,L}$ be L-smooth convex functions with $0 < L < \infty$, and X_t be trajectories generated by the gradient flow, and $t \to \mu_t$ be a nonnegative differentiable function. We call f_{μ_t} the regularized function of f defined by

$$f_{\mu_t}(X_t) = f(X_t) + \frac{\mu_t}{2}\|X_t - x_0\|^2,$$

with an optimal point $x_{\mu_t,\star}$. We notice that $f_{\mu_t}(x_\star) \geqslant f_{\mu_t}(x_{\mu_t,\star}) \geqslant f(x_{\mu_t,\star}) \geqslant f(x_\star)$.

Constant regularization. First, we assume a constant regularization term $\mu_t = \mu$.

Lemma 4. *Let X_t be a trajectory generated the gradient flow (2.3) originating from μ-strongly convex functions. Then, it holds that*

$$f(X_t) - f_\star \leqslant e^{-2\mu t}(f(x_0) - f_\star) + \frac{\mu}{2}\|x_0 - x_\star\|^2, \tag{2.33}$$

and $f_\mu(x_\star) \geqslant f_\mu(x_{\mu,\star}) \geqslant f(x_{\mu,\star}) \geqslant f(x_\star)$.

Proof. From Theorem 10, since $f_\mu(x_{\mu,\star}) \geqslant f(x_\star)$ and $f_\mu(x_0) = f(x_0)$, we have that:

$$f_\mu(X_t) - f_\mu(x_{\mu,\star}) \leqslant e^{-2\mu t}(f_\mu(x_0) - f_\mu(x_{\mu,\star})) \leqslant e^{-2\mu t}(f(x_0) - f(x_\star)).$$

Then, using the inequality $f_\mu(x_\star) \geqslant f_\mu(x_{\mu,\star})$, we have that:

$$\begin{aligned} f_\mu(X_t) - f_\mu(x_{\mu,\star}) &\geqslant f_\mu(X_t) - f_\mu(x_\star), \\ &= f(X_t) + \frac{\mu}{2}\|X_t - x_0\|^2 - f(x_\star) - \frac{\mu}{2}\|x_\star - x_0\|^2, \\ &\geqslant f(X_t) - f(x_\star) - \frac{\mu}{2}\|x_0 - x_\star\|^2. \end{aligned}$$

Rearranging all terms, we get the result. □

The analysis in Lemma 4 follows the proof discrete time [Lessard et al., 2016].

Assume now that $\|x_0 - x_\star\|^2 \leqslant R^2$. To reach an ϵ-accuracy, we can take $\mu = \frac{\epsilon}{R^2}$. The precision ϵ is achieved for $t \geqslant \log(\frac{LR^2}{2\epsilon})\frac{R^2}{2\epsilon}$ (as in discrete-time). Unfortunately, this complexity is worse than the complexity of the gradient flow originating from strongly convex functions for the objective $f(X_t) - f_\star$, that is $O(\frac{R^2}{2\epsilon})$. We will now adjust the regularization term μ_t depending on the time horizon.

Time-varying regularization. We assume now μ_t to vary with respect to time t. Lyapunov functions for this flow are taken in the class of quadratic functions:

$$\mathcal{V}_{a_t,c_t}(X_t, t) = a_t \cdot (f_{\mu_t}(X_t) - f_{\mu_t,*}) + c_t \cdot \|X_t - x_\star\|^2, \tag{2.34}$$

where c_t, a_t are continuously differentiable with respect to time. Verifying linear convergence guarantees using quadratic Lyapunov functions (2.34) can be cast as an LMI.

Theorem 19. *Let $\mu_t, a_t, c_t \geqslant 0$ be continuously differentiable functions with respect to time. The following assertions are equivalent:*

- *The inequality $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t, t) \leqslant 0$ is satisfied for all dimensions $d \in \mathrm{N}$, all functions $f \in \mathcal{F}_{0,\infty}$ and all trajectories X_t generated by the gradient flows (2.3).*

- *There exist* $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ *verifying*

$$\begin{pmatrix} \frac{\dot{\mu}_t}{2} a_t & -\frac{\dot{\mu}_t}{2} a_t & 0 \\ -\frac{\dot{\mu}_t}{2} a_t & \dot{c}_t + \frac{\dot{\mu}_t}{2} a_t - \frac{\mu_t}{2}(\lambda_t^{(1)} + \lambda_t^{(2)}) & -c_t + \frac{\lambda_t^{(1)}}{2} \\ 0 & -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preceq 0, \ \dot{a} = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

Proof. Again, the search for a Lyapunov function can be formulated as a minimization problem,

$$0 \geqslant \max_{f \in \mathcal{F}_{0,\infty}, d \in \mathbf{N}, X_t \in \mathrm{R}^d} \frac{d}{dt} \mathcal{V}_{a_t, c_t}(X_t, t), \\ \text{subject to } \dot{X}_t = -\nabla f_{\mu_t}(X_t).$$

To achieve convergence guarantees on f, we introduce the Gram matrix $G = P^\top P$, with $P = [x_0 - x_\star, \ X_t - x_\star, \nabla f_{\mu_t}(X_t)]$, allowing to exploit the strong convexity of the function f_{μ_t}. Noticing that,

$$\frac{d}{dt} \mathcal{V}_{a_t, c_t}(X_t, t) = \dot{a}_t (f_{\mu_t}(X_t) - f_{\mu,*} + \dot{c}_t \|X_t - x_\star\|^2 + a_t \frac{\dot{\mu}_t}{2} \|X_t - x_\star\|^2 \\ - 2a_t \|\nabla f_{\mu_t}(X_t)\|^2 + 2c_t \langle X_t - x_\star, \nabla f_{\mu_t}(X_t) \rangle.$$

We conclude by following the reasoning from Theorem 11. □

In Theorem 19, when taking $c_t = 0$, it follows that $a_t \leqslant a_0 e^{2 \int_0^t \mu_s ds}$, leading to a convergence guarantee:

$$f(X_t) - f_\star \leqslant e^{-2 \int_0^t \mu_s ds} (f(x_0) - f_\star) + \frac{\mu_t}{2} \|x_0 - x_\star\|^2.$$

When $\mu_t = \mu$ is a non-increasing function, we retrieve previous results from Equation (2.33). This formulation allows to chose μ_t as a function of the horizon time. For example, if $\mu_t = \frac{\alpha}{t+1}$, we retrieve the convergence guarantee for convex functions (Corollary 11),

$$f(X_t) - f_\star \leqslant \frac{1}{(t+1)^{2\alpha}} (f(x_0) - f_\star) + \frac{\alpha}{2(t+1)} \|x_0 - x_\star\|^2.$$

Let us consider SDEs approximating SGD (2.24) applied to regularized functions f_{μ_t}. It is possible to derive a convergence guarantee using Theorem 19:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant e^{-2 \int_0^t h_s \mu_s ds} (f(x_0) - f_\star) + \frac{\mu_t}{2} \|x_0 - x_\star\|^2 + \gamma \int_0^t \mathrm{Tr}(\Sigma(X_s)) h_s^2 e^{-2(\int_s^t h\mu)} ds.$$

Optimizing with the parametric forms $\mu_t = \frac{1}{t^\alpha}$, $h_t = \frac{1}{t^{1-\alpha}}$, convergence to a stationary point is obtained with $\alpha = \frac{1}{3}$, as obtained in Corollary 4.

To conclude, this analysis allows to choose a smoothing parameter adapted to the step size and to link continuously the function to its regularized version.

2.D Using PEPit for computing worst-case guarantees

We illustrate below how to compute a worst-case convergence guarantee using PEPit, an open-source `python` package [Goujaud et al., 2024] designed to simplify access to the PEP framework. In short, PEPit allows to write first-order methods (or continuous-time models) as one would implement them, while the package handles the SDP modeling part. It relies on standard solvers, such as CVXPY [Diamond and Boyd, 2016], SCS [O'Donoghue et al., 2016] and MOSEK [ApS, 2022]. The package already contains about 75 examples, that can be found in

the documentation https://pepit.readthedocs.io/en/0.3.2/. In this section, we provide a tutorial example of implementing a convergence guarantee for the gradient flow originating from strongly convex functions.

Let $f \in \mathcal{F}_{\mu,\infty}$ and let us consider the gradient flow:

$$\dot{X}_t = -\nabla f(X_t), X_0 = x_0 \in \mathbb{R}^d.$$

Our goal is to compute the largest value $\tau_\star$ verifying:

$$\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau_\star \mathcal{V}_{a,c}(X_t), \tag{2.35}$$

for $a = 1, c = 0$, that is for $\mathcal{V}_{a,c}(X_t) = f(X_t) - f_\star$. Recall the formulation as an optimization problem from (2.7), that let four main ingredients appear:

$$\begin{aligned} -\tau_\star = \max_{X_t \in \mathbb{R}^d, d \in \mathrm{N}} & -\langle \nabla f(X_t), \dot{X}_t \rangle, && \text{(performance measure)} \\ \text{such that } & f(X_t) - f_\star = 1, && \text{(initial condition)} \\ & \dot{X}_t = -\nabla f(X_t), && \text{(algorithm)} \\ & f \in \mathcal{F}_{\mu,\infty}. && \text{(class of functions)} \end{aligned}$$

Imports. First, we need to do the right `python` imports.

```
from PEPit import PEP
from PEPit.functions import StronglyConvexFunction
```

Initialization of PEPit. We first, we initialize a PEP object, which allows to manipulate the ingredients of the PEP (such as functions and iterates).

```
problem = PEP()
```

For the sake of the example, we pick some simple values for the problem class and algorithmic parameters, for which we perform the worst-case analysis below.

```
# Specify the strong convexity parameter
mu = 0.1
```

Specifying the problem class. Next, we specify our assumptions on the **class of functions** containing the function to be optimized, and instantiate a corresponding object. Here, we consider an μ-strongly convex function.

```
# Declare a mu-strongly convex function named "func"
func = problem.declare_function(
        StronglyConvexFunction,
        mu=mu), # Strong convexity param.
```

We also define $x_\star$ as an optimal point and $f_\star$ as the optimal value.

```
# Start by defining its unique optimal point xs = x_* and
# corresponding function value fs = f_*
xs = func.stationary_point()
fs = func.value(xs)
```

Gradient flow initialization. Third, we can instantiate the starting points for the gradient flow that we will run, and specify **an initial condition** on a point. To this end, a starting point X_t is introduced, and a bound on the initial distance between those points is specified as $f(X_t) - f(x_*) \leqslant 1$.

```
# Declare the starting point
xt = problem.set_initial_point()
gt, ft = func.oracle(xt)

# Set the initial condition
problem.set_initial_condition(ft - fs == 1)
```

Gradient flow implementation. In this fourth step, we specify the **algorithm** in a natural format. In this example, we simply use the iterates (which are PEPit objects) as if we had to implement gradient descent in practice using a simple loop. Gradient flow may be described as follows,

$$\dot{X}_t = -\nabla f(X_t). \tag{2.36}$$

```
# Implement the gradient flow
xt_dot = -gt
```

Setting up the performance measure. It is crucial for the worst-case analysis to specify the **performance metric** for which we wish to compute a worst-case performance. In this example, we wish to compute the worst-case value of $\frac{d}{dt}\mathcal{V}(X_t) = \langle \dot{X}_t, \nabla f(X_t) \rangle$, which we specify as follows.

```
# Set the performance metric to the function value accuracy
problem.set_performance_metric(gt * xt_dot)
```

Solving the PEP The last natural stage in the process is to solve the corresponding PEP. This is done via the following line, which will ask PEPit to perform the modelling steps and to call an appropriate SDP solver (which should be installed beforehand) to perform the worst-case analysis.

```
# Solve the PEP with verbose = 1
verbose = 1
pepit_tau = problem.solve(verbose=verbose)
```

Output. Running this code for $\mu = 0.1$, we obtain the following output.

```
(PEPit) Setting up the problem: size of the main PSD matrix: 3x3
(PEPit) Setting up the problem: performance measure is minimum of 1 element(s)
(PEPit) Setting up the problem: Adding initial conditions and general
constraints ...
(PEPit) Setting up the problem: initial conditions and general constraints
(1 constraint(s) added)
(PEPit) Setting up the problem: interpolation conditions for 1 function(s)
                function 1 : Adding 2 scalar constraint(s) ...
                function 1 : 2 scalar constraint(s) added
(PEPit) Setting up the problem: constraints for 0 function(s)
(PEPit) Compiling SDP
```

```
(PEPit) Calling SDP solver
(PEPit) Solver status: optimal (solver: SCS); optimal value: -0.200037839892888
```

Note that the size of the SDP is 3×3 instead of 2×2, which is due to the fact that we introduced a term for $\dot{X}_t$. Running this code for different values of μ allows to recover Figure 2.1a.

CHAPTER 3

Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and LASSO

Objectives

Greedy first-order methods, such as coordinate descent with Gauss-Southwell rule or matching pursuit, have become popular in optimization due to their natural tendency to propose sparse solutions and their refined convergence guarantees. In this work, we propose a principled approach to generating (regularized) matching pursuit algorithms adapted to the geometry of the problem at hand, as well as their convergence guarantees. Building on these results, we derive approximate convergence guarantees and describe a transition phenomenon in the convergence of (regularized) matching pursuit from underparametrized to overparametrized models.

This chapter is under minor revisions for publication in the *Journal of Machine Learning Research* [Moucer et al., 2024b].

Codes. All codes for numerical results are provided at `https://github.com/CMoucer/Geometry_Dependent_Matching_Pursuit`.

Contents

1.1 General context **1**
1.1.1 ML framework 1
1.1.2 Global outlines 2
1.2 First-order methods in convex optimization **3**
1.2.1 Basics of convex optimization 3
1.2.2 First-order methods 6
1.2.3 A continuous-time interpretation for gradient descent 10
1.3 Computer-assisted worst-case analyses **12**
1.3.1 Explaining example: the gradient flow 12
1.3.2 Discrete-time versions 15
1.4 Concentration inequalities **17**
1.4.1 Standard concentration inequalities 17
1.4.2 Concentration inequalities for the sum of independent random variables 18
1.4.3 Subgaussian properties 20

3.1 Introduction

Many natural problems from machine learning and data science take the form of an ℓ_1-regularized minimization problem:

$$\min_{\alpha \in \mathbb{R}^d} \{F(\alpha) + H(\alpha) \triangleq f(P\alpha) + \lambda \|\alpha\|_1\}, \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth strongly convex function, $P \in \mathbb{R}^{n \times d}$ and n, d respectively denote the number of samples and the dimension of the problem. Typically, in the vanilla least-squares regression problem, $H(\alpha) = 0$ and $F(\alpha) = f(P\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$, and P corresponds to the input data, $y \in \mathbb{R}^n$ to the labels, d to the number of features (or parameters) and n the number of observations. If in addition $H(\alpha) = \lambda\|\alpha\|_1$, Problem (3.1) is exactly the LASSO problem [Tibshirani, 1996], that belongs to more general variational problems appearing in Fenchel duality theory [Bauschke and Combettes, 2017, Section 15.3]. Problem (3.1) is often compared to its constrained counterpart,

$$\min_{\alpha \in \mathbb{R}^d} F(\alpha), \text{ such that } \|\alpha\|_1 \leqslant R, \tag{3.2}$$

where λ may be seen as the Lagrange multiplier associated to the constraint $\|\alpha\|_1 \leqslant R$ with $R > 0$. Problems (3.1) and (3.2) arise when looking for sparsity patterns, such as in signal processing where we aim for models depending on a small number of variables, or for trace-norm regularized problems, when looking for low-rank solutions [Dudik et al., 2012]. In particular, Problem (3.1) is a popular way to induce sparsity on the solution for a well-chosen range of λ. Thus, ℓ_1-penalization (or constraint) is strongly connected to sparsity and can be seen as a convex substitute for ℓ_0-penalization problems [Candès and Tao, 2005, Section 1.2] for performing feature selection.

First-order methods have become popular to solve optimization Problems (3.1) and (3.2), due to their low cost per iteration and to the limited accuracy requirements in machine learning [Bottou et al., 2018, Section 7 and 8]. Within these methods, different algorithms might be used, whose choice depends on the properties of functions F and H. For instance, a first-order method may rely on the gradient or the proximal operator [Parikh and Boyd, 2013] as in the proximal gradient method, or the linear minimization oracle as in the Frank-Wolfe algorithm [Jaggi, 2013] for the constrained version (3.2). These methods often benefit from convergence guarantees.

In the context of sparsity, traditional first-order methods, such as the proximal gradient, often identify sparse structures in an unkown time [Iutzeler and Malick, 2020, Section 5]. Boosting strategies (also known as matching pursuit) have been developed to ensure sparse representations of approximate solutions [Mallat and Zhang, 1993, Tropp, 2004]. At each iteration, a possibly new atom (also referred to as a weak-learner in the boosting literature, or a coordinate in the context of coordinate descent) is greedily selected as a best candidate among a set of atoms, and combined to past iterates. While boosting benefits from strong statistical properties [Tropp, 2004], from an optimization perspective, their convergence analyses often rely on extra statistical assumptions [Zhang, 2011a]. More recently, randomized and greedy coordinate descent methods have gained interest due to their low-cost per iteration even in high dimension [Nesterov, 2012b] and to their implicit induced sparsity [Beck and Tetruashvili, 2013, Fang et al., 2020].

Correspondences have been highlighted between first-order methods and boosting strategies for non-regularized minimization problems ($\lambda = 0$), leading to convergence guarantees independent of traditional statistical assumptions. For example, coordinate descent has been interpreted as matching pursuit [Locatello et al., 2018], as well as Frank-Wolfe algorithms [Jaggi, 2013, Locatello et al., 2017] for constrained Problems (3.2), by formulating them as minimizers of well-chosen quadratic upper approximations. These analyses strongly rely on a well-chosen geometry, characterized by a gauge function [Friedlander et al., 2014]. To our knowledge, this

comparison was only drawn for non-regularized problems for which $\lambda = 0$ [Locatello et al., 2018] and for constrained problems [Sun and Bach, 2022].

Problem (3.1) in $\mathbb{R}^d$ can be naturally formulated as an optimization problem in $\mathbb{R}^n$, letting the gauge geometry appear,

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^d, x \in \mathbb{R}^n} \; f(x) + \lambda \|\alpha\|_1, \text{ such that } x = P\alpha, \\ &= \min_{x \in \mathbb{R}^n} \; f(x) + \lambda \inf_{\alpha \in \mathbb{R}^d,\; x = P\alpha} \|\alpha\|_1, \\ &= \min_{x \in \mathbb{R}^n} f(x) + \lambda \gamma_{\mathcal{P}}(x), \end{aligned} \tag{3.3}$$

where the gauge function is defined as $\gamma_{\mathcal{P}}(x) \triangleq \inf_{\alpha \in \mathbb{R}^d,\; x = P\alpha} \|\alpha\|_1$ with $\mathcal{P} = \operatorname{conv}(\{\pm P_{:,i}, i = 1, \dots, d\})$ the centrally symmetric convex hull of the columns of P. Gauge functions may be seen as generalized versions of the ℓ_1-norm, providing a sparse representation $\alpha \in \mathbb{R}^d$ of a vector $x \in \mathbb{R}^n$ with respect to a set of atoms. Under some assumptions on P, the gauge function may be a norm, as we will see in Section 3.2. Let us for example take $\mathcal{P} = \operatorname{conv}(\pm e_i)$, then $\gamma_{\mathcal{P}}(x) = \|x\|_1$.

Due to the connection between optimizing in $\mathbb{R}^d$ and in $\mathbb{R}^n$, it is possible to derive algorithms adapted to one geometry or to the other, and to formulate geometry-adapted convergence guarantees. For the ℓ_1-geometry, Nutini et al. [2018b, Section 4] analyzed greedy coordinate descent by considering strong convexity with respect to the ℓ_1-norm, and formulated the strong convexity parameter as an optimization problem [Nutini et al., 2018b, Appendix 4.1]. More generally, d'Aspremont et al. [2018, Section 2] extended smoothness and strong convexity with respect to the gauge $\gamma_{\mathcal{P}}$, which led to formulations of the smoothness parameter as an optimization problem in the work of Sun and Bach [2022, Section 2.5]. These optimization problems are often hard to solve (yet, they have closed-form reformulation in some cases).

The main idea of this work is to propose a principled view on gradient boosting methods, that are obtained by minimizing a smoothness upper bound with respect to the ℓ_1-norm. This methodology leads to a new boosting strategy for regularized problems, that benefits from (sub)linear convergence properties. Unlike former methods, such as orthogonal matching pursuit under restricted isometry property (RIP) by Zhang [2011b], convergence analysis is performed without statistical assumptions on the data. Convergence guarantees let appear parameters characterizing the class of functions and the geometries of optimization problems (3.1) in $\mathbb{R}^d$ and (3.3) in $\mathbb{R}^n$, but remain mostly intractable. To this end, we compute *a priori* refined estimates of convergence rates for boosting methods applied to a particular least-squares problem. We develop two approaches for computing on the one hand deterministic estimates using SDP relaxations [Goemans and Williamson, 1995], and on the other hand high probability bounds using random matrix theory. As a result, we observe a transition phase in the convergence rate of gradient descent (resp. coordinate descent), depending on (n, d). Surprisingly, we conclude that for a fixed number of samples n, adding features (dimension d) in a random feature experiments improves their convergence, which may be compared to the double descent phenomenon [Belkin et al., 2019] for the generalization error. Building on these results, we experimentally highlight a transition phase for the proximal gradient and regularized matching pursuit on a LASSO problem, depending on the value for λ. Finally, we define an *ultimate method*, enjoying linear convergence both in the underparametrized ($n \gg d$) and in the overparametrized ($n \ll d$) regime, that is nonetheless not a boosting method (it may indeed add more than one atom per iteration).

3.1.1 Prior works

Boosting algorithms. Boosting strategies, also known as matching pursuit in signal processing, have been initiated in the context of sparse recovery [Mallat and Zhang, 1993], and

extended to the fitting of weak-learners with 'gradient boosting' techniques such as Adaboost by Freund and Schapire [1999]. Matching pursuit (MP) algorithms produce sparse combinations of atoms by picking a direction from a set of atoms using information on the gradient. Boosting algorithms are suited to both constrained models, with for example orthogonal matching pursuit [Sheng Chen and Luo, 1989, Tropp, 2004, Zhang, 2011b] or greedy algorithms [Tewari et al., 2011], as well as to unconstrained (penalized) optimization problems, with for example the vanilla boosting strategy of [Zhang, 2011a], that minimizes a well-chosen quadratic upper-bound. Recently, Locatello et al. [2017] have unified the framework for matching pursuit and Frank-Wolfe algorithms [Frank and Wolfe, 1956] leading to non-statistical convergence guarantees for matching pursuit.

Coordinate descent. Coordinate descent has gained interest due to the increasing access to large amounts of data, and thereby to the use of large-scale optimization models. Tseng [2001] opened the path to convergence guarantees for proximal coordinate descent on composite minimization problems [Tseng and Yun, 2009]. Nesterov [2012b] derived global guarantees for coordinate gradient descent applied on convex objectives, paving the way to families of randomized coordinate updates [Richtárik and Takáč, 2014], and greedy updates [Beck and Tetruashvili, 2013]. Yet, these analyses often lead to dimension-dependent convergence guarantees. Nutini et al. [2018b] provided the first convergence guarantee of greedy coordinate descent (or coordinate descent with Gauss-Southwell rule) without dependence in the dimension, formulating the update as the minimization of a smoothness upper bound with respect to the ℓ_1-norm. More precisely, they showed a significantly better performance of greedy coordinate descent compared to randomized coordinate descent. However, the analysis did not extend well to proximal coordinate descent, letting a dependence in the dimension appear in the convergence bound. This led to refined techniques such as the greedy update of Karimireddy et al. [2019], with dimension-independent convergence guarantees. Finally, these methods often present the benefit of an induced sparsity, that can be linked to the ℓ_1-norm. Locatello et al. [2017] interpreted steepest coordinate descent as a matching pursuit algorithm, where the atoms corresponds to the unitary directions. More precisely, steepest coordinate descent may be seen as the minimization of a smoothness upper bound with respect to the ℓ_1-norm. Considering gauge functions, coordinate descent can be extended to producing solutions sparse with respect to atoms, as Sun and Bach [2022] did with the generalized conditional gradient method [Bach, 2015].

Refined convergence guarantees. Sparse optimization often reveals a gap between theoretical convergence guarantees and observed behaviors. The LASSO has been widely studied for statistical recovery. From an optimizovation point of view, most of the analyses depend on the statistical recovery efficiency. For constrained optimization problems, [Zhang, 2011b] proposed a forward-backward greedy algorithm for which he derived convergence guarantees under RIP. Similarly, Agarwal et al. [2010] analyzed the proximal gradient and the projected gradient under restricted strong convexity and smoothness, that comes directly from restricted eigenvalue conditions [Raskutti et al., 2010], that appear for example for random Gaussian matrices. A recent focus on average-case analysis of optimization methods under random matrices was initiated by Pedregosa and Scieur [2020], coming from the convergence analysis of the simplex method [Borgwardt, 1987, Spielman and Teng, 2001]. On the contrary, other works improved global convergence guarantees considering well-chosen geometries. For separable quadratics, Nutini et al. [2018b, Section 4.1] have computed explicitly the strong convexity parameter in the ℓ_1-geometry. Generalizing unitary atoms from the ℓ_1-geometry to atoms, Sun and Bach [2022, Section 2.5] formulated smoothness and strong convexity with respect to gauge functions as optimization problems. However in most cases, since these parameters are hard to compute, both strong convexity and smoothness parameters remains formulated in the ℓ_2-norm. This often leads to additional terms in convergence guarantees, coming from the norm equivalence [Nutini et al., 2018b, Appendix 4] or from the geometry such as the pyramidal width [Lacoste-Julien and Jaggi, 2015] or the directional width [Locatello et al., 2017] in Frank-Wolfe techniques, or to the

Hoffman constant [Hoffman, 1957] for linear mappings with strongly convex functions [Necoara et al., 2019, Karimi et al., 2016, Guille-Escuret et al., 2021]. From a computational perspective, Massias et al. [2017] first introduced variants of coordinate descent with GS rule using working set strategy (focusing on certain relevant features), showing improvement over advanced solvers. Massias et al. [2018] introduced advanced coordinate descent solvers that incorporate screening (eliminating irrelevant features), working set techniques and stopping time strategies, which were later improved by Bertrand et al. [2022].

3.1.2 Assumptions

Convex optimization framework. In this work, functions f into consideration are convex, differentiable and Problem (3.3) admits at least one global minimizer $x_\star \in \mathbb{R}^n$. Functions $F(\cdot) = f(P\cdot) : \mathbb{R}^d \to \mathbb{R}$ benefit from the same properties. We restrict ourselves to the analysis of first-order methods (linear combinations of past iterates and gradients).

In this chapter, functions f may be smooth with respect to a generic norm $\|\cdot\|_{\mathbb{R}^n}$, if they verify for all $x, y \in \mathbb{R}^n$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L^f}{2}\|y - x\|^2_{\mathbb{R}^n}. \tag{3.4}$$

Functions $F(\cdot) = f(P\cdot)$ are therefore smooth with respect for any norm $\|\cdot\|_{\mathbb{R}^d}$ with $L^F \leqslant L^f L^{\mathcal{P}}$, where L is defined such that for all $\alpha, \beta \in \mathbb{R}^d$, $\|P(\alpha - \beta)\|^2_{\mathbb{R}^n} \leqslant L^{\mathcal{P}}\|\alpha - \beta\|^2_{\mathbb{R}^d}$, that is $L^{\mathcal{P}} = \sup_{\|\beta\|_{\mathbb{R}^d} \leqslant 1} \|P\beta\|^2_{\mathbb{R}^n}$. For least-squares, functions F are exactly smooth with $L^F = L^f L^{\mathcal{P}}$. In addition, functions f are strongly convex with respect to a norm $\|\cdot\|_{\mathbb{R}^n}$, if for all $x, y \in \mathbb{R}^n$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu^f}{2}\|y - x\|^2_{\mathbb{R}^n}. \tag{3.5}$$

Functions $F(\cdot) = f(P\cdot)$ do not always inherit strong convexity. For example, for least-squares, functions F are not strongly convex as soon as the number of samples n is lower than the dimension d. The 'natural' strong convexity parameter of functions F is given by $\mu^F = \mu^f \mu^{\mathcal{P}}$, with $\mu^{\mathcal{P}} = \inf_{\|\beta\|_{\mathbb{R}^d} \geqslant 1} \|P\beta\|^2_{\mathbb{R}^n}$ and may indeed be zero. As we will see in Section 3.2.4, F however inherits the Łojasiewicz property with parameter $\mu^F_L > 0$, such that for all $\beta \in \mathbb{R}^d$,

$$\frac{1}{2}\|\nabla F(\beta)\|^2_{\mathbb{R}^d,\star} \geqslant \mu^F_L (F(\beta) - F_\star), \tag{3.6}$$

where $\|\cdot\|_{\mathbb{R}^d,\star}$ is the dual norm for $\|\cdot\|_{\mathbb{R}^d}$.

Random matrices. A part of this work is devoted to approximating the strong convexity and smoothness parameters of f and F. We consider on the one hand relaxed formulations for strong convexity and smoothness parameters with respect to the data P (i.e., geometry). On the other hand, we propose high probability bounds of these parameters, relying on random matrix theory. Random matrices often appears in statistical assumptions, such as with restricted isometry property [Candès and Tao, 2005] or the restricted eigenvalue condition [Raskutti et al., 2010]. In the machine learning literature, random matrices appear in average-case analysis for quadratics [Pedregosa and Scieur, 2020] with the Marchenko-Pastur distribution [Marchenko and Pastur, 1967], or when studying the double descent phenomena for the generalization error [Belkin et al., 2019, Mei and Montanari, 2022, Bach, 2023a] for Gaussian data. Most of the time, these analyses let two regimes appear, depending (among others) on the number of samples n and the dimension d. Throughout this work, we thus consider two regimes depending on the linear mapping structure $P \in \mathbb{R}^{n \times d}$: the *underparametrized* (respectively *overparametrized*) regime, characterized by matrices $P \in \mathbb{R}^{n \times d}$ for which $n \geqslant d$ (resp. $d \geqslant n$) and $P^\top P$ (resp. $PP^\top$) is invertible. Note that

the invertibility of $PP^\top$ (resp. $P^\top P$) in the overparametrized (resp. underparametrized) regime can be obtained by adding sufficiently random noise. More assumptions on P and $P^\top P$ will be made across this study.

3.2 A transition phase for linear regression

We begin with the study of a linear regression problem, where problem (3.7) is a special case of the optimization Problem (3.1) with $\lambda = 0$,

$$\min_{\alpha \in \mathbb{R}^d} \{F(\alpha) = f(P\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2\}, \tag{3.7}$$

where $P \in \mathbb{R}^{n \times d}$, and n, d respectively denotes the number of samples and the dimension.

In this section, we focus on describing the convergence regimes of gradient descent in the ℓ_2-geometry and coordinate descent with the Gauss-Southwell (GS) rule [Karimi et al., 2016, Nutini et al., 2018b] in the ℓ_1-geometry. More precisely, we interpret gradient descent and coordinate descent as the minimizers of smoothness upper bound with respect to well-chosen norms, that is, as optimization problems in the geometry under consideration. This interpretation leads to linear convergence both in the underparametrized and the overparametrized regime, letting smoothness and strong convexity parameters appear, that are adapted to the geometry. For characterizing convergence properties of these methods, we provide estimates of these quantities. A first technique developed in this work is based on an SDP relaxation, and leads to deterministic estimates. A second technique, inspired from statistical assumptions and average-case analysis, leads to high probability bounds under statistical assumptions on the data. These estimates let a transition phase appear between the underparametrized and overparametrized regimes, that we illustrate in particular in a random feature experiment. Finally, we interpret coordinate descent as a matching pursuit algorithm depending on the geometry P.

First, let us compute estimates of smoothness and strong convexity parameters by formulating their computation as optimization problems in a generic norm for the least-squares minimization (3.7). In this context, f is $\frac{1}{n}$-smooth $\frac{1}{n}$-strongly convex with respect to the norm $\|\cdot\|_2$. Thus, F is L^F-smooth with respect to an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^d$, with $L^F = \frac{1}{n}\sup_{\|\beta\|^2 \leqslant 1} \|P\beta\|_2^2$. In addition, the function is (possibly) μ^F-strongly convex, with a parameter μ^F explicited in Lemma 5 and possibly equal to 0 (especially when dimension $d < n$).

Lemma 5. *Let $F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$, where $P \in \mathbb{R}^{n \times d}$. Then, F is μ^F-strongly convex with respect to a norm $\|\cdot\|$ with,*

$$\mu^F = \frac{1}{n} \inf_{\|\beta\|^2 \geqslant 1} \|P\beta\|_2^2 \qquad \text{and} \qquad \frac{1}{\mu^F} = n \sup_{\|P\beta\|^2 \leqslant 1} \|\beta\|_2^2.$$

Proof. Let us recall the definition for strong convexity (3.5), for all $\alpha, \nu \in \mathbb{R}^d$, $F(\alpha) \geqslant F(\nu) + \langle \nabla F(\nu), \alpha - \nu\rangle + \frac{\mu^F}{2}\|\alpha - \nu\|^2$. Since F is a quadratic, the left-hand side of the inequality can be rephrased into, for all $\beta \in R^d$, $\|P\beta\|_2^2 \geqslant \mu^F\|\beta\|^2$, from which both formulations follow. □

In Lemma 5, we formulate μ^F, the strong convexity parameter for F, as a nonconvex minimization problem, with a convex objective and concave constraints. Such a problem is usually costly to solve. The function F also verifies the Łojasiewicz inequality (3.6) with μ^{F_L}. Again μ_L^F is formulated as an optimization problem.

Lemma 6. *Let $F : \alpha \in \mathbb{R}^d \mapsto \frac{1}{2n}\|P\alpha - y\|_2^2$, where $P \in \mathbb{R}^{n \times d}$. Then, F verifies the Łojasiewicz*

inequality (3.6) with respect to a (dual) norm $\|\cdot\|_\star$, with

$$\mu_L^F = \frac{1}{n} \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\star^2 \qquad \text{and} \qquad \frac{1}{\mu_L^F} = n \sup_{\|P^\top P\beta\|_\star^2 \leqslant 1} \|P\beta\|_2^2.$$

Proof. Let $\alpha_\star$ be a minimizer for F such that $y = P\alpha_\star$ and $F(\alpha_\star) = 0$. Then, it holds that for all $\alpha \in \mathbb{R}^d$: $\|\nabla F(\alpha)\|_\star^2 = \frac{1}{n^2}\|P^\top P(\alpha - \alpha_\star)\|_\star^2 \geqslant \frac{1}{n} \frac{\|P^\top P(\alpha-\alpha_\star)\|_\star}{\|P(\alpha-\alpha_\star)\|_2^2} \frac{1}{n}\|P(\alpha-\alpha_\star)\|_2^2$. Thus, $\frac{\|\nabla F(\alpha)\|_\star^2}{F(\alpha)} \geqslant \mu_L^F = \frac{1}{n} \inf_{\beta\in\mathbb{R}^d, \|P\beta\|_2^2 \geqslant 1} \|P^\top P\alpha\|_\star$, which is the Łojasiewicz inequality with parameter μ_L^F. □

Again in Lemma 6, μ_L^F is formulated as a (nonconvex) minimization problem. The two quantities μ^F and μ_L^F are compared in Lemma 7, with equality in the underparametized regime in which $P^\top P$ is invertible.

Lemma 7. *Let $F = \frac{1}{2n}\|P\alpha - y\|_2^2$. Then, we have that $\mu_L^F \geqslant \mu^F$ for μ^F (resp. μ_L^F) defined in Lemma 5 (resp. Lemma 6). If $P^\top P$ is invertible, $\mu_L^F = \mu^F$.*

Proof. Let us consider the squared-root formulations of μ^F and μ_L^F given in Lemma 5 and Lemma 6.

$$\frac{1}{\sqrt{n\mu^F}} = \sup_{\|P\beta\|_2 \leqslant 1} \|\beta\| = \sup_{\|z\|_\star \leqslant 1, \|P\beta\|_2 \leqslant 1} \langle \beta, z\rangle,$$
$$\frac{1}{\sqrt{n\mu_L^F}} = \sup_{\|P^\top P\nu\|_\star \leqslant 1} \|P\nu\|_2 = \sup_{\|P^\top P\nu\|_\star \leqslant 1, \|P\beta\|_2 \leqslant 1} \langle P\beta, P\nu\rangle = \sup_{\|P^\top P\nu\|_\star \leqslant 1, \|P\beta\|_2 \leqslant 1} \langle \beta, P^\top P\nu\rangle.$$

Since $\mathrm{Im}(P^\top P) \subset \mathbb{R}^d$, we have $\frac{1}{\sqrt{n\mu^F}} \geqslant \frac{1}{\sqrt{n\mu_L^F}}$, and therefore $\mu_L^F \geqslant \mu^F$. In the special case where $P^\top P$ is invertible, $\mathrm{Im}(P^\top P) = \mathbb{R}^d$, and $\mu_L^F = \mu^F$. □

In the next sections, we see the role of these parameters in the convergence guarantees of gradient descent and steepest coordinate descent, both in the underparametrized and overparametrized regime. We then propose deterministic estimates for μ^F and μ_L^F, as well as high probability bounds based on a simple random model for P.

3.2.1 Gradient descent in the ℓ_2-geometry

We are interested in the convergence of gradient descent in the underparametrized and the overparametrized regimes. Assume $\mathbb{R}^d$ is equipped with the ℓ_2-norm. The function F is convex, L_2^F-smooth with respect to the norm ℓ_2, with $L_2^F = \frac{1}{n}\lambda_{\max}(P^\top P)$. A common interpretation of gradient descent with fixed step size $\gamma = \frac{1}{L_2^F}$ comes from the minimization of a quadratic (smoothness) upper bound on F:

$$\alpha_1 = \alpha_0 - \frac{1}{L_2^F}\nabla F(\alpha_0) = \alpha_0 - \frac{1}{L_2^F} P^\top(P\alpha_0 - y). \tag{3.8}$$

In the underparametrized regime, the function F is μ_2^F-strongly convex with respect to the ℓ_2-norm, with $\mu_2^F = \lambda_{\min}(\frac{P^\top P}{n}) > 0$. As a result, gradient descent (3.8) converges linearly. However, in the overparametrized regime in which $d \geqslant n$, $\mu_2^F = 0$, F is not strongly convex. Yet, gradient descent still converges linearly [Bolte et al., 2010], since quadratics benefit from the Łojasiewicz inequality, with $\mu_{L,2}^F = \frac{1}{n}\lambda_{\min}(PP^\top) > 0$.

Proposition 3. *Let F be convex, L_2^F-smooth with respect to the norm $\|\cdot\|_2$, be μ_2^F-strongly convex and verify a Łojasiewicz inequality with parameter $\mu_{2,L}^F$, with $0 \leqslant \mu_2^F \leqslant L_2^F$ and $0 \leqslant \mu_{2,L}^F \leqslant L_2^F$.*

Let $(\alpha_k)_{k\in\mathbb{N}}$ be generated by gradient descent in (3.8) starting from $\alpha_0 \in \mathbb{R}^d$. The sequence verifies:

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_2^F, \mu_{2,L}^F)}{L_2^F}\right)^k (F(\alpha_0) - F_\star),$$

where $\mu_2^F = \lambda_{\min}(PP^\top/n)$, $\mu_{2,L}^F = \lambda_{\max}(P^\top P/n)$ and $L_2^F = \lambda_{\max}(P^\top P/n)$

Proof. See appendix 3.B. □

Convergence speeds obtained in Proposition 3 depend on $\lambda_{\max}(P^\top P/n)$, $\lambda_{\max}(P^\top P/n)$ and $\lambda_{\min}(PP^\top/n)$. In the case where P is generated randomly, we can derive estimates of these extremal eigenvalues, avoiding a full computation of the extremal eigenvalues, and thus, an approximate convergence guarantee of the method. In the following, we consider random data P, with i.i.d. entries having the same variance, so that $P^\top P$ and $PP^\top$ have a limiting Marchenko-Pastur distribution [Marchenko and Pastur, 1967], whose extremal eigenvalues are known. This distribution generalizes the Wishart distribution of $P^\top P$ and $PP^\top$, obtained from Gaussian data P.

Theorem 20 (Limits of extreme eigenvalues - Theorem 5.11 [Bai and Silverstein, 2010])**.** *Assume $P \in \mathbb{R}^{n\times d}$, where each entry is an i.i.d. random variable with mean 0, variance σ^2, $\mathbb{E}[P_{i,j}^4] < +\infty$ and let $H = \frac{1}{n}P^\top P$. If $\frac{d}{n} \to r \in (0,\infty)$, then we have almost surely that*

$$\lambda_{\min}(H) \to \sigma^2(1-\sqrt{r})^2,$$
$$\lambda_{\max}(H) \to \sigma^2(1+\sqrt{r})^2.$$

Combining Theorems 20 with Proposition 3, we obtain natural estimates of the convergence properties in the underparametrized and the overparametrized regimes for random P.

Corollary 6. *Under the same assumptions than Theorem 20 and assuming $\mathbb{E}[\|P\|^4] < +\infty$, gradient descent with step size $\gamma = \frac{1}{L_2^F}$ converges linearly to the optimum. Then, if $\frac{d}{n} \to r \in (0,\infty)$, we have almost surely that*

- *in the underparametrized regime ($r \ll 1$): $1 - \frac{\lambda_{\min}(P^\top P)}{\lambda_{\max}(P^\top P)} \to 1 - \frac{(1-\sqrt{r})^2}{(1+\sqrt{r})^2}$,*
- *in the overparametrized regime ($r \gg 1$): $1 - \frac{\lambda_{\min}(PP^\top)}{\lambda_{\max}(P^\top P)} \to 1 - r\frac{(1-\sqrt{1/r})^2}{(1+\sqrt{r})^2}$.*

Proof. Applying the Marchenko-Pastur Theorem 20 with Proposition 3 leads to the result. □

Given that $\frac{d}{n} \to r$, we deduce approximate convergence guarantees from Corollary 6: in the underparametrized regime ($\frac{d}{n} \to r \ll 1$), we have $1 - \frac{\lambda_{\min}(P^\top P)}{\lambda_{\max}(P^\top P)} = 4\sqrt{r} + o(\sqrt{r}) \approx 4\sqrt{\frac{d}{n}}$, and in the overparametrized regime ($\frac{d}{n} \to r \gg 1$), $1 - \frac{\lambda_{\min}(PP^\top)}{\lambda_{\max}(P^\top P)} = 4\sqrt{1/r} + o(\sqrt{1/r}) \approx 4\sqrt{\frac{n}{d}}$. These approximate convergence rates should be to compared to the average-case analysis of Pedregosa and Scieur [2020] for least-squares problems, and to the polynomial-based analysis for convergence of gossip developed by Berthier et al. [2020]. Depending on the distribution under consideration, Scieur et Pedregosa developed average-case optimal accelerated methods, whose limit in the number of iterations happens to be the Polyak-Momentum [Scieur and Pedregosa, 2020]. Its worst-case convergence guarantee verifies $\frac{\sqrt{L_2^F}-\sqrt{\mu_2^F}}{\sqrt{L_2^F}+\sqrt{\mu_2^F}} \approx \sqrt{\frac{d}{n}}$, can be compared to Nesterov's accelerated gradient method [Nesterov, 1983] with $1 - \sqrt{\frac{\mu_2^F}{L_2^F}} \approx 2\sqrt{\frac{d}{n}}$ and to gradient descent with $1 - \frac{\mu_2^F}{L_2^F} \approx 4\sqrt{\frac{d}{n}}$. When considering their average-case guarantees, only a polynomial sublinear term is added [Paquette et al., 2023, Table 2] to the worst-case guarantee, without

major modifications in the linear term. In other words, Polyak-Momentum (resp. Nesterov's accelerated gradient method) converges four times (resp. twice) as fast as gradient descent with fixed step sizes.

We conclude from Corollary 6 that convergence of gradient descent depends on the degree of underparametrization (resp. overparametrization). The more independent samples (resp. features), the better the convergence. While the advantage of adding independent samples is well known for improving both learning and convergence speed, it appears that the larger the set of features, the better the convergence. We are now going the study approximate convergence guarantees of coordinate descent, that appears in the ℓ_1-geometry.

3.2.2 Gauss-Southwell coordinate descent in the ℓ_1-geometry

Similar to gradient descent, we study convergence guarantees of coordinate descent based on the Gauss-Southwell (GS) rule. The GS-rule can be obtained from a smoothness upper bound with respect to the ℓ_1-norm, as shown by Nutini et al. [2018b, Section 4], for all $\alpha_0, \alpha \in \mathbb{R}^d$,

$$F(\alpha) \leqslant F(\alpha_0) + \langle \nabla F(\alpha_0), \alpha - \alpha_0 \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_0\|_1^2. \tag{3.9}$$

From this inequality, we compute $L_1^F = \frac{1}{n} \max_{\alpha \in \mathbb{R}^d, \|\alpha\|_1 = 1} \|Pz\|_2^2 = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ (the maximization problem attains its optimum on the extremal point of the simplex). Gauss-Southwell coordinate descent follows by minimizing over $\alpha \in \mathbb{R}^d$, for a fixed $\alpha_0 \in \mathbb{R}^d$,

$$\begin{aligned} i_0 &= \arg\max_{k=1,\dots,d} |\nabla_{i_k} F(\alpha_0)|, \\ \alpha_1 &= \alpha_0 - \frac{1}{L_1^F} \nabla_{i_0} F(\alpha_0) e_{i_0}. \end{aligned} \tag{3.10}$$

As for gradient descent, its convergence speed depends on the parametrization regime. Depending on the (n, d), F may be μ_1^F-strongly convex, or verify the Łojasiewicz inequality with parameter $\mu_{1,L}^F$. Both μ_1^F and $\mu_{1,L}^F$ can be formulated as optimization problems for computing explicit estimates. It follows from the strong convexity characterization given in Lemma 5 with the norm $\|\cdot\|_1$, that $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$, and from Lemma 6 with norm $\|\cdot\|$ for the Łojasiewicz inequality $\mu_{1,L}^F = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$.

In the regimes under consideration, Proposition 4 states that coordinate descent converges linearly, as already proven by Karimi et al. [2016, Theorem 1].

Proposition 4. *[Karimi et al., 2016, Theorem 1] Let F be convex, L_1^F-smooth with respect to the norm $\|\cdot\|_1$, be μ_1^F-strongly convex and verify the Łojasiewicz inequality with $\mu_{1,L}^F$, where $0 \leqslant \mu_{1,L}^F \leqslant L_1^F$ and $0 \leqslant \mu_1^F \leqslant L_1^F$. Let (α_k) be generated by coordinate gradient descent (3.10) starting from $\alpha_0 \in \mathbb{R}^d$. The sequence verifies:*

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_1^F, \mu_{1,L}^F)}{L_1^F}\right)^k (F(\alpha_0) - F_\star)$$

Proof. See Appendix 3.B. □

The convergence guarantee provided in Proposition 4 depends on μ_1^F and $\mu_{1,L}^F$ and hence complicated to compute. Although $L_1^F = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ has a closed-form solution, $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$ and $\mu_{1,L}^F = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$ are formulated as nonconvex minimization problems.

In the following, we construct estimates to these quantities, so that they may be computed a priori. First, we provide SDP relaxations for the optimization problems defining μ_1^F and $\mu_{1,L}^F$,

that may differ from the exact solution. We thus propose to construct high probability bounds for μ_1^F and $\mu_{1,L}^F$, assuming randomly generated data.

SDP relaxations. Building on the formulation of μ_1^F and $\mu_{1,L}^F$ as optimization problems, we rephrase them into relaxed SDPs.

Proposition 5. *Let $P \in \mathbb{R}^{n\times d}$, and let us define $\mu_1^F = \frac{1}{n}\inf_{\|z\|_1^2\geqslant 1}\|Pz\|_2^2$ and $\mu_{1,L}^F = \inf_{\|P\beta\|_2^2\geqslant 1}\|P^\top P\beta\|_\infty^2$. Then the following inequality holds*

- *in the underparametrized regime, $\frac{1}{n\tilde{\mu}_1^F} = \sup_{X\succcurlyeq 0}\operatorname{Tr}((P^\top P)^{-1}X)$, s.t. $\operatorname{diag}(X)\leqslant 1$,*

$$1-\frac{\pi}{2}\frac{\tilde{\mu}_1^F}{L_1^F}\leqslant 1-\frac{\mu_1^F}{L_1^F}\leqslant 1-\frac{\tilde{\mu}_1^F}{L_1^F},$$

- *in the overparametrized regime, $\frac{1}{n\tilde{\mu}_{1,L}^F} = \sup_{X\succcurlyeq 0}\operatorname{Tr}(P^\top PX)$ s.t. $\|P^\top PXP^\top P\|_\infty\leqslant 1$,*

$$1-\frac{\mu_{1,L}^F}{L_1^F}\leqslant 1-\frac{\tilde{\mu}_{1,L}^F}{L_1^F},$$

where $L_1^F = \frac{1}{n}\max_{i=1,\dots,d}\|P_{:,i}\|_2^2$. In addition, we still have that $\tilde{\mu}_1^F\leqslant\tilde{\mu}_{1,L}^F$.

Proof. See Appendix 3.C.1. □

In Proposition 5, we find out SDP relaxations that yield an deterministic estimate for μ_1^F, and an exact lower bound for $\mu_{1,L}^F$. Yet, the larger n, d, the longer the computation of these SDPs.

High probability bounds. We now assume that P is randomly generated, as in the ℓ_2-geometry. Under subgaussian assumptions, we derive in Proposition 6 high probability bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F. More precisely, we prove that L_1^F concentrates around the variance σ^2, μ_1^F around $\frac{\sigma^2}{d}$ and $\mu_{1,L}^F$ around $\frac{\sigma^2}{n}$ with subgaussian tails.

Proposition 6. *Let $P\in\mathbb{R}^{n\times d}$, with $P_i\in\mathbb{R}^d$ i.i.d. subgaussian such that $\mathbb{E}[P_{i,j}]=0$, $\mathbb{E}[P_{i,j}^2]=\sigma^2$. There exists absolute constants $C, C_1, C_2, C_3, C_4, K>0$ such that,*

- *For all $u\geqslant 2K^2\sqrt{\frac{C_1\log(d)}{n}}$,*

$$\left(1+C_2K^2\frac{1}{\sqrt{n}}-t\right)^2\leqslant\frac{L_1^F}{\sigma^2}\leqslant\left(1+2K^2\sqrt{\frac{C_1\log(d)}{n}}+t\right)^2,$$

  holds with probability $1-e^{-\frac{C}{\sigma^2K^4}\min(u_1(t),u_2(t))}$ where $u_1(t)=\log(d)\sigma^2(t+\frac{C_2K^2}{\sqrt{n}})^2$ and $u_2(t)=d\sigma^2(t-2K^2\sqrt{\frac{C_1\log(d)}{n}})^2$.

- *For all $t\geqslant 0$, it holds with probability $1-2\exp(-t^2)$,*

$$\left(1-C_3K^2\left(\sqrt{\frac{d}{n}}+\frac{t}{\sqrt{n}}\right)\right)^2\leqslant\mu_1^F\frac{d}{\sigma^2}\leqslant\left(1+C_3K^2\left(\sqrt{\frac{1}{n}}+\frac{t}{\sqrt{dn}}\right)\right)^2.$$

- *For all $t\geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}$, it holds with probability $1-2\exp(-\min(t^2,u_2(t))$,*

$$\left(1-C_4K^2\left(\sqrt{\frac{n}{d}}+\frac{t}{\sqrt{d}}\right)\right)^2\leqslant\mu_{1,L}^F\frac{n}{\sigma^2}\leqslant\left(1+2K^2\sqrt{\frac{C_1\log(d)}{n}}+t\right)^2.$$

The constant $K > 0$ characterizes subgaussian vectors of P (and defined in Appendix 3.C.2).

Proof. See Appendix 3.C.2. □

Compared with Proposition 3, Proposition 6 provides concentration inequalities for L_1^F, μ_1^F and $\mu_{1,L}^F$ depending on dimension d, the variance σ^2, the number of samples n and absolute constants.

From Proposition 4, we have seen that coordinate descent with GS-rule (3.10) converges in function values with a rate $1 - \frac{\max(\mu_{1,L}^F, \mu_1^F)}{L_1^F}$. In the overparametrized regime (resp. underparametrized), we conclude in Coroallary 7 with limiting concentration of the convergence rate for large dimensions (resp. large number of samples).

Corollary 7. *Let $P \in \mathbb{R}^{n \times d}$, with $P_i \in \mathbb{R}^d$ i.i.d. subgaussian such that $\mathbb{E}[P_{i,j}] = 0$, $\mathbb{E}[P_{i,j}] = \sigma^2$. Then,*

- *in the underparametrized regime, when $n \to \infty$, the quantity $1 - \frac{\mu_1^F}{L_1^F}$ concentrates in $1 - \frac{1}{d} + O(\frac{1}{\sqrt{n}})$ with subgaussian tails,*
- *in the overparametrized regime, when $d \to \infty$ and $\frac{\log(d)}{n} \to 0$, the quantity $1 - \frac{1}{\mu_{1,L}^F}$ concentrates in $1 - \frac{1}{n} + O(\frac{1}{\sqrt{d}}) + O(\sqrt{\frac{\log(d)}{n}})$ with subgaussian tails.*

Proof. See the proof in Appendix 3.C.3. □

For large overparametrized models (resp. underparametrized), the convergence guarantee of coordinate descent with GS rule concentrates to $(1 - \frac{1}{n})$ (resp. $(1 - \frac{1}{d})$), that is independent of the dimension d (resp. of the number of samples). Note that the condition $\log(d) \ll n$ is indeed reasonable, since e^n grows quickly (when $n = 50$, $e^n \approx 5 \times 10^{21}$). A numerical comparison for the expected and exact lower bounds for μ_1^F and $\mu_1^{L,F}$ is provided in Appendix 3.B. Unlike the approximate convergence guarantees for gradient descent in the underparametrized regime (resp. overparametrized) detailed in Corollary 6, the convergence rate of coordinate descent with GS-rule improves up to $1 - \frac{1}{d}$ (resp. $1 - \frac{1}{n}$) for a large number of samples (resp. features).

As for gradient descent in the ℓ_2-geometry, we have formulated coordinate descent with GS rule as the minimization of the smoothness upper bound with respect to the ℓ_1-norm, leading its linear convergence in both the underparametrized and overparametrized regime. For a linear regression problem, nor the strong convexity parameter neither the Łojasiewicz in the ℓ_1-geometry benefit from a closed-form formulation (but it did in the ℓ_2-geometry). In a first approach, we approximate these quantities by SDPs, that may take longer computation in large models (either in the number of samples or the dimension). Instead of that, we consider randomly generated matrices P to approximate these parameters. Under subgaussian data, it appears μ_1^F, $\mu_{1,L}^F$ and L_1^F concentrate to there expectation with subgaussian tails. In the next section, we perform numerical experiments showing the transition phase between the two regimes.

3.2.3 A transition phase phenomenon: experimental results

We compare approximate convergence guarantees to numerical experimental convergence for gradient descent from Corollary 6 and for coordinate descent from Corollary 7. More precisely, we verify the expected transition phase in (n, d): in the overparametrized (respectively underparametrized) regime, the larger the dimension (resp. the number of samples), the better the convergence. To this end, we perform a few experiments on several datasets: least-squares problems obtained either with synthetic Gaussian vectors or from the Leukemia dataset or from random features, described below.

Synthetic quadratics. We consider several least-squares problems (3.7), where the number of samples $n = 50$ is fixed, and the number of dimension varies so that both the overparametrized and the underparametrized regimes are explored. In this model, the feature matrix P into consideration is generated such that $P_{:,i} \sim \mathcal{N}(0, I_d)$ are i.i.d, $\alpha_\star \in \{-1, 1\}^d$ has a sparsity (that is, the number of non-zero entries) equal to $s = 8 < d$, and $y = P\alpha_\star + \epsilon$, where $\epsilon_i \sim \mathcal{N}(0, \sigma)$.
Leukemia dataset. We consider the standard Leukemia dataset [Golub et al., 1999], where $n = 72$ and $d = 7129$. Again, we consider submatrices, so that the dimensions vary from the underparametrized regime to the overparametrized one. For each model, P has zero mean and features with unit variance.
Random features. We consider the example of random features for a fixed prediction model. We consider the regression model $\hat{a} = \arg\min_{a \in \mathbb{R}^d} \frac{1}{2n} \|y - f(P, a, \theta)\|_2^2$, where the family of models is given by $\mathcal{F}(\theta) = \{f(P, a, \theta) = \sum_{i=1}^d a_i \sigma(\langle \theta_{:,i}, P_{:,i} \rangle) = \phi_P(\theta)^\top a, a \in \mathbb{R}^d\}$, where $\theta \in \mathbb{R}^{d \times m} \sim \mathcal{N}(0, \nu^2)$, and $\sigma(\cdot) = \max(0, \cdot)$. In this experiment, we increase the number of features d (from 10 to 1000) while the initial data taken from the leukemia dataset is such that $n = 72$, $m = 200$ and $\theta_i \sim \mathcal{N}(0, I_m)$. In comparison to experiments on synthetic quadratics and on the leukemia dataset, the model does not vary in random features: all models converge to the same optimal solution y.

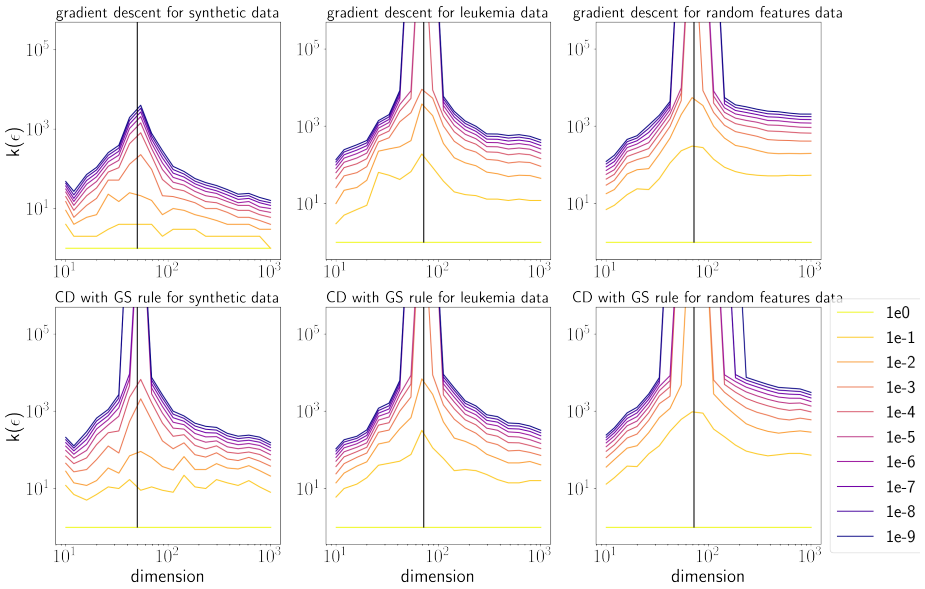


Figure 3.1: ϵ-curve for gradient descent (top) and coordinate descent with the GS-rule (bottom), for the three models: synthetic quadratics (on the left) with $n = 50$, the leukemia dataset (in the middle) with $n = 72$, a random feature model (on the right) with $n = 72$. The number of samples n is indicated by the black line.

In Figure 3.1, we plot the iteration number $k(\epsilon)$ at which a certain accuracy ϵ is reached for the three models describes above, both for gradient descent and coordinate descent with GS-rule. We consider accuracies $\epsilon = \{10^{-0}, \dots, 10^{-10}\}$ and we refer to these curves by ϵ-curves. For the three models, both steepest coordinate descent and gradient descent converge faster when $n \gg d$ (resp. $d \gg n$) in the underparametrized (resp. overparametrized) regime in Figure 3.1. For $n \approx d$, convergence slows down and tends to be sublinear, as expected from the theory for smooth convex functions. In other words, we observe a transition phenomenon for dimensions $d \approx n$. For the random feature models, a double descent phenomena was empirically highlighted by Belkin et al. [2019], and formalized by Mei and Montanari [2022]. For a fixed prediction model, as the number of features increase, the excess risk follows is U-shaped for

underparametrized optimization models and goes down for overparametrized models. As for the excess rick, we observe a transition at $d \approx n$ as well as a better precision for overparametrized models. Contrary to the generalization error, underparametrized models ($d \ll n$) performs well even when $\frac{d}{n} \to 0$ and are not U-shaped. We refer to this phenomenon as a transition phase for gradient and coordinate descent.

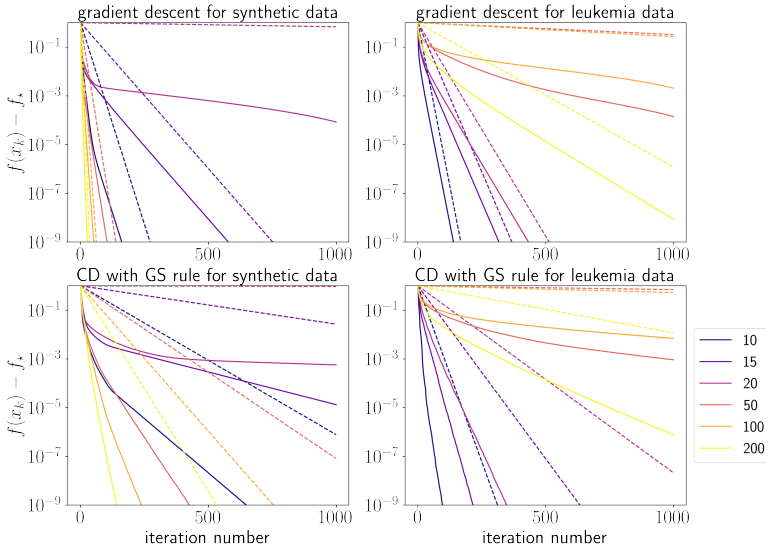


Figure 3.2: Convergence in function value for gradient descent and coordinate descent with GS rule, on synthetic quadratics ($n = 20$) and on the leukemia dataset ($n = 72$), for different values for d. Dashed lines: comparison to the approximate convergence guarantees from Corollary 6 for synthetic quadratics, and to high probability estimates for the leukemia dataset from Proposition 6.

In Figure 3.2, we compare the exact and approximated upper bound to the convergence guarantee in function value for gradient descent and coordinate descent with the GS rule. For gradient descent, the theoretical approximation guarantee from Corollary 6 matches the observed convergence behavior of gradient descent. For steepest coordinate descent, we compare its convergence in function values to the exact upper bound obtained from the SDP relaxation in Proposition 5 for 'small' values of d and n, and to its high probability bound otherwise. In both cases, we numerically recover that convergence is improved as the dimension increases.

3.2.4 Coordinate descent is an instance of matching pursuit

Coordinate descent, as well as gradient descent, converge linearly both in the underparametrized and overparametrized regime, as provided by Proposition 4, respectively due to strong convexity and the Łojasiewicz property. Given a certain structure on f, we prove that $F(\cdot) = f(P\cdot)$ inherits some regularity properties from f and that coordinate descent can be interpreted as a matching pursuit algorithm in either $\mathbb{R}^d$ or $\mathbb{R}^n$. Now, let us consider the more general formulation,

$$\min_{\alpha \in \mathbb{R}^d} F(\alpha) = f(P\alpha),$$

where f is L^f-smooth, μ^f-strongly convex and F is L_1^F-smooth with respect to the ℓ_1-norm.

Underparametrized regime. F is μ_1^F-strongly convex (since $P^\top P$ is invertible). The connection between coordinate descent with GS-rule (3.10) and matching pursuit was highlighted by Locatello et al. [2018]. Considering the set of unitary direction $\mathcal{A} = \text{conv}(\{\pm e_i, i = 1, \dots, d\})$,

that is the ℓ_1 unit ball, coordinate descent may be rewritten as a matching pursuit:

$$e_{i_0} \in -\mathrm{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)) = -\operatorname*{arg\,min}_{i=1,\dots,d} \nabla F(\alpha_0)^\top e_i,$$
$$\alpha_1 = \alpha_0 - \frac{1}{L_1^F} \nabla F(\alpha_0)^\top e_{i_0},$$

where $\mathrm{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)) = \inf_{z\in\mathcal{A}} \langle \nabla F(\alpha_0), z\rangle$. Steepest coordinate descent converges linearly from Proposition 4, with the same convergence guarantee as in the context of matching pursuit [Locatello et al., 2018, Theorem 5].

Overparametrized regime. F does not inherit strong convexity. Yet, for least-squares, F but does inherit some structure from f (see Lemma 6). We prove that coordinate descent can be interpreted as a matching pursuit algorithm in $\mathbb{R}^n$. Recall the gauge function, for $x \in \mathbb{R}^n$, $\gamma_{\mathcal{P}}(x) = \inf_{\alpha\in\mathbb{R}^d, x=P\alpha} \|\alpha\|_1$. Lemma 8 ensures $\gamma_{\mathcal{P}}(\cdot)$ is a norm in the overparametrized regime.

Lemma 8. *Let $\alpha \in \mathbb{R}^d \to P\alpha \in \mathbb{R}^n$ be a surjection in $\mathbb{R}^d$, and $\mathcal{P} = \mathrm{conv}(P)$ be centrally symmetric. The function $\gamma_{\mathcal{P}}(\cdot)$ is a norm, and its dual norm is $\gamma^\star_{\mathcal{P}}(\cdot) = \sup_{s\in\mathcal{P}} \langle s, \cdot\rangle = \|P^\top \cdot\|_\infty$.*

Proof. See Appendix 3.D.1. □

Let f be convex, L_2^f-smooth and μ_2^f-strongly convex with respect to the ℓ_2-norm. We define $L_{\mathcal{P}}^f = L_2^f \sup_{j=1,\dots,d} \|P_j\|_2^2$ and $\mu_{\mathcal{P}}^f = \mu_2^f \inf_{z\in\mathbb{R}^n} \|P^\top z\|_\infty^2$, such that $\|z\|_2^2 = 1$. Then, f is convex, $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. By definition, for all $x \in \mathbb{R}^d$, $L_2^f \|x\|_2^2 \leqslant L_{\mathcal{P}}^f \gamma_{\mathcal{P}}(x)^2$ and $\mu_2^f \|x\|_2^2 \geqslant \mu_{\mathcal{P}}^f \gamma_{\mathcal{P}}(x)^2$. Thus f is $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex with respect to $\gamma_{\mathcal{P}}$ (that is a norm in this regime).

As before, our estimates for smoothness (resp. strong convexity) parameters are obtained by an optimization problem. They appear to be closely related to the parameters for least-squares from Lemma 5 and 6. In the context of least-squares where $f(x) = \frac{1}{2n}\|x - y\|_2^2$ for $x \in \mathbb{R}^n$, we indeed have that $L_2^f = \mu_2^f = \frac{1}{n}$. For the ℓ_1-norm, that $L_{\mathcal{P}}^f = L_1^F$ as defined in (3.9), and $\mu_{\mathcal{P}}^f = \mu_{1,L}^F$ as soon as $PP^\top$ is invertible (which is the case here). Multiplying (3.10) by P, noticing that $\min_{e,\|e\|_1=1} \langle \nabla F(\alpha), e\rangle = \min_{p\in\mathcal{P}} \langle \nabla f(x), p\rangle$, coordinate descent with the GS-rule on F can be formulated as matching pursuit on f,

$$\begin{aligned} z_0 &\in \mathrm{LMO}_{\mathcal{P}}(\nabla f(x_0)), \\ x_1 &= x_0 - \frac{1}{L_{\mathcal{P}}^f} \langle \nabla f(x_0), z_0\rangle z_0. \end{aligned} \tag{3.11}$$

Let x_k be generated by matching pursuit (3.11), starting from $x_0 \in \mathbb{R}^n$ for $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex functions, then, Locatello et al. [2018, Theorem 5] proved linear convergence of the sequence with

$$f(x_k) - f_\star \leqslant \left(1 - \frac{\mu_{\mathcal{P}}^f}{L_{\mathcal{P}}^f}\right)(f(x_0) - f_\star). \tag{3.12}$$

By construction, since $x_k = P\alpha_k$, we have that $F(\alpha_k) - F_\star \leqslant (1 - \frac{\mu_{\mathcal{P}}^f}{L_{\mathcal{P}}^f})(F(\alpha_0) - F_\star) = (1 - \frac{\mu_{1,L}^F}{L_1^F})(F(\alpha_0) - F_\star)$. The same result could have been derived from Proposition 4 and the observation that strongly convex functions composed with a linear mapping verify a Łojasiewicz-inequality.

Lemma 9. *Let f be $\mu_{\mathcal{P}}^f$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. Then, F verifies a Łojasiewicz inequality with parameters $\mu_{\mathcal{P}}^f$, that is for all $\alpha \in \mathbb{R}^d$,*

$$\frac{1}{2}\|\nabla F(\alpha)\|_\infty \geqslant \mu_{\mathcal{P}}^f (F(\alpha) - F_\star).$$

Proof. Let $x \in \mathbb{R}^n$, $f_\star \geqslant f(x) - \sup_z \langle -\nabla f(x), y - x\rangle - \frac{\mu_{\mathcal{P}}^f}{2}\gamma_{\mathcal{P}}^2(y-x) \geqslant f(x) - (\frac{\mu_{\mathcal{P}}^f}{2}\gamma_{\mathcal{P}}^2(\cdot))^\star(-\nabla f(x)) \geqslant f(x) - \frac{1}{2\mu_{\mathcal{P}}^f}\|P^\top \nabla f(x)\|_\infty^2$. Since $F(\cdot) = f(P\cdot)$, the inequality is obtained by taking $x = P\alpha$ and since $\nabla F(\alpha) = P^\top \nabla f(P\alpha)$. □

Lemma 9 corresponds to the result of Karimi et al. [2016, Appendix B], that let a Hoffman constant appear (that is in their context equal to the smallest non-zero eigenvalue of P), as defined in Necoara et al. [2019, Section 3 and 4.1] by $\theta(P) = \max_{z, \|P^\top z\|_\infty = 1} \|z\|_2^2$. They indeed proved that F verifies a Łojasiewicz inequality, for all $\alpha \in \mathbb{R}^d$, $\frac{1}{2}\|\nabla F(\alpha_k)\|_2^2 \geqslant \theta(P)\mu^F(F(\alpha_k) - F_\star)$.

Depending on the parametrization regime, we have proven that coordinate descent may be formulated as a (possibly rebased) matching pursuit method. In the underparametrized regime on the one hand, since F inherits all regularity properties from f, the atoms are defined by the Euclidean basis and the matching pursuit is formulated in $\mathbb{R}^d$. On the other hand in the overparametrized regime, the introduction of a well-chosen gauge function $\gamma_{\mathcal{P}}$ allows to formulated coordinate descent as a matching pursuit algorithm in $\mathbb{R}^n$, and to perform a convergence analysis using the strong convexity assumption on f. Again, global values of the smoothness and strong-convexity parameters can be formulated as optimization problems depending on the gauge. The gauge let also appear how F inherits some structure from f. In the next sections, we generalize this framework for analyzing penalized linear models.

3.3 A transition phase for penalized linear models

We now consider the penalized linear model,

$$\min_{\alpha \in \mathbb{R}^d} \{G(\alpha) \triangleq f(P\alpha) + \lambda\|\alpha\|_1\}, \tag{3.13}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is L_2^f-smooth, μ_2^f-strongly convex, (and thus, the function $F : \mathbb{R}^d \to \mathbb{R}$ such that $F(\cdot) = f(P\cdot)$, is L_2^F smooth), $P \in \mathbb{R}^{n \times d}$, $\lambda > 0$ and where $H(\alpha) = \lambda\|\alpha\|_1$ is closed convex and proper. In this section, we derive a new matching pursuit algorithm for a ℓ_1-regularized model, that we compare to proximal coordinate descent with GS rule and to the proximal gradient descent. Building on the results of Section 3.2, we derive convergence guarantees depending on the properties of f and P, and notice a strong connection to the proximal coordinate descent with GS rule. Yet, in the overparametrized regime, neither the proximal gradient nor the regularized matching pursuit benefits from linear convergence. Instead of that, we describe experimentally the role of λ in the LASSO, as a continuous mapping between low-rank solutions and full-rank solution to the least-squares.

Proximal gradient descent. Proximal gradient descent, a.k.a. forward-backward (see e.g. [Combettes and Wajs, 2005]) was developed for such 'composite' convex optimization problems. Given a starting point $\alpha_0 \in \mathbb{R}^d$, each iterate is obtained by minimizing a smooth quadratic upper bound on F:

$$G(\alpha) \leqslant F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k\rangle + \frac{L_2^F}{2}\|\alpha_k - \alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{3.14}$$

Minimizing the right side of the inequality yields the proximal gradient method as follows:

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_2^F}{2} \|\alpha - \alpha_k\|_2^2 + \lambda \|\alpha\|_1.$$

The proximal gradient method converges sublinearly if F is smooth and convex, and linearly if F is in addition μ_2^F-strongly convex, such as in the underparametrized regime. Then, the sequence α_k starting from $\alpha_0 \in \mathbb{R}^d$ verifies [Taylor et al., 2018b, Theorem 2.1],

$$G(\alpha_k) - G_\star \leqslant \left(1 - \frac{\mu_2^F}{L_2^F}\right)^k (G(\alpha_0) - G_\star).$$

Coordinate descent. In practice, (randomized) coordinate gradient descent is widely used to avoid computing the full gradient (that costs $O(d)$), and is particularly suited to sparse regression problems. Nutini et al. [2018b] analyzed coordinate descent with the Gauss-Southwell selection rule, that tends to perform better than randomized coordinate descent.

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \nabla_{i_k} F(\alpha_k)(\alpha^{(i_k)} - \alpha_k^{(i_k)}) + \frac{L_2^F}{2}(\alpha^{(i_k)} - \alpha_k^{(i_k)})^2 + \lambda |\alpha^{(i_k)}|, \tag{3.15}$$

where $i_k = \arg\min_l \min_{t \in \mathbb{R}} \nabla_l F(P\alpha_k)(t - \alpha^{(l)}) + \frac{L_2^F}{2}(t - \alpha^{(l)})^2 + \lambda |t|$ corresponds to the GS rule. [Nutini et al., 2018b, Appendix 8] proved that coordinate descent with the Gauss-Southwell rule makes at least as much progress as randomized coordinate descent,

$$G(\alpha_{k+1}) - G_\star \leqslant \left(1 - \frac{\mu_2^F}{d L_2^F}\right) (G(\alpha_k) - G_\star).$$

A refinement, that let a sublinear dependence in the parameter μ_1^F appear, is mentioned in Nutini et al. [2018b, Appendix 8]. Coordinate descent with GS-rule is closely related to matching pursuit as for nonpenalized models, as detailed in Appendix 3.E, where we formulate this method as a 'nearly' matching pursuit algorithm.

In the following, we derive a matching pursuit procedure for ℓ_1-regularized problems (3.13), that we compare to classical boosting algorithm and coordinate descent with GS-rule. After that, we compute convergence guarantees for smooth (possibly) strong convex functions. Finally, we interpret the convergence regimes as a function of the penalty λ.

3.3.1 Regularized matching pursuit

We propose a new regularized matching pursuit algorithm based on the ℓ_1-geometry. The main idea is to replace the ℓ_2-norm in the minimization Problem (3.14) leading to the proximal gradient by a ℓ_1-norm. Let F be convex, L_1^F-smooth, as for coordinate descent with GS rule in the linear regression problem from Section 3.2. We define the penalized matching pursuit method starting from $\alpha_0 \in \mathbb{R}^d$ as the sequence minimizing smoothness with respect to the ℓ_1-norm at each iteration:

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\alpha\|_1. \tag{3.16}$$

Whereas the optimization steps in proximal gradient descent (3.14) and proximal coordinate descent with the GS rule (3.15) can be decomposed coordinate-wise, the function $\alpha \to \|\alpha\|_1^2$, that is not separable. Based on the same upper bound (3.16), Song et al. [2017, Algorithm 1] generalized greedy coordinate descent with the "SOft ThresOlding PrOjection" (SOTOPO) algorithm using a reweighted least-squares formulation (Appendix 3.A). However, their methods

is neither a coordinate-based method, nor a boosting method. We propose instead a regularized matching pursuit algorithm that draws a clean connection to boosting and proximal coordinate descent.

In the following, we formulate this optimization step (3.16) as a matching pursuit algorithm, that only calls for a linear minimization oracle. Using a variational trick detailed in Appendix 3.A to approach $\|\beta\|_1^2$, we begin by formulating Problem (3.16) starting from $\alpha_k \in \mathbb{R}^d$ as a separable optimization problem,

$$\begin{aligned} V_\star &= \min_{\beta \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \beta \rangle + \frac{L_1^F}{2} \|\beta\|_1^2 + \lambda \|\beta + \alpha_k\|_1, \\ &= \min_{\beta \in \mathbb{R}^d} \max_{z \geqslant 0} \langle \nabla F(\alpha_k), \beta \rangle - \frac{z^2}{2L_1^F} + z\|\beta\|_1 + \lambda \|\beta + \alpha_k\|_1, \\ &= \max_{z \geqslant 0} -\frac{z^2}{2L_1^F} + \min_{\beta \in \mathbb{R}^d} \sum_{i=1}^d \{\nabla_i F(\alpha_k)\beta^{(i)} + z|\beta^{(i)}| + \lambda|\beta^{(i)} + \alpha_k^{(i)}|\}. \end{aligned}$$

At the optimum, $z = L_1^F \|\beta\|_1$. The problem is now separable in each coordinate $\beta^{(i)}$, and can be reduced to an optimization problem in $z \geqslant 0$ in Lemma 10.

Lemma 10. *The optimization step (3.16) can be reformulated as*

$$V_\star = \max_{z_{\min} \leqslant z} \left\{ h(z) \triangleq -\frac{z^2}{2L_1^F} + \sum_{i \in I} \min\left(\lambda|\alpha_k^{(i)}|, -\nabla_i F(\alpha_k)\alpha_k^{(i)} + z|\alpha_k^{(i)}|\right) \right\}, \tag{3.17}$$

where $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ *and where* $I = \{i, \alpha_k^{(i)} \neq 0\}$ *is the set of active atoms.*

Proof. The function $\phi_i(z, \beta^{(i)}) = \nabla_i F(\alpha_k)\beta^{(i)} + z|\beta^{(i)}| + \lambda|\beta^{(i)} + \alpha_k^{(i)}|$ is lower bounded if $z \geqslant |\nabla_i F(\alpha_k)| - \lambda$ for all $i \in I$. Then, $\phi_i(z, \cdot)$ attains its minima in $\beta^{(i)} = 0$ with $\phi_i(z, 0) = \lambda|\alpha_k^{(i)}|$ or in $\beta^{(i)} = -\alpha_k^{(i)}$ with $\phi_i(z, -\alpha_k^{(i)}) = -\nabla_i F(\alpha_k)\alpha_k^{(i)} + z|\alpha_k^{(i)}|$ or in every possible value for $\beta^{(i)}$ if $z = \pm\nabla_i F(\alpha_k) - \lambda$ with $\phi_i(z, 0) = \lambda|\alpha_k^{(i)}|$. □

Lemma 10 leads to a convex constrained optimization problem in $\mathbb{R}^+$, whose objective is the sum of a quadratic and piecewise linear functions whose slope coefficients changing at $z_i = \lambda + \nabla_i F(\alpha_k)\alpha_k^{(i)}/|\alpha_k^{(i)}|$. Its constraints in $z_{\min}$ includes the LMO. By construction, the LMO given by $z_{\min}$ may correspond to several atoms β_j such that $j \in \arg\min_i |\nabla_i F(\alpha_k)|$. Since we aim at solving Problem (3.16) by constructing a solution as sparse as possible, we introduce Assumption 1.

Assumption 1. *The algorithm only selects one atom corresponding to the LMO, that is* $i_{\min} \in \arg\max_i |\nabla_i F(\alpha_k)|$, *such that* $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$.

In the following lemmas, we compute the minimum of h explicitly. Under Assumption 1, Lemma 11 first deals with the situation in which the objective is quadratic, that is for all $i \in I$, $z_i \geqslant z_{\min}$.

Lemma 11. *Let* $(z_\star, \beta_\star)$ *be a solution to (3.17), assume* $\{i, z_i \geqslant z_{\min}\} = \emptyset$ *and verify Assumption 1. Then* $z_\star = z_{\min}$, $\beta_\star^{(i_{\min})} = -\text{sign}(\nabla_{i_{\min}} F(\alpha_k)) \frac{z_{\min}}{L_1^F}$ *and* $\beta_\star^{(i)} = 0$ *for* $i \neq i_{\min}$.

Proof. The objective is quadratic and attains its minimum at $z_{\min} = L_1^F |\beta^{(i_{\min})}|$. □

In the context of Lemma 11 and Assumption 1, only the atom given by the LMO in $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ can be added to the set of active atoms. Now, we assume the objective is piecewise quadratic, that is $\mathcal{S} = \{i, z_i \geqslant z_{\min}\} \neq \emptyset$.

Lemma 12. *Let $z_\star, \beta_\star$ be a solution of (3.17) and assume $\mathcal{S} = \{i, z_i \geqslant z_{\min}\} \neq \emptyset$ and verify Assumption 1. There are four possible solutions to Problem (3.17),*

- *If $\frac{dh}{dz}(z_{\min}) \leqslant 0$, then $z_\star = z_{\min}$.*
  In addition, $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i \geqslant z_\star, \\ 0 & \text{if } z_i \leqslant z_\star, \\ -\operatorname{sign}(\nabla_{i_{\min}} F(\alpha_k)) \frac{z_{\min} - \sum_{i \in \mathcal{S}} |\alpha_k^{(i)}|}{L_1^F} & \text{if } i = i_{\min}. \end{cases}$
- *If there exists $k \in \mathcal{S}$ such that $\frac{dh}{dz}(z_k^+) \geqslant 0$ and $\frac{dh}{dz}(z_{k+1}^-) \leqslant 0$, then $z_\star \in]z_k, z_{k+1}[$. In addition,* $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i \geqslant z_\star, \\ 0 & \text{if } z_i \leqslant z_\star. \end{cases}$
- *If there exists $k \in \mathcal{S}$ such that $\frac{dh}{dz}(z_k^-) \geqslant 0$ and $\frac{dh}{dz}(z_k^+) \leqslant 0$ then $z_\star = z_k$. In addition,*
  $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i > z_\star, \\ 0 & \text{if } z_i < z_\star, \\ -\operatorname{sign}(\alpha_k^{(i)}) \left(\frac{z_k}{L_1^F} - \sum_{i, z_i > z_k} |\alpha_k^{(i)}| \right) & \text{if } i = k. \end{cases}$
- *If $\frac{dh}{dz}(z_{|I|}) > 0$, then for all $i \in I$, $z_\star > z_i$ and* $\beta_\star^{(i)} = \begin{cases} -\operatorname{sign}(\alpha_k^{(i)}) \frac{z_i}{L_1^F} & \text{if } i = |I|, \\ 0 & \text{otherwise.} \end{cases}$

Proof. The function h is strictly concave and piecewise quadratic on $[z_i, z_{i+1}]$. The solution to the optimization Problem (3.17) is thus obtained by studying the sign of $\frac{dh}{dz}(\cdot)$ at z_i^- and z_i^+. By construction of the solution given in the proof of Lemma 10, for all i such that $z_\star > z_i$ (resp. $z_\star < z_i$), then $\beta^{(i)} = 0$ (resp. $\beta^{(i)} = -\alpha_k^{(i)}$). Finally, we have $z_\star = L_1^F \|\beta_\star\|_1$ which gives the solution for $z_\star = z_{\min}$ or $z_\star = z_k$. □

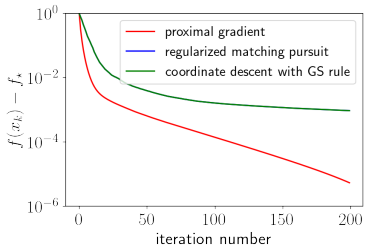
Lemmas 11 and 12 provides a closed form solution by calling only for the linear minimization oracle $\min_i |\nabla_i F(\alpha_k)|$, and performing $O(|I|)$ operations on the active atoms. From that, we deduce Algorithm 1. In short, at each iteration, Algorithm 1 performs one of the three possible actions: either one new atom is added (at most) by calling the $\mathrm{LMO}(\nabla F(\alpha_k)) = \arg\max_i |\nabla_i F(\alpha_k)| = \arg\max_{p \in \mathcal{P}} p^\top \nabla f(P\alpha_k)$ while some active atoms may be set to zero, or one active atom may be optimized while some active atoms may be set to zero, or some active atoms are set to zero (but none is added nor optimized). To sum it up, at each iteration, it constructs the next iterate using only past active atoms plus possibly a new one generated by the LMO. Therefore, Algorithm 1 belongs to the family of boosting algorithms. We refer to it as the regularized matching pursuit (RMP).

Compared to the boosting approach of Zhang et al. [2012] for metric-norm regularization or to the generalized conditional gradient [Bach, 2015, Sun and Bach, 2022], active atoms are not modified uniformly since only some of them may be reduced to zero. The SOTOPO method of Song et al. [2017] minimizes the same upper bound with respect to the ℓ_1-norm (3.16). Yet, it is resolved with a different variational formulation, that does not let a linear minimization oracle appear. Compared to proximal coordinate descent with GS-rule (3.15) applied with L_1^F (instead of L_2^F), the regularized matching pursuit happens to often follow exactly the same path when starting from zero (but not when starting from a nonzero point), as observed in Figure 3.3. This suggests a connection between regularized matching pursuit and proximal coordinate descent, as proven by Locatello et al. [2018] for gradient descent and steepest coordinate descent.

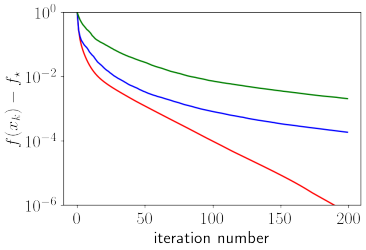
In Figure 3.3, the RMP appears to converge linearly in the underparametrized regime, and sublinearly in the overparametrized regime. We compute some convergence guarantees in the next section.

Algorithm 1 Regularized matching pursuit (RMP)

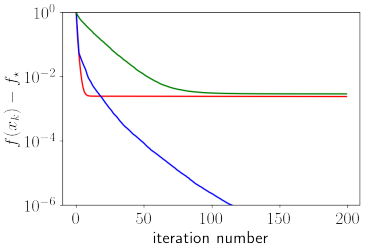
$\alpha \in \mathbf{R}^d$, $N \in \mathbf{N}$
for $k \in [0, \dots, N]$ **do**
  $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ and $i_{\min} = \arg\max_i |\nabla_i F(\alpha_k)|$
  For $\alpha_k^{(i)} \neq 0$, compute $z_i = \lambda + \frac{\alpha_k^{(i)}}{|\alpha_k^{(i)}|} \nabla_i F(\alpha_k)$ such that $z_{i+1} \geqslant z_i$
  if $\{i, z_i \geqslant z_{\min}\} = \emptyset$ **then**
    $\beta^{(i_{\min})} = -\text{sign}(\nabla_{i_{\min}} F(\alpha_k)) \frac{z_{\min}}{L_1^F}$
  else
    Compute $u = \text{argmin}_i \{z_i \geqslant z_{\min}\}$ and for $i \in [u, v]$, compute $\frac{dh}{dz}(z_i)$
    if $\frac{dh}{dz}(z_{\min}) \leqslant 0$ or $\frac{dh}{dz}(z_u) \leqslant 0$ **then**
      For $i \in [u, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
      If $\frac{dh}{dz}(z_{\min}) \leqslant 0$, **then** $\beta^{(i_{\min})} = -\text{sign}(\nabla_{i_{\min}} F(\alpha_k))(\frac{z_{\min}}{L_1^F} - \sum_{i=u}^{v} |\alpha_k^{(i)}|)$
    else
      $n = \arg\max\{i, i \in [u, v-1],\ \frac{dh}{dz}(z_i^+), \frac{dh}{dz}(z_{i+1}^-) \geqslant 0\}$
      if $n = v - 1$ **then**
        $\beta^{(v)} = -\text{sign}(\alpha_v) \frac{1}{L_1^F} (\lambda + \frac{\alpha_v}{|\alpha_v|} \nabla_v F(\alpha_k))$
      else
        For $i \in [n+1, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
        If $\frac{dh}{dz}(z_n^+) \leqslant 0$, **then** $\beta^{(n)} = -\text{sign}(\alpha_k^{(n)})(\frac{1}{L_1^F} \left(\lambda + \frac{\alpha_k^{(n)}}{|\alpha_k^{(n)}|} \nabla_n F(\alpha_k) \right) - \sum_{i=n+1}^{v} |\alpha_k^{(i)}|)$
      end if
    end if
  end if
  $\alpha_{k+1} = \alpha_k + \beta$
end for



(a) $d = 30$, starting from zero.



(b) $d = 30$, starting from a non zero point.



(c) $d = 500$, starting from a non zero point.

Figure 3.3: Convergence in function value of the proximal gradient descent, coordinate descent with Gauss-Soutwhell rule and with $L = L_1^F$ (instead of $L = L_2^F$) and of the regularized matching pursuit, for synthetic quadratics (see Section 3.2.3) with $n = 50$, $s = 8$, $\lambda = 0.001$, $\sigma = 0.5$. RMP and coordinate descent with GS-rule match exactly in these examples.

3.3.2 Convergence guarantee

We now establish convergence guarantees for the RMP, both for strongly convex and non-strongly convex functions. We consider a more general composite minimization problem,

$$\min_{\alpha \in \mathbb{R}^d} \Big\{ G(\alpha) \triangleq F(\alpha) + H(\alpha) \Big\}, \tag{3.18}$$

where H is closed, convex, proper, and where F is L_1^F-smooth and (possibly) μ_1^F-strongly convex with respect to the ℓ_1-norm. If in addition, F is a linear mapping, and $H(\cdot) = \|\cdot\|_1$, this is exactly the original optimization Problem (3.1). We evaluate the convergence guarantees of a generalized version of the RMP (the GRMP), that is not always a boosting method,

$$\alpha_{k+1} = \arg\min_{\alpha \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + H(\alpha). \tag{3.19}$$

As we will see, our proofs are similar to those for randomized coordinate descent [Richtárik and Takáč, 2014, Theorem 5, 7].

3.3.2.1 Strongly convex functions

Let us assume that F is L_1^F-smooth and μ_1^F-strongly convex, typically in the underparametrized regime. Similarly to coordinate gradient descent with GS rule which converges linearly in this context [Nutini et al., 2018b], regularized matching pursuit is formulated as the minimization of the smoothness upper bound with respect to the ℓ_1-norm. Therefore, it benefits from linear convergence guarantees, detailed below.

Proposition 7. *[Nutini, 2018, Appendix A.8] Let F be convex, L_1^F-smooth with respect to the ℓ_1-norm, and μ_1^F-strongly convex with respect to the ℓ_1-norm. Then, the sequence (α_k) generated by (3.19) verifies,*

$$G(\alpha_{k+1}) - G_\star \leqslant \left(1 - \frac{\mu_1^F}{L_1^F}\right) (G(\alpha_k) - G_\star).$$

Proof. The proof is taken from Nutini [2018, Appendix A.8.] and consists in an optimization step over all trajectories. The argument is inspired from randomized coordinate descent [Richtárik and Takáč, 2014]). □

The RMP is a special case of method (3.19), and verifies the convergence guarantee of Proposition 7. As a conclusion, it beats traditional boosting techniques converging sublinearly,

such as coordinate descent with GS rule (with $1 - \frac{\mu_2^F}{dL_2^F} \leqslant 1 - \frac{\mu_1^F}{L_1^F}$), or the generalized conditional gradient that is also adapted to a gauge geometry. In addition, its linear guarantee only depends on the strong convexity and smoothness parameters of F with respect to the ℓ_1-norm. In the special case of the LASSO, the estimates established in Proposition 5 and 6 still apply. In the overparametrized regime however, Figure 3.3 suggests that the method does not converge linearly (since it is stuck at an accuracy of about around 10^{-5}).

3.3.2.2 Smooth convex functions

Let now F be L_1^F-smooth, convex, but not strongly convex (which is verified in the overparametrized regime). Usually, guarantees for splitting methods, such as proximal gradient, states a sublinear convergence guarantee. Similarly in Proposition 8, we prove sublinear convergence for the GRMP. To our knowledge, there is no such result for sublinear convergence for SOTOPO [Song et al., 2017] or for coordinate descent with the Gauss-Southwell rule, which is very close to the GRMP.

Proposition 8. *Let (α_k) be generated by the generalized regularized matching pursuit (3.19), starting from $\alpha_0 \in \mathbb{R}^d$*

$$G(\alpha_k) - G_\star \leqslant \frac{2L_1^F \mathcal{R}_{\alpha_0}^2}{k+1},$$

where $\mathcal{R}_{\alpha_0}^2 = \max_{\alpha \in \mathbb{R}^d} \max_{\alpha_\star \in \mathbb{R}^d} \{\|\alpha - \alpha_\star\|_1^2, \text{ s.t. } G(\alpha) \leqslant G(\alpha_0)\}$.

Proof. This technique is inspired from a proof for sublinear convergence of randomized proximal coordinate descent established by Richtárik and Takáč [2014, Theorem 5]. Let $\alpha_{k+1} \in \mathbb{R}^d$ be a minimizer of the smooth upper bound:

$$\begin{aligned}
G(\alpha_{k+1}) &\leqslant \inf_{\alpha \in \mathbb{R}^d} F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2}\|\alpha - \alpha_k\|_1^2 + H(\alpha), \\
&\leqslant \inf_{\alpha \in \mathbb{R}^d} F(\alpha) + H(\alpha) + \frac{L_1^F}{2}\|\alpha - \alpha_k\|_1^2 \ (= G(\alpha) + \frac{L_1^F}{2}\|\alpha - \alpha_k\|_1^2) \ (F \text{ convex}), \\
&\leqslant \inf_{t \in [0,1]} G(t\alpha_\star + (1-t)\alpha_k) + \frac{L_1^F t^2}{2}\|\alpha_k - \alpha_\star\|_1^2, \\
&\leqslant \inf_{t \in [0,1]} G(\alpha_k) - t(G(\alpha_k) - G_\star) + \frac{L_1^F t^2}{2}\|\alpha_k - \alpha_\star\|_1^2 \ (\text{convexity of } H, F), \\
G(\alpha_{k+1}) - G_\star &\leqslant \inf_{t \in [0,1]} (1-t)(G(\alpha_k) - G_\star) + \frac{L_1^F t^2}{2}\|\alpha_k - \alpha_\star\|_1^2.
\end{aligned}$$

The solution of this minimization problem is given by $t_\star = \min(1, \frac{G(\alpha_k) - G_\star}{L_1^F \|\alpha_k - \alpha_\star\|_1^2})$. We conclude the minimization bound, depending on the sign of $G(\alpha_k) - G_\star - L_1^F \|\alpha_k - \alpha_\star\|_1^2$:

$$G(\alpha_{k+1}) - G_\star \leqslant \max\left(1 - \frac{G(\alpha_k) - G_\star}{2L_1^F \|\alpha_k - \alpha_\star\|_1^2}, \frac{1}{2}\right)(G(\alpha_k) - G_\star).$$

As a first conclusion, notice that $G(\alpha_k) - G_\star$ is nonincreasing. Recall now that $\mathcal{R}_{\alpha_0}^2 = \max_{\alpha \in \mathbb{R}^d} \max_{\alpha_\star \in \mathbb{R}^d} \{\|\alpha - \alpha_\star\|_1^2, \text{ s.t. } G(\alpha) \leqslant G(\alpha_0)\}$. Then, using the notation $\delta_k = G(\alpha_k) - G_\star$, an upper bound for δ_{k+1} is given by $\delta_{k+1} \leqslant \max\left(1 - \frac{\delta_k}{2L_1^F \mathcal{R}_{\alpha_0}^2}, \frac{1}{2}\right)\delta_k$. Assume now that $\delta_0 \leqslant L_1^F \mathcal{R}_{\alpha_0}^2$ and notice that $\delta_k \leqslant L_1^F \mathcal{R}_{\alpha_0}^2$ since δ_k is nonincreasing. If not, notice that the inequality satisfied at the next iteration $\delta_1 \leqslant \frac{1}{2}\mathcal{R}_{\alpha_0}^2$. Then, we have for $\omega = \frac{1}{2L_1^F \mathcal{R}_{\alpha_0}^2}$, $\delta_{k+1} \leqslant (1 - \delta_k \omega)\delta_k$. Following the same argument as in the proof for sublinear convergence of steepest coordinate descent, detailed in Appendix 3.D, we arrive to a convergence guarantee $G(\alpha_k) - G_\star \leqslant \frac{2L_1^F \mathcal{R}_{\alpha_0}^2}{k+1}$. □

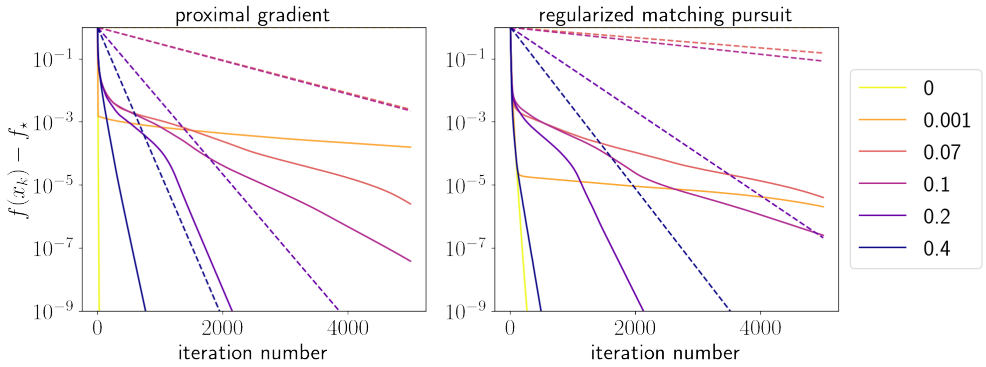


Figure 3.4: Convergence in function values for the proximal gradient on the left and the regularized matching pursuit on the right for $n = 50$, $d = 500$ and a sparsity $s = 8$ and for several penalty λ. Convergence is compared in dashed lines to local convergence guarantee, taken on the support S on the last iterates and the SDP relaxation from Proposition 5.

Proposition 8 provides a sublinear convergence guarantee for the GRMP for non-strongly convex functions. To our knowledge, this is the first sublinear guarantee for a boosting algorithm under classical assumptions from convex optimization. This method does not benefit from linear convergence guarantee. Yet, as we see from the numerical experiments in the next section that the RMP does converge linearly in certain regimes in the case of the ℓ_1-regularized model.

3.3.3 A transition phase depending on λ: experimental results

The RMP algorithm benefits from convergence guarantees similar to those for the proximal gradient: these methods converge linearly under strong convexity assumptions (underparametrized regime) but have sublinear guarantees for smooth convex problems (overparametrized regime). In the context of sparsity though, the proximal gradient descent benefits from linear convergence under additional assumptions on the problem classes such as restricted eigenvalue properties [Raskutti et al., 2010], and for a well-chosen parameter λ [Agarwal et al., 2010, Theorem 2]. In this section, our experiments reveals a transition phenomenon driven by λ on a LASSO problem $F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2 + \lambda\|x\|_1$, where P are synthetic Gaussian data in the overparametrized setting as in Section 3.2.3.

The convergence behavior of proximal gradient descent follows a transition phase, that can be divided into three phases: first, the method converges linearly according to the nonregularized trajectory, then it converges sublinearly, and it converges linearly once the support is identified. Iutzeler and Malick [2020, Theorem 1] prove the proximal gradient identifies the structure of the solution (described by manifolds, or sparsity patterns) after a certain number of steps. For strongly convex functions, Nutini et al. [2018a] bounded the 'active-set'-complexity of the proximal gradient method. The regularized matching pursuit follows the same behavior. It appears that the sparsity of the solution, and the sparsity identification highly depends on the value of λ: the larger λ, the sparser the solution and the quicker the identification. Thanks to this observation, we derive a posteriori guarantee in Figure 3.4, based on the sparsity of the solution to the optimization problem. In Figure 3.4, local strong convexity parameters are given by the estimated of Corollary 6 and Proposition 5. We recover that large parameter λ both induces a stricter sparsity on the solution and a better convergence.

We describe this transition phase numerically in Figure 3.5 by plotting the ϵ-curve (see Section 3.2.3) as a function of λ. For $\lambda = 0$, both methods converge linearly (as expected in the overparametrized regime for gradient descent and coordinate descent with the GS-rule). For 'large' values of λ for which the support is quickly identified, the convergence is linear too. In the intermediary phase however, convergence depends on the effective dimension of the trajectory,

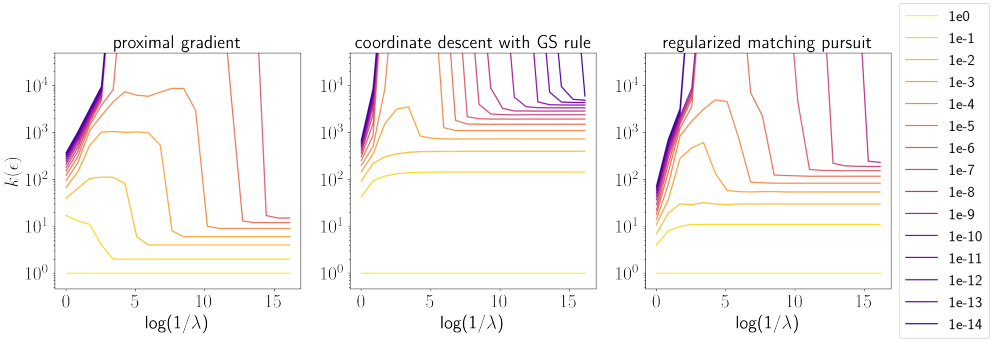


Figure 3.5: ϵ-curve of the proximal gradient, coordinate descent with the GS rule and regularized matching pursuit for a LASSO problem with $d = 500$, $n = 50$, a sparsity level $s = 8$, $\sigma = 0.5$, after $k = 10000$ iterations for several values of λ.

$d_{eff} \approx n$ by construction (and thus, on the effective strong convexity of f long the trajectory). The ϵ-curves can be seen as equivalent in the optimization perspective with the regularization path usually drawn in the context of statistical recovery.

The regularized matching pursuit Algorithm 1 formulation allows some intuition regarding the interplay between λ and the sparsity of the solution. Let $\mathcal{A} = \{i, \alpha_k^{(i)} > 0\}$ be the set of active atoms. Algorithm 1 may reduce an active atom $i \in \mathcal{A}$ to zero if $\|\nabla F(\alpha)\|_\infty - \frac{\alpha_k^{(i)}}{|\alpha_k^{(i)}|}\nabla_i F(\alpha) \leqslant 2\lambda$. The larger λ, the more active directions may be canceled out. The smaller λ ($\lambda \ll \|\nabla F(\alpha)\|_\infty$), the closer is regularized matching pursuit to coordinate descent with GS rule (on the right in Figure 3.5): indeed, only the linear minimization oracle may be added to the set of atoms without modifying other active atoms ($z_i \lesssim z_{\min}$). For $\lambda \approx 0$, the regularized matching pursuit thus converges linearly up to a certain iteration number, which appears with the parallel level lines in Figure 3.5.

Based on the minimization of a smoothness upper bound with respect to the ℓ_1-norm, we have developed a regularized matching pursuit algorithm, that benefits from linear convergence in the underparametrized regime (where F is strongly convex), and sublinear convergence in the overparametrized regime (where F is not strongly convex). Thanks to the ϵ-curve, we numerically described the role of λ on the convergence of the method (and on the sparsity). In the following section, we propose to develop a method suited to the gauge geometry in the overparametrized regime.

3.3.4 An ultimate method adapted to the geometry of regularized models

The regularized matching pursuit 1 was derived from the ℓ_1-geometry. In Section 3.2.4, for non-regularized models, coordinate descent with GS-rule was interpreted as a matching pursuit algorithm in both the underparametrized and overparametrized regime. In what follow, we see that the regularized matching pursuit as developed above does not benefit from this formulation in the overparametrized regime. Instead, we propose an 'ultimate method' for the gauge geometry, that benefits from linear convergence in the overparametrized regime but lacks a simple formulation.

Recall the equivalent regularized minimization problems (3.1) and (3.3),

$$\min_{\alpha \in \mathbb{R}^d} f(P\alpha) + \lambda\|\alpha\|_1 = \min_{x \in \mathbb{R}^n} f(x) + \lambda\gamma_{\mathcal{P}}(x),$$

where $\gamma_{\mathcal{P}}$ is a gauge function as defined in Section 3.2.4, and f is $L^f_{\gamma_{\mathcal{P}}}$-smooth and $\mu^f_{\gamma_{\mathcal{P}}}$-strongly convex with respect to the gauge. The problem in $\mathbb{R}^d$ is reformulated in $\mathbb{R}^n$, of lower dimension.

Remark 7. *This reformulation only requires $\gamma_{\mathcal{P}}$ to be a gauge function, but not specifically to be a norm. Reversely, minimizing a function penalized by a gauge function (or a semi-norm) can be reformulated as minimizing a linear function penalized by and ℓ_1-norm.*

As for the regularized matching pursuit, we formulate an optimization method as the minimization of the smoothness upper bound with respect to the gauge function, starting from $x_0 \in \mathbb{R}^n$:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \langle \nabla f(x_k), x - x_k \rangle + \frac{L^f_{\gamma_{\mathcal{P}}}}{2} \gamma_{\mathcal{P}}(x - x_k)^2 + \lambda \gamma_{\mathcal{P}}(x). \tag{3.20}$$

We refer to this method as the ultimate method for the gauge $\gamma_{\mathcal{P}}$, that is adapted to the geometry of the regularized problem 3.1. Let us reformulate the minimization Problem (3.20) on $\mathbb{R}^n$ into a minimization problem in $\mathbb{R}^d$. Let $x_k = P\alpha_k$ with $\alpha_k \in \mathbb{R}^d$, then

$$\begin{aligned}
&\min_{x \in \mathbb{R}^n} \langle \nabla f(x_k), x - x_k \rangle + \frac{L^f_{\gamma_{\mathcal{P}}}}{2} \gamma_{\mathcal{P}}(x - x_k)^2 + \lambda \gamma_{\mathcal{P}}(x), \\
&= \min_{\alpha, \nu \in \mathbb{R}^d} \langle \nabla f(P\alpha_k), P(\alpha - \alpha_k) \rangle + \frac{L^f_{\gamma_{\mathcal{P}}}}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\nu\|_1, \text{ s.t. } x = P\alpha = P\nu, \\
&= \min_{\alpha, \nu \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L^f_{\gamma_{\mathcal{P}}}}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\nu\|_1, \text{ s.t. } P\alpha = P\nu.
\end{aligned}$$

When $P\alpha = P\nu$ implies $\alpha = \nu$, such as in the underparametrized regime where $P^\top P$ is invertible, the ultimate method for the gauge is equivalent with the regularized matching pursuit (1). However, in the overparametrized regime, $P\alpha = P\nu$ does not imply $\alpha = \nu$ in general. This method does not belong to boosting algorithms due to the evaluation of the gauge function in x and in $x - x_k$ in (3.20). In addition, this minimization problem admits neither a simple closed-form solution in general nor a solution based on the KKT conditions (as we did for regularized matching pursuit). While not directly computable in general, the minimization step (3.20) converges linearly to the optimum, as proven below in Proposition 9.

Proposition 9. *Let f be $L^f_{\gamma_{\mathcal{P}}}$-smooth and $\mu^f_{\gamma_{\mathcal{P}}}$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. The ultimate method (3.20) (x_k) converges linearly with*

$$f(x_k) - f_\star \leqslant \left(1 - \frac{\mu^f_{\gamma_{\mathcal{P}}}}{L^f_{\gamma_{\mathcal{P}}}}\right)^k (f(x_0) - f_\star).$$

Proof. The proof follows exactly the proof for Theorem 7, replacing the function F by f and the norm $\|\cdot\|_1$ by $\gamma_{\mathcal{P}}(\cdot)$. □

Proposition 9 provides a linear convergence guarantee for the ultimate algorithm. We recover the convergence guarantee of coordinate descent with GS rule in the non regularized model (3.12). In Figure 3.6, we solve the optimization step for the ultimate method for the gauge with the solver MOSEK [ApS, 2022] on a LASSO problem. It converges linearly in the overparametrized regime, while the other method are first stuck in a sublinear phase. Compared to the proximal gradient, proximal coordinate descent with GS rule, the regularized matching pursuit and the ultimate method starts with sparse solution, and differs after a small number of iteration (about 30 here). In the special case of the LASSO, it is possible to approximate its convergence guarantee as for the linear regression problem. Noticing that $L^f_{\gamma_{\mathcal{P}}} = L^F_1$ and $\mu^f_{\gamma_{\mathcal{P}}} = \mu^F_1$, the estimate of the convergence guarantee of coordinate descent with GS rule from Proposition 5 apply here. In the Appendix 3.F, we propose an inner loop strategy to avoid the use of an optimization solver, together with the convergence analysis of the outer loop given the precision of the inner loop. Note that this analysis can be extended to constrained models,

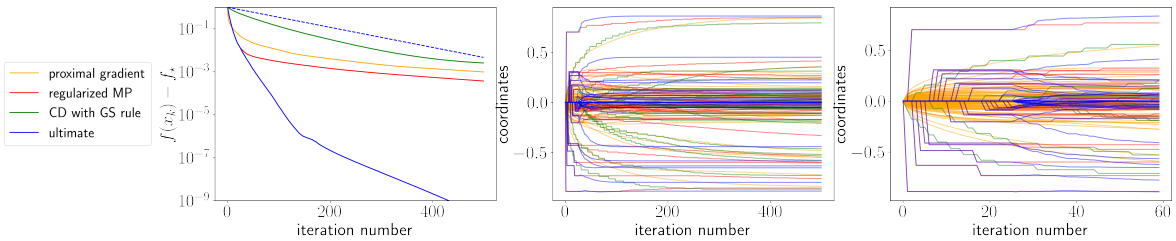


Figure 3.6: Convergence in function value and coordinates as a function of the iteration number for the proximal gradient descent, the proximal coordinate descent with GS rule, the regularized matching pursuit and the ultimate method on a LASSO problem, with $d = 500$, $n = 50$, $s = 8$, $\lambda = 0.2$. The approximate guarantee in provided in dashed lines.

leading to a constrained matching pursuit strategy benefiting from similar properties, as detailed in Appendix 3.G.

Conclusion and future works

In this chapter, we developed a principled view for generating optimization algorithms from the minimization of a smoothness upper bound with respect to a well-chosen norm. For non-regularized models, this procedure leads to coordinate descent with GS-rule, that can be interpreted as a matching pursuit algorithm both in the ℓ_1-geometry for underparametrized models, and in the $\gamma_{\mathcal{P}}$-geometry for overparametrized models. Building on these results, we derive a new regularized matching pursuit algorithm based on the minimization of smoothness with respect to the ℓ_1-norm (whose counterpart is proximal gradient descent in the ℓ_2-geometry). While strongly connected to proximal coordinate descent with GS-rule, the regularized matching pursuit cannot be interpreted as a matching pursuit algorithm in the gauge geometry for overparametrized models and does not converge linearly in this regime. We finally formulate an ultimate method adapted to overparametrized geometries. Yet, this method lacks a closed-form formulation. In numerical experiments, we approximate it using an inner-loop strategy.

From this approach, we obtain refined convergence guarantees for (resp. regularized) matching pursuit (resp. coordinate descent with GS rule), that are adapted to the geometry of the problem under consideration. For linear regression and the LASSO, we derive upper bounds (resp. high probability bounds) for convergence guarantees using SDP relaxations (resp. under statistical assumptions on the data). As a byproduct, convergence guarantees of both gradient descent and steepest coordinate descent applied to least-squares follow a transition phase from the underparametrized to the overparametrized regime. For ℓ_1-regularized models, a similar transition phase for λ appears, and allows to interpret it as a measure of the sparsity of the solution.

Building on these results, we believe it could be of interest to extend this principled approach to accelerated matching pursuit algorithms (and thus, to accelerated coordinate descent algorithms). Some accelerated techniques have already been developed relying on randomly selected coordinates, such as those of Nesterov and Stich [2017] for nonregularized minimization and Fercoq and Richtárik [2015] or Locatello et al. [2018, Section 3] for composite minimization problems. Another interesting line of research could be to understand the connections between the observed transition phase for optimization methods and the double descent phenomenon observed for the generalization error in machine learning.

Appendices

In this document, we provide proofs for the main theorems of the chapter as well as additional experiments offering a comprehensive overview of the main chapter's results. Table 3.1 summarizes the main contributions and results of the Appendix.

Table 3.1: Content of the appendices

Appendix 3.A	Mathematical tools appearing in the proofs: η-tricks and computation of the basis of a kernel.
Appendix 3.B	Proof for linear convergence of matching pursuit (Proposition 4) and gradient descent (Proposition 3).
Appendix 3.C	Estimates for the convergence of steepest coordinate descent for least-squares: SDP relaxations and high-probability bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F, with numerical comparisons to μ_2^F and $\mu_{2,L}^F$.
Appendix 3.D	Matching pursuit in the gauge geometry: properties of the gauge and sublinear convergence guarantee.
Appendix 3.E	Formulation of steepest coordinate descent as a 'nearly' matching pursuit algorithm.
Appendix 3.F	Efficiently computing the ultimate method: an inner loop strategy.
Appendix 3.G	Constrained matching pursuit and convergence properties.

3.A Mathematical tools

3.A.1 The η-tricks: reweighted least-squares formulations

Due to non-smoothness, the ℓ_1-norm is often seen as difficult to optimize. A common way to simplify regularization term containing an ℓ_1-term, such as the LASSO, consists in formulating it as a reweighted least-square problems. We refer to the work of Bach et al. [2012] There are three common formulations:

- the variational formulation [Bach et al., 2012, Section 5]

$$\|x\|_1 = \min_{\eta \in \mathbb{R}^n, \eta \geqslant 0} \frac{1}{2} \sum_{i=1}^n \frac{x_i^2}{\eta_i} + \frac{1}{2} \sum_{i=1}^n \eta_i,$$

- the variational constrained formulation [Bach et al., 2012, Section 1]

$$\|x\|_1^2 = \min_{\eta \in \Delta_n} \sum_{i=1}^n \frac{x_i^2}{\eta_i},$$

  where $\Delta_n = \{\eta \in \mathbb{R}^n, \eta \geqslant 0, \sum_{i=1}^n \eta_i = 1\}$ is the simplex,

- the maximization problem,

$$\|x\|_1 = \max_{\|s\|_\infty \leqslant 1} \langle s, x \rangle.$$

Finally, we notice a useful variational trick, in dimension 1. For $x \in \mathbb{R}^n$,

$$\frac{L}{2}\|x\|_1^2 = \max_{z \geqslant 0} \frac{-z^2}{2L} + z\|x\|_1, \tag{3.21}$$

with the optimum $z_\star = L\|x\|_1$.

3.A.2 Computing the basis of a kernel

To compute the basis of $\mathrm{Ker}(P)$, we perform a QR decomposition on P such that $Q_1^\top Q_1 = I_n$, $Q_2^\top Q_2 = I_{d-n}$ and $Q_1^\top Q_2 = 0$: $P^\top = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}$, where Q_2 is a basis for the nullspace of P. Indeed, let $z \in \mathbb{R}^{d-n}$, and $AQ_2 z = R^\top Q_1^\top Q_2 z = 0$.

3.B Proof for Propositions 3, 4

Let us prove linear convergence of gradient descent with fixed step size and coordinate descent with GS rule in a more general framework. This proof leads to the results of Propositions 3, 4 for the ℓ_2, ℓ_1 norm respectively.

Let F be L^F-smooth with respect to a norm $\|\cdot\|$, (possibly) μ^F-strongly convex with respect to $\|\cdot\|$ and verify the Łojasiewicz inequality with parameter μ_L^F. We consider a method (α_k) starting from $\alpha_0 \in \mathbb{R}^d$, and obtained by minimizing the smoothness quadratic upper bound:

$$F(\alpha_{k+1}) \leqslant F(\alpha_k) + \min_{\alpha \in \mathbb{R}^d} \left(\langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L^F}{2} \|\alpha - \alpha_k\|^2 \right) \leqslant F(\alpha_k) - \frac{1}{2L^F} \|\nabla F(\alpha_k)\|_\star^2.$$

If F is μ^F-strongly convex, then F verifies the Łojasiewicz inequality with parameter μ^F: for all $\alpha \in \mathbb{R}^d$, $\mu^F(F(\alpha) - F_\star) \leqslant \frac{1}{2}\|\nabla F(\alpha)\|_\star^2$. Thus, substracting $F_\star$ on each side of the smoothness inequality, we have

$$F(\alpha_{k+1}) - F_\star \leqslant \left(1 - \frac{\mu^F}{L^F}\right) (F(\alpha_k) - F_\star).$$

Similarly, if F satisfies the Łojasiewicz inequality with parameter μ_L^F, but is not strongly convex ($\mu^F = 0$).

3.C Estimates for the convergence of steepest coordinate descent for least-squares

In Proposition 4, a sequence (α_k) generated by steepest coordinate descent for a linear regression problem has a linear convergence rate in function values,

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_1^F, \mu_{1,L}^F)}{L_1^F}\right)^k (F(\alpha_0) - F_\star).$$

We assume that F is a quadratic, that is $F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$. We first derive SDP relaxations for the optimization problems characterizing μ_1^F, $\mu_{1,L}^F$ and L_1^F, and numerically compare these estimates to μ_2^F and $\mu_{2,L}^F$. Then, we compute inequalities connecting $\mu_{1,L}^F$ and μ_1^F to L_1^F. Thanks to these inequalities, we prove concentration inequalities for $\mu_{1,L}^F$, μ_1^F and L_1^F, and derive approximate convergence guarantees of steepest coordinate descent.

3.C.1 SDP relaxations for μ_1^F and $\mu_{1,L}^F$

We look for exact lower bounds for μ_1^F and $\mu_{1,L}^F$. Both in the overparametrized and underparametrized regime, we are going to reformulate the optimization problems defining $\mu_1^F, \mu_{1,L}^F$ into SDPs, and relax some rank constraints. Then, we compare these estimates to μ_2^F and $\mu_{2,L}^F$ in numerical experiments.

3.C.1.1 Proof for Proposition 5

SDP relaxation for μ_1^F in the underparametrized regime. Recall from Lemma 7 that μ_1^F is non zero in this regime and given by $\frac{1}{\sqrt{n\mu_1^F}} = \max_{\alpha,\|\alpha\|_\infty\leqslant 1} \max_{\nu,\|P\nu\|_2\leqslant 1} \alpha^\top \nu$. Since $P^\top P$ is invertible, we proceed to a change of variable in $P^\top P$:

$$\begin{aligned}\frac{1}{\sqrt{n\mu_1^F}} &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1} \max_{\nu,\|Pz\|_2\leqslant 1} (P^\top Px)^\top z,\\ &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1} \max_{\nu,\|P\nu\|_2\leqslant 1} (P\alpha)^\top (P\nu),\\ &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1} \|P\alpha\|_2,\\ &= \max_{\alpha} \|P(P^\top P)^{-1}\alpha\|_2, \text{ s.t. } \|\alpha\|_\infty \leqslant 1.\end{aligned}$$

A first attempt to compute an exact solution to this problem is to consider $\alpha \in \{-1,1\}^d$ and compute all possible values for the $\|\cdot\|_\infty$-norm. However this computation would require 2^d configurations. Instead, we compute an SDP relaxation of $\frac{1}{\sqrt{\mu_1^F}}$ in Lemma 13.

Lemma 13. *Let the regime be underparametrized ($n \geqslant d$). μ_1^F has a $\frac{\pi}{2}$-approximation:*

$$\tilde{\mu}_1^F \leqslant \mu_1^F \leqslant \frac{\pi}{2}\tilde{\mu}_1^F.$$

where $\tilde{\mu}_1^F$ has an SDP-formulation $\frac{1}{\tilde{\mu}_1^F} = n \max_{X\succcurlyeq 0,\mathrm{diag}(X)\leqslant 1} \mathrm{Tr}(CX)$, with $C = (P^\top P)^{-1}$.

Proof. Let us reformulate an SDP relaxation to this problem. The problem $OPT = \frac{1}{\mu_1^F} = n \max_{\nu,\|\nu\|_\infty^2\leqslant 1} \|P(P^\top P)^{-1}\nu\|_2^2 = n \max_{\nu,\nu_i^2\leqslant 1} \nu^\top (P^\top P)^{-1}\nu$ can be relaxed into an SDP,

$$SDP = \frac{1}{\tilde{\mu}_1^F} = n \max_{Z\succcurlyeq 0} \mathrm{Tr}(CZ), \quad \text{s.t. } \mathrm{diag}(Z) \leqslant 1, \tag{3.22}$$

where $C = (P^\top P)^{-1} \succ 0$. This is exactly the Max-Cut SDP relaxation, for which Goemans and Williamson [1995] proved the approximation $\frac{2}{\pi} SDP \leqslant OPT \leqslant SDP$. □

SDP relaxation for $\mu_{1,L}^F$ in the overparametrized regime. For $\mu_{1,L}^F$, the maximization problem does not correspond to Max-Cut. So it is possible to derive an SDP lower bound, but no SDP-approximation. Recall the formulation of $\mu_{1,L}^F$ as an optimization problem $\mu_{1,L}^F = \frac{1}{n} \inf_{\nu\in\mathbb{R}^d} \|P^\top P\nu\|_\infty^2$, s.t. $\|P\nu\|_2^2 = 1$. In Lemma 14, we derive a lower bound for $\mu_{1,L}^F$ formulated as an SDP.

Lemma 14. *Let $PP^\top$ be invertible, then,*

$$\mu_{1,L}^F \geqslant \frac{1}{n} \inf_{Z\succcurlyeq 0} \|P^\top PZP^\top P\|_\infty, \quad \text{s.t. } \mathrm{Tr}(P^\top PZ) = 1,$$

Proof. We introduce $Z = \nu\nu^\top \in \mathbb{R}^{d\times d}$, where $\mathrm{rank}(Z) = 1$, and reformulate the problem into $\mu_{1,L}^F = \frac{1}{n} \inf_{Z\succcurlyeq 0,\mathrm{rank}(Z)=1,\mathrm{Tr}(P^\top PZ)=1} \|P^\top PZP^\top P\|_\infty$, which can be relaxed as an SDP $\mu_{1,L}^F \geqslant \frac{1}{n} \inf_{Z\succcurlyeq 0,\mathrm{Tr}(P^\top PZ)=1} \|P^\top PZP^\top P\|_\infty$. □

Comparison of $\tilde{\mu}_1^F$ and $\tilde{\mu}_{1,L}^F$. First, let us notice that norm $\|\cdot\|_\infty$ corresponds to the infinite norm on the diagonal of the matrices. Consequently, the constraint $\mathrm{diag}(X) \leqslant 1$ is equivalent with $\|X\|_\infty \leqslant 1$. In addition, in the underparametrized regime, a change of variable $X \to$

$P^\top PZP^\top P$ leads to the reformulation of (3.22) as,

$$\frac{1}{\tilde{\mu}_1^F} = n \max_{X \succcurlyeq 0} \operatorname{Tr}(P^\top PX), \quad \text{s.t. } \|P^\top PXP^\top P\|_\infty \leqslant 1.$$

Thus, we conclude with the fact that $(P^\top P)$ is not invertible in the overparametrized regime, and thus that $\frac{1}{\tilde{\mu}_1^F} \geqslant \frac{1}{\tilde{\mu}_{1,L}^F}$.

3.C.1.2 Numerical comparison

In this section, we compare the estimate $\tilde{\mu}_1^F$ (resp. $\tilde{\mu}_{1,L}^F$) for μ_1^F (resp. $\mu_{1,L}^F$) to μ_2^F and μ_2^F/d (resp. $\mu_{2,L}^F$ and $\mu_{2,L}^F/d$). We consider Gaussian matrices $X \in \mathbb{R}^{n,d}$ such that $X_i \sim \mathcal{N}(0, \Sigma)$ are i.i.d., given four different diagonal variances: a uniform variance $\Sigma = I_d$, a non-uniform variance $\Sigma = \operatorname{Diag}(1, \dots, 1/d)$, a variance with only one small value $\Sigma = \operatorname{Diag}(1, \cdots, 1, \frac{1}{100})$, a variance with only one large value $\Sigma = \operatorname{Diag}(1, \dots, 1, 100)$. These variances are inspired from the work of Nutini et al. [2018b, Section 4.1], who computed explicitly μ_1^F for separable quadratics.

Underparametrized regime: comparison for μ_1^F. In this regime, μ_1^F has an SDP approximation given by $\tilde{\mu}_1^F$, as proven in Appendix 3.C.1. By norm equivalence, $\frac{\mu_2^F}{d} \leqslant \mu_1^F \leqslant \mu_2^F$, as proven by Nutini et al. [2018b, Appendix 4]. The value of the SDP approximation for μ_1^F in Figure 3.7 is very close to its lower bound $\frac{\mu_2^F}{d}$, except for the variance where only one diagonal element of the variance is very large.

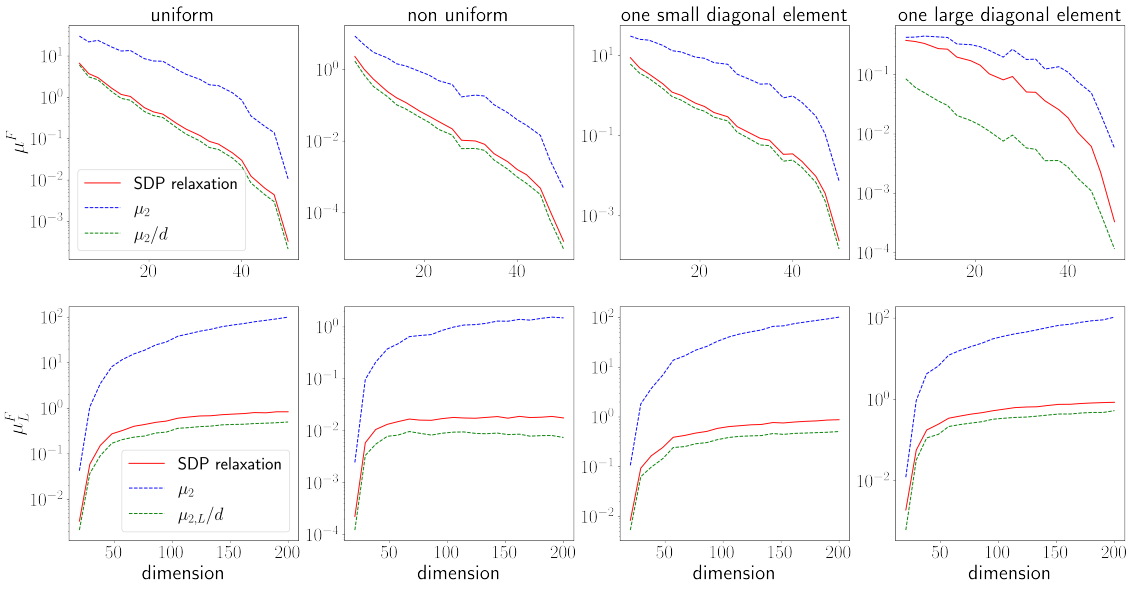


Figure 3.7: Comparison of μ_1^F on the top (resp. $\mu_{1,L}^F$ on the bottom) with μ_2^F and μ_2^F/d (resp. $\mu_{2,L}^F$ and $\mu_{2,L}^F$) for $n = 50$ (resp. $n = 20$) averaged on 5 trials, for Gaussian random data with a variance equal from the left to the right to $\Sigma = I_d$, $\Sigma = \operatorname{diag}(1, 1/2, \dots, 1/d)$, $\Sigma = \operatorname{diag}(1, \dots, 1, 1/100)$ and $\operatorname{diag}(1, \dots, 1, 100)$.

Overparametrized regime: comparison for $\mu_{1,L}^F$. In this regime, $\mu_{1,L}^F$ is lower bounded by $\tilde{\mu}_{1,L}^F$ defined by an SDP (see Appendix 3.C.1. By norm equivalence, we have that $\frac{\mu_{2,L}^F}{d} \leqslant \mu_{1,L}^F \leqslant \mu_{2,L}^F$. The SDP relaxation provides a lower bound for $\mu_{1,L}^F$, that is always close to its lower bound $\frac{\mu_{2,L}^F}{d}$ as observed in Figure 3.7. We observe similar results for a random features model generated from the Leukemia dataset (libsvm) in Figure 3.8.

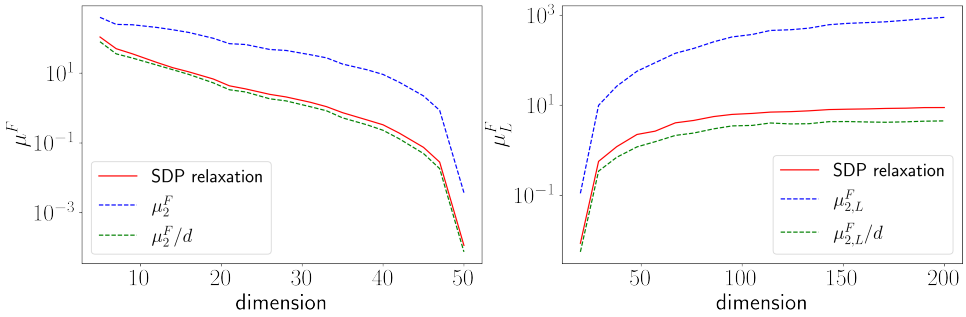


Figure 3.8: Comparison of the SDP relaxation for μ_1^F on the left (resp. $\mu_{1,L}^F$ on the right) with μ_2^F and μ_2^F/d (resp. $\mu_{2,L}^F$ and $\mu_{2,L}^F/d$) for a random feature model generated from the Leukemia dataset with a number of samples $n = 50$ (resp. $n = 20$).

3.C.2 High-probability bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F: proof for Proposition 6

In this proof, we look for concentration inequalities on μ_1^F, $\mu_{1,L}^F$ and L_1^F. To this end, we first establish deterministic inequalities connecting μ_1^F, $\mu_{1,L}^F$ and L_1^F. In a second part, we assume the data P are generated randomly, such that each row is subgaussian. Under some mild assumptions, we derive concentration inequalities on these parameters.

3.C.2.1 Deterministic inequalities connecting μ_1^F, $\mu_{1,L}^F$ and L_1^F

In this proof, we look for inequalities connecting L_1^F, μ_1^F and $\mu_{1,L}^F$ to minimal and maximal eigenvalues of $P^\top P$ and $PP^\top$. We first propose to establish deterministic inequalities characterizing these parameters.

We consider the special case of least-squares, for which $F(\alpha) = \frac{1}{n}\|P\alpha - y\|_2^2$ with $P \in \mathbb{R}^{n\times d}$. We have seen that:

- $L_1^F = \frac{1}{n}\max_{i=1,\dots,d}\|P_{:,i}\|_2^2$,
- $\mu_1^F = \frac{1}{n}\inf_{\eta\in\mathbb{R}^d}\|P\eta\|_2^2$ such that $\|\eta\|_1^2 \geqslant 1$,
- $\mu_{1,L}^F = \frac{1}{n}\sup_{\eta\in\mathbb{R}^d}\|P^\top P\eta\|_\infty^2$such that $\|P\eta\|_2^2 = 1$.

In Lemma 15, we establish deterministic lower and upper bounds for both μ_1^F and $\mu_{1,L}^F$.

Lemma 15. *Let $F(\alpha) = \frac{1}{n}\|P\alpha - y\|_2^2$ with $P \in \mathbb{R}^{n\times d}$. Denote μ_1^F (resp. $\mu_{1,L}^F$) the strongly convex (resp. Łojasiewicz) parameter of F with respect to the ℓ_1-norm. Then, μ_1^F verifies,*

$$\frac{1}{n}\frac{\lambda_{\min}(P^\top P)}{d} \leqslant \mu_1^F \leqslant \frac{1}{n}\frac{\mathbf{1}_d^\top P^\top P\mathbf{1}_d}{d^2},$$
$$\frac{1}{n}\frac{\lambda_{\min}(PP^\top)}{d} \leqslant \mu_{1,L}^F \leqslant \frac{L_1^F}{n}.$$

Proof. From result of Nutini et al. [2018b, Appendix 4], we have by the norm equivalence $\frac{\mu_2^F}{d} \leqslant \mu_1^F$. In the special case of least-squares, $\mu_2^F = \frac{\lambda_{\min}(P^\top P)}{n}$. In addition, we have that $\mu_1^F = \frac{1}{n}\inf_{\eta\in\mathbb{R}^d,\|\eta\|_1^2=1}\|P\eta\|_2^2 \leqslant \frac{1}{n}\frac{\mathbf{1}_d^\top P^\top P\mathbf{1}_d}{d^2}$ by taking $\eta = \frac{\mathbf{1}_d}{d}$.

Let us reformulate $\mu_{1,L}^F$. Using the trick $\forall \nu \in \mathbb{R}^d, \|\nu\|_1^2 = \inf_{\gamma \in \Delta_d} \sum_{i=1}^d \frac{\nu_i^2}{\gamma_i}$,

$$\begin{aligned}
\mu_{1,L}^F &= \frac{1}{n} \inf_{\nu \in \mathbb{R}^n} \|P^\top \nu\|_\infty^2, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\
&= \frac{1}{n} \inf_{\nu \in \mathbb{R}^n} \max_{\eta \in \Delta_d} \nu^\top P \mathrm{Diag}(\eta) P^\top \nu, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\
&= \frac{1}{n} \max_{\eta \in \Delta_d} \inf_{\nu \in \mathbb{R}^n} \nu^\top P \mathrm{Diag}(\eta) P^\top \nu, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\
&= \frac{1}{n} \max_{\eta \in \Delta_d} \lambda_{\min} \left(P \mathrm{Diag}(\eta) P^\top \right).
\end{aligned}$$

For $\eta = \frac{1}{d}\mathbf{1}_d$, we have that $\mu_{1,L}^F \geqslant \frac{1}{n} \frac{\lambda_{\min}(PP^\top)}{d}$. In addition, we can reformulate $\mu_1^{L,F}$ as follows: $\mu_1^{L,F} = \frac{1}{n} \max_{\eta \in \Delta_d} \lambda_{\min} \left(P \mathrm{Diag}(\eta) P^\top \right) = \frac{1}{n} \max_{\eta \in \mathbb{R}^d} \sup_{M \succcurlyeq 0, \mathrm{Tr}(M)=1} \mathrm{Tr}(P \mathrm{Diag}(\eta) P^\top M)$. Thus, $\mu_1^{L,F} = \frac{1}{n} \sup_{M \succcurlyeq 0, \mathrm{Tr}(M)=1} \|P^\top M P\|_\infty \leqslant \frac{1}{n^2} \|P^\top P\|_\infty$, for $M = \frac{I_d}{n} = \frac{L_1^F}{n}$. □

As expected by the norm equivalence (and already highlighted by Nutini et al. [2018b, Appendix 4]), strong convexity parameters verify $\mu_1^F \geqslant \frac{\mu_2^F}{d}$. Lemma 15 states a similar comparison for Łojasiwiecz parameters with $\mu_{1,L}^F \geqslant \frac{\mu_{2,L}^F}{d}$. We recover that $\mu_1^F \geqslant 0$ (resp. $\mu_{1,L}^F \geqslant 0$) in the overparametrized (resp. underparametrized) regime. Under subgaussian data, we now build high probability bounds for $\mu_{1,L}^F$, μ_1^F and L_1^F based on these lower and upper bounds.

3.C.2.2 Generalities on subgaussian data

Before deriving concentration bounds, we detail the subgaussian assumptions on P.

Definition 4. *[Vershynin, 2018, Definition 2.5.2, Proposition 2.5.2] Let X be a subgaussian random with $\tau > 0$, and $\mathbb{E}[X] = 0$. Then, there exists absolute constants $c_1, c_2 > 0$ such that,*

$$\begin{aligned}
&\forall t \in \mathbb{R}, \ \mathbb{E}[e^{tX}] \leqslant e^{c_1 \|X\|_{\Psi_2}^2 t^2}, \\
&\forall t \in \mathbb{R}, \ \mathbb{P}(|X| \geqslant t) \leqslant 2 \exp(-\frac{c_2}{\|X\|_{\Psi_2}^2} t^2),
\end{aligned}$$

In addition, $\|X\|_{\Psi_2} = \inf\{t > 0, \mathbb{E}[e^{X^2/t^2}] \leqslant 2\}$ is the subgaussian norm of X.

Usually, a subgaussian variable X with $\mathbb{E}[X] = 0$, $\mathbb{E}[x^2] = \sigma^2$ is defined with a parameter $\tau > 0$ such that for all $t \in \mathbb{R}$, $\mathbb{E}[\exp(tX)] \leqslant \exp(\frac{t^2\tau^2}{2})$. Then, $\sigma \leqslant \tau = \sqrt{2c_1}\|X\|_{\Psi_2}$. As for gaussian data, Definition 4 can be extended to vectors.

Definition 5. *[Vershynin, 2018, Definition 3.4.1] A random vector $X \in \mathbb{R}^d$ is called subgaussian if the one-dimensional marginals $\langle X, x \rangle$ are subgaussian random variables for all $x \in \mathbb{R}^d$. The subgaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{x \in \mathbb{S}^{d-1}} \|\langle X, x \rangle\|_{\Psi_2}$.*

Throughout this section, we assume the data P to be subgaussian as follows:

Assumption 2 (Subgaussian data)**.** *$P_1, \dots, P_n \in \mathbb{R}^d$ are i.i.d. subgaussian random vectors with $\mathbb{E}[P_i] = 0$, $\mathbb{E}[P_{i,j}^2] = \sigma^2$ and $K = \max_i \|P_i\|_{\Psi_2}$.*

3.C.2.3 A concentration inequality for $\sqrt{L_1^F}$

Our goal is to obtain a concentration inequality for $L_1^F = \frac{1}{n} \max_{i=1,..,d} \|P_{:,i}\|_2^2$, where the data P is generated as in Assumption 2. We rather focus on $\sqrt{L_1^F} = \frac{1}{\sqrt{n}} \max_{i=1,\dots,d} \|P_{:,i}\|_2$.

Theorem 21. *[Vershynin, 2018, Theorem 3.1.1, equation (3.3)] Let $p = (p_1, \dots, p_n) \in \mathbb{R}^n$ be a random vector with independent subgaussian coordinates p_i that satisfy $\mathbb{E}[p_i^2] = 1$, and $K = \max_i \|p_i\|_{\Psi_2}$. Then, there exists $C > 0$ an absolute constant such that, for all $t \in \mathbb{R}$*

$$\mathbb{P}(|\|p\|_2 - \sqrt{n}| \geqslant t) \leqslant 2\exp\left(-\frac{C}{K^4}t^2\right).$$

Equivalently, the concentration result of Theorem 21 can be extended by a linearity argument to random variables with variance $\mathbb{E}[p_i^2] = \sigma^2$: for all $t \in R$,

$$\mathbb{P}(|\|p\|_2 - \sigma\sqrt{n}| \geqslant t) \leqslant 2\exp\left(-\frac{C}{\sigma^2 K^4}t^2\right).$$

In Theorem 21, vectors p correspond to columns of P as defined in Assumption 2. We conclude with the concentration of each norm of the columns $(P_{:,i})$ around $\sigma\sqrt{n}$. More precisely, according to Definition 4, $\|P_{:,i}\|_2^2 - \sigma\sqrt{n}$ are i.i.d. subgaussian variables. In Lemma 16, we provide a lower and an upper bound for the expectation of $\max_{i=1,\dots,d} \|P_{:,i}\|_2$.

Lemma 16. *Let $P \in \mathbb{R}^{n\times d}$ be a collection of subgaussian elements as in Assumption 2. There exist $C_1, C_2 > 0$ absolute constant such that,*

$$C_2 K^2 \sigma \leqslant \mathbb{E}[\max_{i=1,\dots,d}(\|P_{:,i}\|_2 - \sigma\sqrt{n})] \leqslant 2\sigma K^2\sqrt{C_1 \log(d)}.$$

Proof. From Theorem 21 and Definition 4, there exists $C_1 > 0$ an absolute constant such that $Y = \|P\|_2 - \sigma\sqrt{n}$ is subgaussian with for all $t \in \mathbb{R}$, $\mathbb{E}[\exp(tY] \leqslant \exp(C_1 t^2 K^4 \sigma^2) = \exp(v^2 t^2/2)$. Boucheron et al. [Boucheron et al., 2013, Theorem 2.5] derived an upper bound for the expectation to the maximum of independent subgaussian random variables (and more generally, to subgamma random variables): $\mathbb{E}[Y] \leqslant \sqrt{2\log(d)}$. For the left-hand side, first we have the following inequality $\mathbb{E}[\max_{i=1,\dots,d} \|P_{:,i}\|_2] \geqslant \mathbb{E}[\|P_{:,i}\|_2]$. This expectation can be lower bounded using [Vershynin, 2018, Theorem 3.1.1] and $1 + x \leqslant e^x$ for all $x > 0$. □

Proposition 10. *Let P be a subgaussian matrix generated as in Assumption 2 and $L_1^F = \frac{1}{n}\max_{i=1,\dots} \|P_{:,i}\|_2^2$. Then, there exists absolute constant $C, C_1, C_2 > 0$ such that for all $t \geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}$,*

$$\mathbb{P}\left(|\sqrt{L_1^F} - \mathbb{E}[\sqrt{L_1^F}]| \geqslant t\right) \leqslant e^{-\frac{C}{\sigma^2 K^4}\min(u_1(t), u_2(t))}.$$

where $u_1(t) = \log(d)(t + \frac{C_2 K^2\sigma}{\sqrt{n}})^2$ and $u_2(t) = d(t - 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}})^2$.

Proof. Recall that $\sqrt{L_1^F} = \max_{i=1,\dots,d} \|P_{:,i}\|_2$, where $\frac{1}{\sqrt{n}}\|P_{:,i}\|_2 - \sigma$ is subgaussian. For the right side event, notice that $\mathbb{P}\left(\sqrt{L_1^F} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right) \leqslant d\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right)$ by the union bound. In addition, using the bounds on $\mathbb{E}(\max_{i=1,\dots,d} \|P_{:,i}\|_2)$ from Lemma 16 and concentration from Theorem 21, there exists $C_2, C > 0$ such that $\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right) \leqslant d\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} - \sigma \geqslant \frac{C_2 K^2\sigma}{\sqrt{n}} + t\right) \leqslant \exp\left(-\frac{C\log(d)}{\sigma^2 K^4}(t + \frac{C_2 K^2\sigma}{\sqrt{n}})^2\right)$. For the left side event, using independence of the $P_{i,j}$, we conclude from Theorem 21 that there exists an absolute constant

$C_1 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\sqrt{L_1^F} \leqslant -t + \mathbb{E}[\sqrt{L_1^F}]\right) &\leqslant \mathbb{P}\left(\sqrt{L_1^F} \leqslant -t + \sigma + 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}\right), \\ &\leqslant \mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} - \sigma \leqslant -t + 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}\right)^d, \\ &\leqslant \exp\left(-\frac{Cd}{\sigma^2K^4}(t - 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}})^2\right).\end{aligned}$$

□

We conclude from Proposition 10 that $\sqrt{L_1^F}$ concentrates around its mean, that admits lower and upper bounds as in Lemma 16. More precisely, there exist absolute constant $C, C_1, C_2 > 0$ such that for all $t \geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}^2$,

$$C_2K^2\sigma\frac{1}{\sqrt{n}} + \sigma - t \leqslant \sqrt{L_1^F} \leqslant \sigma + 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}} + t,$$

holds with probability $1 - \exp(-\frac{C}{\sigma^2K^4}u(t))$, where $u(\cdot)$ is quadratic by part.

3.C.2.4 Concentration inequality for $\mu_{1,L}^F$ and μ_1^F under subgaussian data.

As we have seen in Lemma 15, μ_1^F and $\mu_{1,L}^F$ have deterministic approximants, closely related to the minimal eigenvalues of $PP^\top$ and $P^\top P$. Under subgaussian Assumption 2, we provide concentration inequalities for μ_1^F and $\mu_{1,L}^F$.

More precisely, recall from Lemma 15 that μ_1^F (resp. $\mu_{1,L}^F$) is lower bounded by $\frac{1}{d}\frac{\lambda_{\min(P^\top P)}}{n}$ (resp. $\frac{1}{n}\frac{\lambda_{\min(PP^\top)}}{d}$). Let us begin by calling a concentration inequality [Vershynin, 2018, Theorem 4.6.1] for eigenvalues of such subgaussian matrices.

Theorem 22. *[Vershynin, 2018, Theorem 4.6.1] Let $P \in \mathbb{R}^{n\times d}$ be a subgaussian matrix generated as in Assumption 2. Then, there exists an absolute constant $C_3 > 0$ such that, for all $t \geqslant 0$,*

$$\sqrt{n} - C_3K^2(\sqrt{d} + t) \leqslant \sqrt{\lambda_{\min}(P^\top P)} \leqslant \sqrt{\lambda_{\max}(P^\top P)} \leqslant \sqrt{n} + C_3K^2(\sqrt{d} + t),$$

with probability at least $1 - 2\exp(-t^2)$, with $K = \max_i \|P_i\|_{\phi_2}$.

Similarly, we deduce from Theorem 22 that there exists $C_4 > 0$ such that for all $t \geqslant 0$,

$$\sqrt{d} - C_4K^2(\sqrt{n} + t) \leqslant \sqrt{\lambda_{\min}(PP^\top)} \leqslant \sqrt{\lambda_{\max}(PP^\top)} \leqslant \sqrt{d} + C_4K^2(\sqrt{n} + t),$$

holds with probability at least $1 - 2\exp(-t^2)$. In particular, it is possible to derive bounds for quadratics of subgaussian data from Theorem 22. We provide a concentration result in Proposition 11 for $\mathbf{1}_d^\top P^\top P\mathbf{1}_d$ that appears in Lemma 15.

Proposition 11. *Let P be generated as in Assumption 2. Then, there exists $C_3 > 0$ such that, for all $t \geqslant 0$,*

$$\frac{1}{\sqrt{d}} - C_3K^2\left(\sqrt{\frac{1}{nd}} + \frac{t}{d\sqrt{n}}\right) \leqslant \sqrt{\frac{\mathbf{1}_d^\top P^\top P\mathbf{1}_d}{d^2n}} \leqslant \frac{1}{\sqrt{d}} + C_3K^2\left(\sqrt{\frac{1}{nd}} + \frac{t}{d\sqrt{n}}\right),$$

holds with probability $1 - 2\exp(-t^2)$.

Proof. This result is directly obtained from Vershynin [2018, Theorem 4.6.1]. Under the same assumptions than in Theorem 22, there exists an absolute constant $C_3 > 0$ such that, for all $t \geqslant 0$, $\|\frac{1}{n}P^\top P - I_d\| \leqslant K^2 \max(\delta, \delta^2)$, where $\delta = C_3K^2(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}})$. From this, Vershynin concludes an approximate isometry for P [Vershynin, 2018, Lemma 4.1.5]. For all $x \in \mathbb{R}^d$, $(1-\delta)\|x\|_2 \leqslant \|Px\|_2 \leqslant (1+\delta)\|x\|_2$. Note that this is exactly the same constant than in Theorem 22. □

Using Theorem 22, Proposition 11 and Lemma 15, we get the expected concentration bounds n μ_1^F and $\mu_{1,L}^F$.

3.C.3 Proof for Corollary 7

Given the convergence guarantee for coordinate descent with GS-rule from Theorem 4, we conclude concentration for the convergence rate. We provide the proof for convergence in the underparametrized regime, where the convergence guarantee is given by $1 - \frac{\mu_1^F}{L_1^F}$, and let the overparametrized regime to the reader. Recall from Proposition 3.C.2 that, there exists absolute constant $C, C_1, C_2, C_3, K > 0$ such that,

$$1 - \frac{1}{d}\frac{\left(1 + C_3K^2(\frac{1}{\sqrt{n}} + \frac{t}{\sqrt{nd}}\right)^2}{\left(1 + C_2K^2\frac{1}{\sqrt{n}} + \frac{t}{\sqrt{n}}\right)^2} \leqslant 1 - \frac{\mu_1^F}{L_1^F} \leqslant 1 - \frac{1}{d}\frac{\left(1 - C_3K^2(\frac{1}{\sqrt{n}} + \frac{t}{\sqrt{nd}}\right)^2}{\left(1 + 2K^2\sqrt{C_1\frac{\log(d)}{n}} + \frac{t}{\sqrt{n}}\right)^2},$$

with probability $p(t) = 1 - 4\exp\left(-\min(t^2, \frac{d\sigma^2}{n}(t - 2K^2\sqrt{C_1\log(d)})^2, \frac{\log(d)\sigma^2}{n}(t + C_2^2K^2)^2)\right)$. Applying a limited development in $\frac{1}{\sqrt{n}}$, the left term is equal to

$$1 - \frac{1}{d}\left(1 + \frac{2}{\sqrt{n}}[(C_3 - C_2)K^2 + \frac{t}{\sqrt{d}} - t] + o(\frac{1}{\sqrt{n}})\right),$$

and the right term,

$$1 - \frac{1}{d}\left(1 - \frac{2}{\sqrt{n}}[K^2(C_3 + 2\sqrt{C_1\log(d)}) + \frac{t}{\sqrt{d}} + t] + o(\frac{1}{\sqrt{n}})\right).$$

These two limited development allows to conclude to the limiting convergence rate for coordinate descent with GS rule in the underparametrized regime.

3.D Matching pursuit in the gauge geometry

3.D.1 The gauge is a norm: proof for Lemma 8

Since $\alpha \to P\alpha$ is surjective, the function $\gamma_{\mathcal{P}}(x)$ is well-defined. Let us prove subbaditivity, positive definiteness and absolute homogeneity.

- Let $t > 0$ and $x \in \mathrm{R}^n$,

$$\gamma(tx) = \inf_{\alpha \in \mathrm{R}^d, tx = P\alpha} \|\alpha\|_1 = \inf_{\tilde{\alpha}(=\frac{\alpha}{t}) \in \mathrm{R}^d, x = P\tilde{\alpha}} \|t\tilde{\alpha}\|_1 = t \inf_{\tilde{\alpha} \in \mathrm{R}^d, x = P\tilde{\alpha}} \|\tilde{\alpha}\|_1 = t\gamma(x).$$

  Since $\mathcal{P} = \mathrm{conv}(P)$ is centrally symmetric, we conclude that for all $t \neq 0$, $\gamma(tx) = |t|\gamma(x)$. Finally, letting $t \to 0$, we conclude that $\gamma(0) = 0$.

- Let $x \in \mathrm{R}^n$ be such that $\gamma(x) = 0$. We have $0 = \inf_{\alpha \in \mathrm{R}^d} \|\alpha\|_1, \text{s.t. } x = P\alpha$. There exists (α_k) a sequence in R^d such that $x = P\alpha_k$ and $\|\alpha_k\|_1 \to 0$, meaning that $\alpha_k \to 0$. By linearity of $P\alpha$, we obtain $x = 0$.
- Let $x, y \in \mathrm{R}^n$, $\gamma(x+y) = \inf_\eta \|\eta\|_1, \text{s.t. } x+y = P\eta$. Let α, β be the minimal representation for x, y, such that $x = P\alpha$ and $y = P\beta$. We have that

$$\gamma(x+y) = \inf_{\eta, P(\alpha+\beta)=P\eta,} \|\eta\|_1 \leqslant \|\alpha+\beta\|_1 \leqslant \|\alpha\|_1 + \|\beta\|_1 = \gamma(x) + \gamma(y).$$

If $\gamma(\cdot)$ is a norm, we compute its dual norm $\gamma^\star(z) = \sup_{x, \gamma(x) \leqslant 1} \langle z, x \rangle = \sup_x \inf_{\lambda \geqslant 0} \langle z, x \rangle + \lambda - \lambda\gamma(x) = \sup_x \inf_{\lambda \geqslant 0} \sup_{\alpha, x = P\alpha} \langle z, x \rangle + \lambda - \lambda\|\alpha\|_1 = \inf_{\lambda \geqslant 0} \lambda + \sup_\alpha \langle \alpha, P^\top z \rangle - \lambda\|\alpha\|_1 = \inf_{\lambda \geqslant 0, \|P^\top z\|_\infty \leqslant \lambda} \lambda = \|P^\top z\|_\infty$.

3.D.2 Proof for sublinear convergence of matching pursuit

We have seen in Section 3.2.4 that matching pursuit converges linearly in both the under-parametrized and overparametrized regime. This result improves the sublinear guarantee of matching pursuit proven by Locatello et al. [2018, Theorem 3] letting a sublevel set radius appear (as usual in the coordinate descent literature).

Theorem 23. *Let f be convex, $L^f_{\gamma_\mathcal{P}}$-smooth with respect to the norm $\gamma_\mathcal{P}(\cdot)$. Then, the sequence verifies for $\mathcal{R} = \max_{x_\star \in X_\star} \max_{x \in \mathbb{R}^d} \gamma_\mathcal{P}(x - x_\star)$, s.t. $f(x) \leqslant f(x_0) < +\infty$,*

$$f(x_k) - f_\star \leqslant \frac{2L^f_{\gamma_\mathcal{P}} \mathcal{R}^2}{k-1}.$$

Proof. The sequence (x_k) verifies the descent lemma:

$$f(x_{k+1}) - f(x_k) \leqslant -\frac{1}{2L^f_{\gamma_\mathcal{P}}} \sigma_\mathcal{P}(\nabla f(x_k))^2.$$

We introduce $\delta_k = f(x_k) - f_\star$, so that $\delta_{k+1} \leqslant \delta_k - \frac{1}{L^f_{\gamma_\mathcal{P}}} \sigma_\mathcal{P}(\nabla f(x_k))^2$. By convexity and Cauchy-Schwarz inequality, we have $\delta_k \leqslant \langle x_k - x_\star, \nabla f(x_k) \rangle \leqslant \gamma_\mathcal{P}(x_k - x_\star) \sigma_\mathcal{P}(\nabla f(x_k))$. Then,

$$\delta_{k+1} \leqslant \delta_k - \frac{1}{L^f_{\gamma_\mathcal{P}} \gamma_\mathcal{P}(x_k - x_\star)^2} \delta_k^2 \leqslant \delta_k - \frac{1}{L^f_{\gamma_\mathcal{P}} \mathcal{R}^2}, \delta_k^2.$$

where $\mathcal{R} = \max_{x_\star \in X_\star, x \in \mathbb{R}^d} \gamma_\mathcal{P}(x_k - x_\star), \text{s.t.} f(x) \leqslant f(x_0)$ is assumed to be finite. Denoting $\omega = \frac{1}{L_\mathcal{P} \mathcal{R}^2}$, and dividing by δ_k, we have that $\frac{1}{\delta_k} + \omega \frac{\delta_k}{\delta_{k+1}} \leqslant \frac{1}{\delta_{k+1}}$. Since δ_k is nonincreasing with k, $\frac{1}{\delta_{k+1}} \geqslant \frac{1}{\delta_k} + \omega$. By summation, $\frac{1}{\delta_k} \geqslant \omega(t-1)$ and the result follows. □

3.E Steepest coordinate descent is 'nearly' a matching pursuit algorithm

Coordinate descent with a Gauss-Southwell rule can be formulated using an LMO. Indeed, recall its formulation,

$$\alpha_{k+1} = \arg\min_{\alpha \in \mathbb{R}^d} \langle \nabla_{i_k} F(\alpha_k) e_{i_k}, \alpha - \alpha_k \rangle + \frac{L^F_2}{2} \|\alpha - \alpha_k\|_2^2 + \lambda |\alpha_{i_k}|,$$

where $i_k = \arg\min_k \min_{t\in\mathbb{R}} P_{:,k}^\top \nabla f(P\alpha)(t-\alpha_k) + \frac{L_2^F}{2}(t-\alpha_k)^2 + \lambda|t|$. We introduce the gauge function $\gamma_{\mathcal{P}}(x) = \inf_{\alpha\in(\mathbb{R}^+)^d} \sum_{i=1}^d \alpha_i P_i$. Then, applying the GS-rule can be formulated as,

$$\begin{aligned} i_k &= \arg\min_k \min_{u+\alpha_k\geqslant 0} P_{:,k}^\top \nabla f(P\alpha)u + \frac{L_2^F}{2}u^2 + \lambda|u+\alpha_k|, \\ &= \arg\min_k \Delta_k = \frac{1}{2L_2^F}(P_{:,k}^\top \nabla f(P\alpha) + \lambda - L_2^F\alpha_k)_+^2 - \frac{1}{2L_2^F}(P_{:,k}^\top \nabla f(P\alpha)+\lambda)^2. \end{aligned}$$

Let $\mathcal{P}$ be the set of atoms, $\mathcal{P}_k = \{k, \alpha_k \neq 0\}$ the set of visited atoms at iteration k. The algorithm consists in computing:

- for non visited atoms $k \in \mathcal{P}\setminus\mathcal{P}$ ($\alpha_k = 0$), $\Delta_k = -\frac{1}{2L_2^F}(-P_{:,k}^\top \nabla f(P\alpha) - \lambda)_+^2$ and this value can be computed using the $\mathrm{LMO}_{\mathcal{P}\setminus\mathcal{P}_k}(\nabla f(P\alpha_k)) = p_{\text{out}}$, leading to $\Delta_{\text{out}} = \frac{1}{2L_2^F}(-p_{\text{out}}^\top \nabla f(P\alpha_k) - \lambda)_+^2$, which costs $O(|\mathcal{P}\setminus\mathcal{P}_k|)$,
- for visited atoms, the objective needs to be computed completely in at most $|\mathcal{P}_k|$ iterations: $\Delta_{\text{in}} = \frac{1}{2L_2^F}\sup_{p\in\mathcal{P}_k}(p^\top\nabla f(P\alpha_k)+\lambda)^2 - (p^\top\nabla f(P\alpha_k)+\lambda - L\alpha_k^p)_+^2$.

Then, we compute i_k corresponding to the minimizer of $\min(\Delta_{\text{in}}, \Delta_{\text{out}})$, and compute the update $\alpha_{k+1}^{i_k} = (\alpha_k^{i_k} - \frac{1}{L_2^F}(p_{i_k}^\top \nabla f(P\alpha_k) + \lambda))_+$.

3.F The ultimate method: an inner loop strategy

The ultimate method is defined as a minimization problem, that has no closed-form solution. To overcome this issue, at each iteration k, we will solve iteratively the inner minimization problem using a randomized alternating minimization technique:

$$\min_{\eta,\beta\in\mathbb{R}^d} \langle P^\top\nabla f(P\alpha), \beta-\alpha\rangle + \frac{L_{\gamma_{\mathcal{P}}}^f}{2}\|\beta-\alpha\|_1^2 + \lambda\|\nu\|_1, \text{ such that } P\beta = P\nu.$$

First, we give the guarantees of an inner loop strategy, and in a second part, the result when applying a randomized alternating minimization technique to the inner loop.

Let us provide the convergence guarantees of the inner loop strategy, following the example of d'Aspremont et al. [2021, Section 5.2]. Let $L(\beta,\nu) = \langle P^\top\nabla f(P\alpha), \beta-\alpha\rangle + \frac{L_{\gamma_{\mathcal{P}}}^f}{2}\|\beta-\alpha\|_1^2 + \lambda\|\nu\|_1$ and $L_\star^\alpha = \min_{\eta,\beta\in\mathbb{R}^d} L(\beta,\nu)$, such that $P\beta = P\nu$. Given an approximate solution to this problem at iteration k, Theorem 24 provides a guarantee on the outer loop.

Theorem 24. *Let $(\tilde\beta_k, \tilde\nu_k)$ with $\tilde x_k = P\tilde\beta_k = P\tilde\nu_k$ be an approximate solution of x_k produced by the ultimate method (3.8) such that $L^k(\tilde\beta_k,\tilde\nu_k) - L_\star^k \leqslant \epsilon_k$, then we have*

$$f(\tilde x_k) - f_\star \leqslant \left(1 - \frac{\mu_{\gamma_{\mathcal{P}}}^f}{L_{\gamma_{\mathcal{P}}}^f}\right)^k (f(x_0) - f_\star) + \sum_{i=0}^{k-1}\left(1-\frac{\mu_{\gamma_{\mathcal{P}}}^f}{L_{\gamma_{\mathcal{P}}}^f}\right)^i \epsilon_{k-i}.$$

Proof.

$$\begin{aligned} f(\tilde x_{k+1}) &\leqslant f(\tilde x_k) + L^k(\tilde\eta_{k+1}, \tilde\beta_{k+1}), \\ &\leqslant f(\tilde x_k) + L_\star^k + \epsilon_k, \\ &\leqslant f(\tilde x_k) - \frac{\mu_{\gamma_{\mathcal{P}}}^f}{L_{\gamma_{\mathcal{P}}}^f}(f(\tilde x_k) - f_\star) + \epsilon_k, \\ f(\tilde x_{k+1}) - f_\star &\leqslant \left(1 - \frac{\mu_{\gamma_{\mathcal{P}}}^f}{L_{\gamma_{\mathcal{P}}}^f}\right)(f(\tilde x_k) - f_\star) + \epsilon_k. \end{aligned}$$

The result is obtained by a direct summation. □

The precision of the outer loop depends on the precision ϵ_k of the inner loop at each iteration k. Given an iterative method with iteration number t to obtain $(\tilde{\beta}_k^t, \tilde{\nu}_k^t)$, Theorem 24 ensures that the better the precision of the inner loop, the better the convergence guarantee of the outer loop. More precisely, if the error is constant $\epsilon_k = \epsilon$, the global convergence guarantee becomes $f(\tilde{x}_k) - f_\star \leqslant \left(1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}}\right)^k (f(x_0) - f_\star) + \epsilon \frac{L^f_{\gamma\mathcal{P}}}{\mu^f_{\gamma\mathcal{P}}} \left(1 - (1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}})^k\right)$. For a linearly decreasing inner precision $\epsilon_i = \left(1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}}\right)^i$, the global convergence rate is exactly $1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}}$. Finally, for a sublinearly inner precision, the global convergence guarantee is given by $F(\tilde{x}_k) - F_\star \leqslant \left(1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}}\right)^k (F(x_0) - F_\star) + \sum_{i=0}^k -1 \left(1 - \frac{\mu^f_{\gamma\mathcal{P}}}{L^f_{\gamma\mathcal{P}}}\right)^i \frac{1}{(k-i+1)^\alpha}$. The inner precision ϵ_k can slow down the global convergence of the ultimate method. However, achieving a low precision in the inner loop can be very costly, especially if the inner method converges sublinearly.

3.F.1 Alternating randomized block coordinate descent

We propose to solve this minimization problem that defines matching pursuit using an alternating minimization technique. Diakonikolas and Orecchia [2018] developed the alternating randomized block coordinate descent (AR-BCD), that generalizes alternating minimization to more than two blocks.

Theorem 25. *[Diakonikolas and Orecchia, 2018, Theorem 3.4] Let x_k be generated by AR-BCD with a distribution $(p_i)_{i=1,\dots,n-1}$ over $n-1$ $L^{f,i}$-smooth blocks, and n be the non-smooth block. Then, for $R_i = \max_{x\in\mathbb{R}^d, f(x)\leqslant f(x_0)} \|x_{\star,i} - x_i\|^2$,*

$$\mathbb{E}[f(x_{k+1})] - f_\star \leqslant \frac{2}{k+3} \left(\sum_{i=1}^{n-1} \frac{L^{f,i}}{p_i} R_i \right).$$

The AR-BCD algorithm converges sublinearly for a non-smooth objective, according to Theorem 25. Let us first reformulate the minimization problem as an unconstrained problem, calling $\alpha = \alpha_k$:

$$\begin{aligned}
&\min_{\eta,\beta\in\mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \beta - \alpha\rangle + \frac{L^f_{\gamma\mathcal{P}}}{2}\|\beta - \alpha\|_1^2 + \lambda\|\eta\|_1, \text{ such that } P\beta = P\eta,\\
&= \min_{\eta,k\in\mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \eta - \alpha\rangle + \frac{L^f_{\gamma\mathcal{P}}}{2}\|\eta + k - \alpha\|_1^2 + \lambda\|\eta\|_1, \text{ such that } k \in \mathrm{Ker}(P),\\
&= \min_{\eta\in\mathbb{R}^d, z\in\mathbb{R}^{d_Q}} \langle P^\top \nabla f(P\alpha), \eta - \alpha\rangle + \frac{L^f_{\gamma\mathcal{P}}}{2}\|\eta + Qz - \alpha\|_1^2 + \lambda\|\eta\|_1,
\end{aligned}$$

where Q is a basis for $\mathrm{Ker}(P)$ and $d_Q = \dim(\mathrm{Ker(Q)})$, obtained using a QR decomposition. We now use the eta-trick on $\|\cdot\|_1^2$ and an other η-trick on $\|\cdot\|_1$ (see 3.A.1) , which leads to the equivalent minimization problem,

$$\begin{aligned}
\min_{\eta\in\mathbb{R}^d, z\in\mathbb{R}^r} \min_{\gamma\in\Delta_d, \theta\geqslant 0} G(\eta, z, \gamma, \theta) = \langle P^\top \nabla f(P\alpha), \eta - \alpha\rangle + \frac{\lambda}{2}\left(\eta^\top \mathrm{Diag}(\theta)^{-1}\eta + \mathrm{Diag}(\theta)1\right)\\
+ \frac{L^f_{\gamma\mathcal{P}}}{2}(\eta + Qz - \alpha)^\top \mathrm{Diag}(\gamma)^{-1}(\eta + Qz - \alpha).
\end{aligned}$$

Remark 8. *Note that without applying the second η-trick on $\|\cdot\|_1$, the objective function is*

smooth with respect to z, no non-smooth with respect to η, γ, which prevents us from using the alternating minimization technique.

Each coordinate can be solved as follows:

$$\begin{aligned}
z_{opt} &= \left(Q^\top \mathrm{Diag}(\gamma_k)^{-1} Q\right)^{-1} Q^\top \mathrm{Diag}(\gamma_k)^{-1}(\alpha - \eta), \\
\eta_{opt} &= \mathrm{Diag}(L^f_{\gamma_\mathcal{P}}/\gamma + \lambda/\theta)^{-1}\left(L^f_{\gamma_\mathcal{P}} \mathrm{Diag}(\gamma)^{-1}(\alpha - Qz) - P^\top \nabla f(P\alpha)\right), \\
\gamma_{opt} &= \frac{|\eta^i + (Qz)^i - \alpha^i|}{\|\eta + (Qz) - \alpha\|_1}, \\
\theta_{opt} &= |\eta|.
\end{aligned}$$

Let us rewrite our problem into $\min_{\eta\in\mathbb{R}^d, z\in\mathbb{R}^r} \min_{\gamma\in\Delta_d, \theta\geqslant 0} G(\eta, z, \gamma, \theta) = G(\eta, z, \beta) = G(\xi)$ where $G(\cdot)$ is smooth with respect to η, z but non smooth with respect to $\beta = (\gamma, \theta)^\top$. We perform AR-BCD with probabilities p_1 for η (respectively $p_2 = 1 - p_1$ for z). Let us rewrite $(\eta, z, \beta) = (\eta_1, \eta_2, \eta_3)$, and let $S_i(\xi)$ be the set of points that differs from ξ only over block i. AR-BCD is given at iteration k by:

$$\begin{aligned}
&\text{Pick } i_k \in \{1, 2\} \text{ with probability } p_{i_k}, \\
\tilde{\xi}_{k+1} &= \operatorname*{arg\,min}_{\xi\in S_{i_k}(\xi_k)} G(\xi), \\
\xi_{k+1} &= \operatorname*{arg\,min}_{\xi\in S_3(\tilde{\xi}_{k+1})} G(\xi).
\end{aligned}$$

3.F.2 Alternating minimization

$$\min_{\eta\in\mathbb{R}^d, z\in\mathbb{R}^{dQ}} \min_{\gamma\in\Delta_d} \langle P^\top \nabla f(P\alpha), \eta - \alpha\rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2}(\eta + Qz - \alpha)^\top \mathrm{Diag}(\gamma)^{-1}(\eta + Qz - \alpha) + \lambda\|\eta\|_1.$$

The objective function $F(\eta, z, \gamma)$ is jointly convex in (η, z, γ) (for $\gamma > 0$). This problem can be solved using alternating minimization (which is stronger than coordinate descent). Starting from $\eta_0 \in \mathbb{R}^d, z_0 \in \mathbb{R}^r (r = d - \mathrm{rg}(P) = r, \gamma_0 \in \Delta_d$,

$$\begin{aligned}
z_{k+1} &= \operatorname*{arg\,min}_z F(\eta_k, z, \gamma_k), \\
\gamma_{k+1} &= \operatorname*{arg\,min}_\gamma F(\eta_k, z_{k+1}, \gamma), \\
\eta_{k+1} &= \operatorname*{arg\,min}_\eta F(\eta, z_{k+1}, \gamma_{k+1}).
\end{aligned}$$

In this context, it corresponds to computing

$$\begin{aligned}
z_{k+1} &= \left(Q^\top \mathrm{Diag}(\gamma_k)^{-1} Q\right)^{-1} Q^\top \mathrm{Diag}(\gamma_k)^{-1}(\alpha - \eta_k), \\
\eta_{k+1} &= S_{\lambda/L^f_{\gamma_\mathcal{P}}\gamma_k}\left(\alpha - \frac{1}{L^f_{\gamma_\mathcal{P}}} \mathrm{Diag}(\gamma_k) P^\top \nabla f(P\alpha) - Qz)\right), \\
\gamma_{k+1} &= \frac{|\eta^i_{k+1} + (Qz_{k+1})^i - \alpha^i|}{\|\eta_{k+1} + (Qz_{k+1}) - \alpha\|_1}.
\end{aligned}$$

3.F.3 Experimental results

In Figure 3.9, we apply an alternating randomized block coordinate descent (AR-BCD) technique to solve a well-chosen inner-loop optimization problem, which was developed by Diakonikolas

and Orecchia [2018]. The inner loop strategy is developed in 3.F.1, as well as its convergence guarantee. We also derive an alternating minimization method on three blocks, as detailed in 3.F.2, simpler than AR-BCD, for which there is no convergence guarantee.

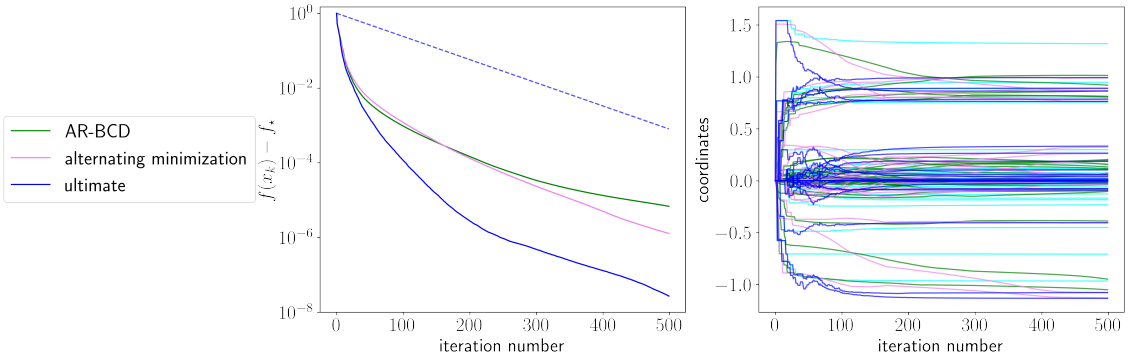


Figure 3.9: Convergence in function value for the ultimate method computed with an alternating minimization technique, AR-BCD, and with the solver MOSEK, compared to the regularized matching pursuit. Parameters are given by $d = 50$, $n = 20$, $s = 8$ and $\lambda = 0.001$, with an inner loop with exponential precision.

3.G Constrained matching pursuit.

In this section, we concentrate on a constrained minimization problem:

$$\min_{x, \gamma_{\mathcal{P}}(x) \leqslant 1} f(x), \tag{3.23}$$

where $\gamma_{\mathcal{P}}$ is gauge function as defined in Section 3.2.4 for the symmetric set $\mathcal{P} = \text{conv}\{P_{:,i}, i = 1, \dots, d\}$ and where f is $L_{\gamma_{\mathcal{P}}}$-smooth and $\mu_{\gamma_{\mathcal{P}}}$-strongly convex. Again, we reformulate the problem in $\mathbb{R}^d$:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n,\ \alpha \in (\mathbb{R}^+)^d} f(x), \text{ such that } x = P\alpha \text{ and } \sum_{i=1}^{d} \alpha^{(i)} \leqslant 1, \\ &= \min_{\alpha \in \Delta_d} F(\alpha), \end{aligned} \tag{3.24}$$

where $\Delta_d = \{\alpha \in \mathbb{R}^d, \sum_{i=1}^d \alpha^{(i)} = 1, 0 \leqslant \alpha_i\}$ is the simplex in $\mathbb{R}^d$, and where $F(\cdot) = f(P\cdot)$ is L_1^F-smooth and μ_1^F-strongly convex with respect to the ℓ_1-norm.

As for the regularized matching pursuit (3.16) and the ultimate method 3.20, we formulate an optimization method as the minimization of the smoothness upper bound with respect to the ℓ_1-norm:

$$\alpha_{k+1} = \min_{\alpha \in \Delta_d, \alpha_k \in \Delta_d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2. \tag{3.25}$$

As we will see, this minimization procedure leads to a matching pursuit procedure, that can be compared to variants of the Frank-Wolfe algorithm [Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015, Jaggi, 2013]. It is therefore denoted *constrained matching pursuit*. In addition, its convergence guarantees can be derived from the convergence analyses of the regularized matching pursuit (3.16) and the ultimate method 3.20.

3.G.1 Constrained matching pursuit

In the following, we reformulate the update defined in (3.25) as a matching pursuit algorithm.

Calculation to derive the constrained matching pursuit algorithm. In what follows, we derive the steps leading to the constrained matching pursuit Algorithm 2. Denoting $p^{(i)} = P_{:,i} \in \mathbb{R}^n$ the rows of P, the update takes the form:

$$\min_{\beta, \sum_i \beta^{(i)}=0,\ \alpha_k^{(i)}+\beta^{(i)} \geqslant 0} \sum_i \langle \nabla f(x_k), p^{(i)} \beta^{(i)} \rangle + \frac{L}{2} \|\beta\|_1^2. \tag{3.26}$$

To solve this problem, we propose to first relax the constraint $\alpha_k^{(i)} + \beta^{(i)} \geqslant 0$, and then enforce it. The minimization problem (3.26) takes the form:

$$\begin{aligned}
&\max_{z \geqslant 0} \frac{-z^2}{2L} + \min_{\beta, \sum_i \beta^{(i)}=0} \sum_i \langle \nabla f(x_k) p^{(i)} \beta^{(i)} \rangle + z|\beta^{(i)}|, \\
&= \max_{z \geqslant 0, \lambda} \frac{-z^2}{2L} + \min_{\beta} \sum_i \langle \nabla f(x_k) p^{(i)} \beta^{(i)} \rangle + z|\beta^{(i)}| + \lambda \beta^{(i)}, \\
&= \max_{z, \geqslant 0, \lambda} \frac{-z^2}{2L}, \quad \text{such that } z \geqslant \sup_i |\lambda + \nabla f(x_k)^\top p^{(i)}|, \\
&= \min_{\lambda} -\frac{1}{2L} \left(\sup_i |\lambda + \nabla f(x_k)^\top p^{(i)}| \right)^2,
\end{aligned}$$

whose optimum is achieved at $z_\star = \frac{\sup_i \langle \nabla f(x), p^{(i)} \rangle - \inf_i \langle \nabla f(x_k), p^{(i)} \rangle}{2}$ and $\lambda_\star = -\frac{\sup_i \langle \nabla f(x_k), p^{(i)} \rangle + \inf_i \langle \nabla f(x_k), p^{(i)} \rangle}{2}$, where $z_\star = L\|\beta\|_1$.

Remark 9. *This formulation is closely related to the linear minimization oracle (LMO), that is defined for all $q \in \mathbb{R}^n$ by:*

$$\mathrm{LMO}_{\mathcal{P}}(q) = \arg\min_{p \in \mathcal{P}} \langle p, q \rangle.$$

Indeed, it holds that $\max_{p \in \mathcal{P}} \langle \nabla f(x_k), p \rangle = \min_{p \in \mathcal{P}} \nabla f(x_k), p \rangle$ by symmetry of $\mathcal{P}$. Among others, the Frank-Wolfe algorithm [Frank and Wolfe, 1956] and its variants [Lacoste-Julien and Jaggi, 2015, Locatello et al., 2017] query such an LMO at each iteration.

Denoting I_+ and I_- the ensemble such that $\beta_{I_+} \geqslant 0$ and $\beta_{I_-} \leqslant 0$, it follows that:

$$z_\star = L(\|\beta_{I_-}\|_1 + \|\beta_{I_+}\|_1) = 2L \sum_{i \in I_+} \beta^{(i)} = -\sum_{i \in I_-} \beta^{(i)}.$$

We then determine the sign of $\beta^{(i)}$ using the rule:

- $\beta^{(i)} \leqslant 0$ if $\langle \nabla f(x_k), p^{(i)} \rangle + \lambda - z \geqslant 0$ (i.e. $\langle \nabla f(x_k), p^{(i)} \rangle = \sup_j \langle \nabla f(x_k), p^{(j)} \rangle$, and $i \in I_-$),
- $\beta^{(i)} \geqslant 0$ if $\langle \nabla f(x_k), p^{(i)} \rangle + \lambda + z \leqslant 0$ (i.e. $\langle \nabla f(x_k), p^{(i)} \rangle = \inf_j \langle \nabla f(x_k), p^{(j)} \rangle$, and $i \in I_+$).

Finally, we add the constraint $\alpha_k^{(i)} + \beta^{(i)} \geqslant 0$ for selecting the directions to update. Indeed, when the LMO is attained in several atoms $p^{(i)}$, Problem (3.26) may admit several solutions. We propose a procedure to compute $\beta^{(i)}$ by first arranging $i \in I_-$ such that $\alpha_k^{(i_0)} \geqslant \ldots \geqslant \alpha_k^{(i_1)}$. Then, For all $i \in I_-$ such that $\alpha_k^{(i_0)} = \alpha_k^{(i)}$, we have a separable minimization problem in $\beta^{(i)}$. Therefore we assume that all of such $\beta^{(i)}$ are equals and compute $\beta^{(i)} = (\alpha_k^{(i)} - \frac{z_\star}{2KL})_+ - \alpha_k^{(i)}$ where $K = \mathrm{Card}\{i \in I_-, \alpha_k^{(i)} = \alpha_k^{(i_0)}\}$. Finally, if $\alpha_k^{(i_0)} < \frac{z_\star}{2KL}$, perform the same operation for i_1 and $\sum_{i \in I_-, \alpha_k^{(i)} = \alpha_k^{(i_1)}} \beta^{(i)} = -\frac{z_\star}{2L} + K\alpha_k^{(i_0)}$.

From these steps, we deduce the constrained matching pursuit, formulated in Algorithm 2.

Algorithm 2 Constrained matching pursuit (CMP)

Initial point: $x_0 \in \mathcal{P}$
for $k \in [1, ..., N]$ **do**
  Decompose $x_k = \sum_{i \in I_k} \alpha_k^{(i)} p^{(i)}$ with $\alpha^{(i)} > 0$
  $\{s_k^i\}_{i \in I_+} = \text{LMO}_{\mathcal{P}}(\nabla f(x_k))$
  $\{v_k^{(i)}\}_{i \in I_-} = \arg\max_{p \in I_k} \langle \nabla f(x_k), p \rangle$
  $z_\star = \frac{1}{2} \langle \nabla f(x_k), v_k - s_k \rangle$
  Put in order I_- such that $\alpha_k^{(i_0)} \leqslant \ldots \geqslant \alpha_k^{(i_d)}$ such that $\langle \nabla f(x_k), p^{(i)} \rangle = \langle \nabla f(x_k), v_k \rangle$ (maximal direction)
  Put in order I_+ such that $\alpha_k^{(j_0)} \leqslant \ldots \leqslant \alpha_k^{(j_d)}$ such that $\langle \nabla f(x_k), p^{(j)} \rangle = \langle \nabla f(x_k), s_k \rangle$ (minimal direction)
  $z_{\text{neg}} = \frac{z_\star}{2L \text{Card}\left(i \in I_-, \alpha_k^{(i)} = \alpha_k^{(i_0)} \geqslant 0\right)}$
  Compute first the maximal modification by considering I_-.
  if $\alpha_k^{(i_0)} - z_{\text{neg}} \geqslant 0$ **then**
    $\beta_{\text{neg}}^{(i_0)} = -z_{\text{neg}}$
  else
    while $\alpha_k^{(i_l)} - z_{\text{neg}} < 0$ starting from $l = 0$ and $z_{\text{neg}} \geqslant 0$ **do**
      for $i \in I_-$ such that $\alpha_k^{(i)} = \alpha_k^{(i_l)}$ **do**
        $\beta_{\text{neg}}^{(i)} = -\alpha_k^{(i_l)}$
      end for
      $z_{\text{neg}} = \frac{\text{Card}\left(i \in I_-, \alpha^{(i)} = \alpha_k^{(i_l)}\right)}{\text{Card}\left(i \in I_-, \alpha_k^{(i_l)} = \alpha_k^{(i_{l+1})}\right)} (z_{\text{neg}} - \alpha_k^{(i_l)})$
      $l = l + 1$
    end while
    if $\alpha_k^{(i_l)} - z_{\text{neg}} \geqslant 0$ and $z_{\text{neg}} > 0$ **then**
      for $i \in I_-$ such that $\alpha_k^{(i)} = \alpha_k^{(i_l)}$ **do**
        $\beta_{\text{neg}}^{(i_l)} = -z_{\text{neg}}$
      end for
    end if
  end if
  Compute the update for I_+
  $z_+ = -\sum_{i \in I_-} \beta^{(i),-}$
  $\beta_{pos}^{(i_0)} = z_+$ (take the minimal element in I_+)
  Compute the final update : $x = x + \sum_{i \in I} \beta^{(i)} p^{(i)}$
end for

Algorithm 2 is a matching pursuit procedure: it incrementally selects a limited number of directions, identified by the linear minimization procedure, and modifies the already selected atoms, potentially setting some of them to zero. To understand the intuition behind this method, consider a scenario where only one direction is chosen at which the LMO is attained. This leads to a simplified version of the constrained matching pursuit Algorithm 3.

Algorithm 3 selects only two directions : the direction identified by the LMO at $\nabla f(x_k)$, and a direction identified using the already selected atoms $\arg\max_{i \in I_k} \langle \nabla f(x_k), p_i \rangle$. It turns out that this simplified version corresponds to the Pairwise Frank-Wolfe Algorithm (PWA), as introduced by Lacoste-Julien and Jaggi [2015, Algorithm 2], except that PWA exhibits a linesearch strategy after identifying the direction $s_k - v_k$ and let a parameter L_2^F instead of L_1^F. Similarly, Pedregosa et al. [2020] provided a backtracking linesearch using the same update direction.

Algorithm 3 Simplified constrained matching pursuit

Initial point: $x_0 \in \mathcal{P}$
for $k \in [0, ..., N]$ **do**
$x_k = \sum_{i \in I_k} \alpha^{(i)} p^{(i)}$ where $I_k = \{i \in [1, \dots, d], \alpha_k^{(i)} > 0\}$
$s_k = \mathrm{LMO}_{\mathcal{P}}(\nabla f(x_k)) = \mathrm{argmin}_{p \in \mathcal{P}} \langle \nabla f(x_k), p \rangle$
$v_k = \arg\max_{i \in I_k} \langle \nabla f(x_k), p_i \rangle$
$z_\star = \frac{1}{2} \langle \nabla f(x_k), v_k - s_k \rangle$
$x_k = x_k + \left((\alpha_{v_k} - \frac{z_\star}{2L_2^F})_+ - \alpha_{v_k} \right) (s_k - v_k)$
end for

3.G.2 Convergence guarantees

This section is devoted to deriving convergence guarantees for the constrained matching pursuit algorithm, and comparing it to the convergence guarantee of PFW [Lacoste-Julien and Jaggi, 2015]. It appears that its convergence guarantee can be deduced from the convergence analysis for the ultimate method (3.20) and the generalized regularized matching pursuit (3.19).

First, let us reformulate Problem 3.23 as a penalized problem:

$$\min_{\alpha \in \mathbb{R}^d} F(\alpha) + \phi(\|\alpha\|_1),$$

$\phi(t) = i_{t \leqslant 1}$ is an indicator function (by symmetry of $\mathcal{P}$). In particular, $\phi \dot{\|} \cdot \|_1$ is a closed convex and proper function. Thereby, Proposition 7 and Proposition 8 extend directly to the constrained matching pursuit algorithm, as stated in the following corollary.

Corollary 8. *Let F be L_1^F-smooth and convex with respect to the ℓ_1-norm. Then, the sequence (α_k) generated by 3.25 verifies:*

$$F(\alpha_k) - F_\star \leqslant \frac{2L_1^F \mathcal{R}_{\alpha_0}^2}{k+1},$$

where $\mathcal{R}_{\alpha_0}^2 = \max_{\alpha \in \mathbb{R}^d} \max_{\alpha_\star \in \mathbb{R}^d} \{\|\alpha - \alpha_\star\|_1^2, F(\alpha_k) \leqslant F(\alpha_0)\}$.

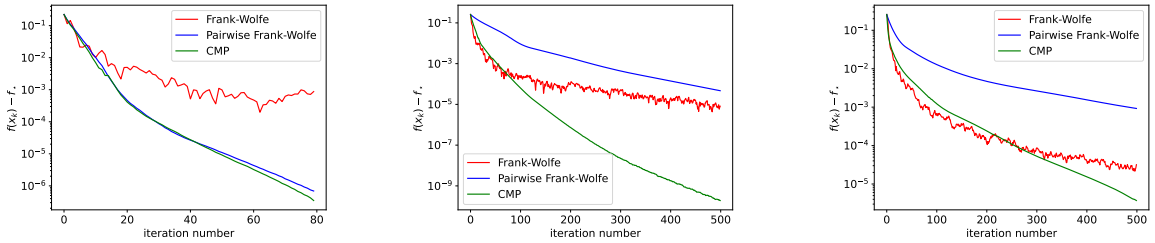
If in addition, F is μ_1^F-strongly convex with respect to the ℓ_1-norm, then the sequence (α_k) verifies:

$$F(\alpha_{k+1}) - F_\star \leqslant (1 - \frac{\mu_1^F}{L_1^F})(F(\alpha_k) - F_\star).$$

Compared to Pairwise Frank-Wolfe, our approach benefits from a more comprehensive approach of the proof, that derives mostly from the construction of the method as the minimization of the smoothness quadratic upper bound. In their convergence analysis, Lacoste-Julien and Jaggi [2015, Section 3] introduced the pyramidal width, that can be interpreted as the condition number of $\mathcal{P}$ and is often intractable. For simple cases, such as the simplex or the unit ball, the authors exhibited a $O(\frac{1}{d})$-term [Lacoste-Julien and Jaggi, 2015, Appendix B.1]. While it is difficult to compare this pyramidal width to the exact value for μ_1^F, our approach benefits from refined approximation of μ_1^F, as proven in Section 3.2.2. Figure 3.10 provides a numerical experiment for strongly-convex functions F.

As expected, Figure 3.10 shows a linear convergence behavior in the underparametrized regime. Surprisingly yet, we observe a similar behavior in the overparametrized regime even for a nonsparse solution (we expect a sublinear convergence guarantee from Corollary 8). We left to a future work the understanding of this behavior.

As for the regularized matching pursuit, the constrained matching pursuit (3.25) does not converge linearly when f is $L_{\gamma_{\mathcal{P}}}$-smooth and $\mu_{\gamma_{\mathcal{P}}}$-strongly convex. Indeed, the *ultimate con-*



(a) $d = 50$, $n = 100$, $s = 8$. (b) $d = 500$, $n = 50$, $s = 8$. (c) $d = 500$, $n = 50$, $s = 250$.

Figure 3.10: Convergence in function values of the constrained matching pursuit, Pairwise Frank-Wolfe and Frank-Wolfe for a linear regression problem with a noise $\sigma = 0.5$ and a regularization with $\mu = 0.001$ as a function of the iteration number.

strained method benefiting from linear convergence properties in the overparametrized regime would be:

$$x_{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min} \ \langle \nabla f(x_k), x - x_k \rangle + \frac{L_{\gamma_\mathcal{P}}}{2} \gamma_\mathcal{P}(x - x_k)^2,$$
$$\text{such that } \gamma_\mathcal{P}(x) \leqslant 1.$$

Reformulated in $\mathbb{R}^d$, this minimization problem is equivalent with:

$$\min_{\alpha \in \mathbb{R}^d} \ \langle F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_{\gamma_\mathcal{P}}}{2} \|\alpha - \alpha_k\|_1^2,$$
$$\text{such that } \|\nu\|_1 \leqslant 1,$$
$$P\alpha = P\nu.$$

Again, the constraint $P\alpha = P\nu$ simplifies to $\alpha = \nu$ in the underparametrized regime, but not in the overparametrized regime. The further study of this ultimate constrained method is left to future works.

CHAPTER 4

Constructive approaches to concentration inequalities with independent random variables

Objectives

Concentration inequalities, a major tool in probability theory, quantify how much a random variable deviates from a certain quantity. This chapter proposes a systematic convex optimization approach to studying and generating concentration inequalities with independent random variables. Specifically, we extend the generalized problem of moments to independent random variables.
We first introduce a variational approach that extends classical moment-generating functions, focusing particularly on first-order moment conditions. Second, we develop a polynomial approach, based on a hierarchy of sum-of-square approximations, to extend these techniques to higher-moment conditions. Building on these advancements, we refine Hoeffding's, Bennett's and Bernstein's inequalities, providing improved worst-case guarantees compared to existing results.

This chapter is submitted to *Mathematics of Operations Research*.

Codes. All codes are provided at `https://github.com/CMoucer/ConcentrationInequalities`. We use standard solvers SCS [O'Donoghue et al., 2016] and MOSEK [ApS, 2022].

Contents

2.1 Introduction **22**
2.1.1 Lyapunov functions 24
2.1.2 Prior works 24
2.1.3 Contributions and organization 25
2.1.4 Assumptions 25
2.2 A principled approach to Lyapunov functions for gradient flows **26**
2.2.1 The gradient flow 26
2.2.2 Accelerated gradient flows 30
2.2.3 Higher-order convergence and time dilation 34
2.3 SDEs for modeling SGD **36**
2.3.1 Lyapunov functions do not always extend to the stochastic setting 37
2.3.2 Diminishing the step size is a key to success 39

2.3.3 Averaging for larger step sizes . . . 40
2.4 Accelerating the gradient flow . . . 43
2.4.1 Averaging does not preserve convergence rates . . . 43
2.4.2 Diminishing step sizes do not help preserve acceleration . . . 44
2.5 Conclusion and future work . . . 45
Appendices . . . 46
2.A Proof for ODEs . . . 46
2.A.1 Proofs for Theorem 11 . . . 46
2.A.2 Proof for Corollary 3 . . . 46
2.B Proofs for SDEs . . . 47
2.B.1 Proof for Theorem 15 . . . 47
2.B.2 Proof for Theorem 17 . . . 48
2.C Strongly convex regularization . . . 48
2.D Using PEPit for computing worst-case guarantees . . . 50

Introduction.

Concentration inequalities have emerged as a major tool in probability theory, finding applications in learning theory [Devroye et al., 1996], in random matrix theory [Tao, 2012] or statistical physics or mechanics [Dembo and Zeitouni, 1998]. These inequalities quantify how much a random variable deviates from a certain quantity, usually its mean. Classical examples include Markov's inequality for bounding probabilities of deviations from 0 or Chebyshev's inequalities for deviation from the mean. In machine learning and statistics, where the data are often assumed to be independent and identically distributed (i.i.d.), basic inequalities such as Hoeffding's inequality, Bennett's inequality or Bernstein's inequality [Boucheron et al., 2013, Chapter 6] are extensively used, e.g., for characterizing generalization properties of machine learning algorithms [Bach, 2023b]. For instance, Hoeffding's inequality [Hoeffding, 1963] states that for $X_1, \dots, X_n$ i.i.d. random variables taking their values almost surely in $[0,1]$, the sum $\sum_{i=1}^n X_i$ has a subgaussian tail with deviation $t \geqslant 0$:

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geqslant nt\right) \leqslant \exp\left(-2nt^2\right).$$

In this work, we provide a principled approach to concentration inequalities for independent univariate random variables with finite moments. Let $\mathcal{X}$ be a subset of $\mathbb{R}$ and $\mathcal{P}(\mathcal{X})$ be the set of distributions on $\mathcal{X}$. Given $X_1, \dots, X_n$ independent random variables generated from distributions $p_1, \dots, p_n \in \mathcal{P}(\mathcal{X})$, we formulate the generalized problem of moments for independent random variables as follows:

$$\begin{aligned} \rho_n &= \sup_{p_1,\dots,p_n \in \mathcal{P}(\mathcal{X})} \mathbb{E}_{p_1,\dots,p_n}[F(X_1,\dots,X_n)] \text{ such that } \forall i, \mathbb{E}_{p_i}[g_i(X_i)] = \mu_i, \\ &= \sup_{p_1,\dots,p_n \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}^n} F(x_1,\dots,x_n) dp_1(x_1) \cdots dp_n(x_n) \text{ such that } \forall i, \int_{\mathcal{X}} g_i(x_i) dp_i(x_i) = \mu_i, \end{aligned} \tag{4.1}$$

for some functions $F : x \in \mathcal{X}^n \mapsto \mathbb{R}^+$ and $g_i : x \in \mathcal{X} \mapsto \mathbb{R}^m$. For instance, Hoeffding's inequality involves $F(x) = \mathbf{1}_{\sum_{i=1}^n x_i \geqslant nt + \sum_{i=1}^n \mathbb{E}[X_i]}$, and therefore $\mathbb{E}_{p_1,\dots,p_n}[F(x)] = \mathbb{P}(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geqslant nt)$. Problem (4.1) is an infinite-dimensional non-convex problem. Without further assumptions on F and g, minimizing with respect to distributions p_i is often intractable.

This problem is closely related to the generalized problem of moments formalized by Lasserre [2008], which extends the traditional problem of moments [Landau, 1998] that seeks a measure matching a given set of moments. The search for optimal multivariate Chebyshev's inequalities began in the 1960s [Marshall and Olkin, 1960, Isii, 1962]. Isii [1962, 1964], along with Karlin

and Studden [1966], formalized the pursuit of sharp inequalities by framing it as an optimization problem. Let $\mathcal{Y} \subset \mathbb{R}^d$ and $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions on $\mathcal{Y}$. The generalized problem of moments takes the form

$$\begin{aligned} \rho &= \sup_{p \in \mathcal{P}(\mathcal{Y})} \mathbb{E}_p[F(X)], \text{ such that } \mathbb{E}_p[g(X)] = \mu \\ &= \sup_{p \in \mathcal{P}(\mathcal{Y})} \int_{\mathcal{Y}} F(x) dp(x) \text{ such that } \int_{\mathcal{Y}} g(x) dp(x) = \mu, \end{aligned} \tag{4.2}$$

where $g : \mathcal{Y} \mapsto \mathbb{R}^m$. Compared to Problem (4.1), the generalized problem of moments optimizes over distributions $p \in \mathcal{P}(\mathcal{Y})$ that are not necessarily products of their marginals. Under mild assumptions on the moment vector μ, strong duality holds [Bertsimas and Popescu, 2005, Theorem 2.2]. Its Lagrangian relaxation was first formulated by Isii [1964] as follows:

$$\begin{aligned} \rho &= \inf_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^m} \alpha + \beta^\top \mu \text{ such that } \forall x \in \mathcal{Y}, F(x) \leqslant \alpha + \beta^\top g(x), \\ &= \inf_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^m} \alpha + \beta^\top \mu + \sup_{x \in \mathcal{Y}} \{F(x) - (\alpha + \beta^\top g(x))\}. \end{aligned} \tag{4.3}$$

where $\alpha \in \mathbb{R}$ corresponds to the dual variable associated to the constraint $\int_{\mathcal{Y}} dp(x) = 1$, and $\beta \in \mathbb{R}^m$ to the constraint $\int_{\mathcal{Y}} g(x) dp(x) = \mu$. The dual Problem (4.3) is convex as it is expressed as the pointwise supremum of affine functions. Yet, it is unclear how to deal with the constraint "$\forall x \in \mathcal{Y}, F(x) \leqslant \alpha + \beta^\top g(x)$", because it corresponds to infinitely many linear constraints in (α, β).

Those problems are now traditionally approached via convex reformulations or approximations using semidefinite programming (SDP). Bertsimas and Popescu [2005] first investigated optimal bounds for $\mathbb{E}_p[F(X)] = \mathbb{P}(X \in S)$ assuming $\mathcal{X}$ and S to be semi-algebraic sets. For univariate random variables, they efficiently solved it using a single SDP, allowing them to derive tight bounds. For multivariate random variables, they proposed a series of semidefinite relaxations using sum-of-square representations (SoS). More generally, Lasserre [2008] investigated the generalized problem of moments and derived a hierarchy of SDPs converging to the optimal value (extending the methodology developed for approximating global optimization problems [Lasserre, 2001]). Simultaneously, Vandenberghe et al. [2007], Comanor et al. [2006] reformulated the generalized Chebyshev inequality as linear matrix inequalities (LMI) using an S-procedure.

These SDP-based approaches, along with tight guarantees, often allow reconstructing corresponding worst-case distributions [Bertsimas and Popescu, 2005, Section 5.1] [Vandenberghe et al., 2007, Section 2.2]. These extremal distributions turns out to be discrete [Rogosinski, 1958, Theorem 1] and may even be specified with $m + 2$ Dirac, where m is the number of constraints in (4.2). When the distributions in the problem under consideration are continuous with additional properties like symmetry or unimodality, these bounds may no longer be sharp. Therefore, Popescu [2005] generalized Chebyshev's inequality to convex classes of distributions generated by an appropriate parametric family of distributions. Building on Choquet's theory and conic duality, they provided an SDP reformulation (resp. approximation) of the generalized problem of moments for such univariate (resp. multivariate) distributions. From this framework, Van Parys et al. [2015] extended Gauss inequalities to multivariate unimodal distributions and outlined a methodology for computing worst-case unimodal distributions related to this problem.

Research questions and assumptions. Throughout this work, we assume $X_1, \ldots, X_n$ to be independent random variables in $\mathcal{X} \subset \mathbb{R}$ with finite moments $\forall i, \mathbb{E}_{p_i}[g(X_i)] = \mu_i \in \mathbb{R}^m$. For all i, we define $\mathcal{P}_{\mu_i}(\mathcal{X}) = \{p_i \in \mathcal{P}(\mathcal{X}), \int_{\mathcal{X}} g_i(x_i) dp_i(x_i) = \mu_i\}$ the set of (univariate) distributions on $\mathcal{X}$ with moment μ_i. If in addition, $X_1, \ldots, X_n$ follows the same distributions with moments $\mu_1 = \mu_2 = \cdots = \mu_n$, they are said to be independent and identically distributed (i.i.d.). We

assume that F is an indicator function, that is for S a compact semi-algebraic subset of $\mathcal{X}$, $F(x) = \mathbf{1}_{x \in S}$. The generalized moment problem for independent variables takes the form:

$$\begin{aligned} \rho_n &= \sup_{\forall i, p_i \in \mathcal{P}(\mathcal{X})} \int_{x \in \mathcal{X}^n} \mathbf{1}_{x \in S} dp_1(x_1) \cdots dp_n(x_n) \text{ such that } \forall i, \int_{\mathcal{X}} g(x_i) dp_i(x_i) = \mu_i, \\ &= \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{x \in \mathcal{X}^n} \mathbf{1}_{x \in S} dp_1(x_1) \cdots dp_n(x_n). \end{aligned} \tag{4.4}$$

In this formulation, independence is encoded by explicitly accounting for the constraint "$\forall x \in \mathcal{X}^n, q(x) = p_1(x_1) \cdots p_n(x_n)$". Classical concentration inequalities correspond to upper bounds to these problems for specific choices of g_i's. It is natural to wonder how tight such bounds are. To this end, we propose constructive and tractable approaches to upper bounding this problem, along with a comparison to existing bounds. Then, we raise the issue of reconstructing worst-case distributions that can potentially match these bounds in some scenarios.

Contributions. We propose a constructive approach for computing concentration inequalities of independent variables given a set of moments. To this end, we start by considering variational formulations to Problem (4.4):

$$\rho_n^{\mathcal{H}} = \inf_{H \in \mathcal{H}} \sup_{\forall i, p_i \in \mathcal{P}(\mu_i)} \int_{\mathcal{X}^n} H(x) dp_1(x) \cdots dp_n(x) \text{ such that } \forall x \in \mathcal{X}^n, F(x) \leqslant H(x), \tag{4.5}$$

where $\mathcal{H}$ is a well-chosen set of functions. It is straightforward that Problem (4.5) yields an upper bound to the generalized problem of moments for independent random variables, that is:

$$\rho_n \leqslant \rho_n^{\mathcal{H}}.$$

Efficiently leveraging the variational formulation (4.5) requires strategies to enforce the constraint $\forall x \in \mathcal{X}^n, F(x) \leqslant H(x)$ and to bound the objective. The choice of the function families $\mathcal{H}$ ensure these two requirements are satisfied. We propose two natural and complementary strategies based on convex optimization, each relying on different choices for $\mathcal{H}$.

1. The first strategy revolves around a family of product-functions $U(\cdot) = \prod_{i=1}^n u_i(x_i)$ inspired from classical probability proofs and variational probabilistic inference [Jaakkola and Jordan, 1999]. Given such a function, we define a *separable approach* by formulating Problem (4.4) as n univariate subproblems. We then develop a *variational approach* by optimizing over the family of product-functions. When F is log-convex and first-order moments are finite with $\forall i, \mathbb{E}[g_i(X_i)] = \mathbb{E}[X_i] = \mu_i$, the variational approach takes the form of a finite-dimensional convex optimization reformulation, that can be efficiently solved. This strategy significantly improves Hoeffding's inequality for small values of n and accurately meets the asymptotic large deviations for large n. Moreover, this approach allows the reconstruction of distributions involved at the optimum of the upper bound Problem (4.5). However, when it comes to higher-order finite moments, the variational approach formulates as a nonconvex problem that cannot be solved efficiently anymore.

2. The second strategy relies on a family of polynomial upper bounds, which is particularly suited to higher-moment conditions. Referred to as a *polynomial approach*, this strategy formulates as a non-convex optimization problem that can be effectively approximated by a series of sum-of-square formulation [Lasserre, 2008]. This approach contributes to refining Bernstein's and Bennett's inequality, which are fundamental probabilistic bounds.

3. Finally, we extend the polynomial approach to a *feature-based approach* relying on broader families of upper bounds $\mathcal{H}$. Compared to the variational approach, this method refines

Hoeffding's inequality using higher-order polynomials when applied to two random variables. While this methodology introduces finer approximations $\rho_n \leqslant \rho_n^{\mathcal{H}}$, it often requires well-chosen relaxations to approximate $\rho_n^{\mathcal{H}}$.

Outline of the chapter. Section 4.1 focuses on approximating Problem (4.4) with a variational approach from a family of product-functions, that is particularly suited to first-order moment assumptions. We formally compare these bounds to existing results. In Section 4.2, we study in-depth the variational approach associated with Hoeffding's inequality. Furthermore, we propose a methodology for reconstructing distributions that match the variational optimization problem in the worst-case scenario. Section 4.3 considers a family of polynomial upper bounds and approximate the resulting upper optimization problem using sum-of-square formulation. Thereby, we derive numerical evaluations of Bennett's and Bernstein's inequality. Finally, we introduce a feature-based framework that expands the scope of upper bounds families, providing a comprehensive approach to Hoeffding's inequality, that includes both the polynomial and variational methodologies.

4.1 A variational approach based on product-functions.

Classical functions $F(\cdot)$ usually correspond to the probability tails of a sum of independent random variables, specifically, $F(X) = \mathbf{1}_{\sum_{i=1}^n X_i \in S}$ representing $\mathbb{P}(\sum_{i=1}^n X_i \in S)$. This applies, among others, to Hoeffding's, Bennett's, and Bernstein's inequalities (see, e.g., [Boucheron et al., 2013, Bach, 2023b, Vershynin, 2018] and references therein). Their proofs typically rely on the Cramér-Chernoff technique, which essentially combines the exponential Chernoff's inequality with the independence of random variables. Specifically, they use moment-generating functions as follows

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geqslant nt\right) \leqslant \inf_{\lambda \geqslant 0} e^{-\lambda nt} \mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}] = \inf_{\lambda \geqslant 0} \prod_{i=1}^n e^{-\lambda(\mu_i + t)} \mathbb{E}[e^{\lambda X_i}],$$

and then optimize over $\lambda \geqslant 0$ for obtaining the smallest possible valid upper bound within this family. A natural, but richer, family of inequalities for obtaining concentration bounds involves constructing upper bounding functions as products of univariate functions, which are classical in probabilistic variational inference:

$$\mathcal{U} = \left\{ U : \mathcal{X}^n \mapsto \mathbb{R}^+ \text{ such that } \forall x \in \mathcal{X}^n, U(x) = \prod_{i=1}^n u_i(x_i) \text{ and } \forall i, u_i : \mathcal{X} \mapsto \mathbb{R}^+ \right\}. \tag{4.6}$$

In the Cramér-Chernoff technique, the functions u_i correspond to moment-generating functions with $\forall x_i \in \mathcal{X},\ u_i(x_i) = e^{\lambda(x_i - \mu_i - t)}$ and serve as natural upper bounds to the indicator function $\mathbf{1}_{\sum_{i=1}(x_i - \mu_i) \geqslant nt} \leqslant \prod_{i=1}^n e^{\lambda(x_i - \mu_i - t)}$.

Throughout this section, we assume the existence of a function $U \in \mathcal{U}$ as defined in (4.6) such that $\forall x \in \mathcal{X}^n, F(x) \leqslant U(x)$, from which we define two strategies. First, we introduce a separable approach, which generalizes classical probability proofs. Second, we optimize over the family of product-functions (4.6) and show how it formulates as a convex optimization problem. Depending on the moments and product-functions under consideration, we demonstrate that these approaches yield tractable upper approximations.

4.1.1 The separable approach.

Let $U \in \mathcal{U}$ be a product-function verifying $\forall x \in \mathcal{X}^n,\ F(x) \leqslant U(x)$. This section studies the properties of the optimization problem:

$$\rho_n^U = \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} U(x_1, \dots x_n) dp_1(x_1) \cdots dp_n(x_n).$$

This problem naturally provides an upper bound to the generalized problem of moments for independent random variables (4.4), specifically $\rho_n \leqslant \rho_n^U$. By definition, the function $U \in \mathcal{U}$ has a product structure $\forall x \in \mathcal{X}^n, U(x) = \prod_{i=1}^n u_i(x_i)$. This allows the separation of the integral over $\mathcal{X}^n$ into n integrals over $\mathcal{X}$, thus decoupling the optimization problem into n independent optimization problems over $p_i \in \mathcal{P}(\mathcal{X})$. In other words, the family $\mathcal{U}$ aligns with the structure imposed by independence:

$$\rho_n^U = \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \prod_{i=1}^n u_i(x_i) dp_1(x_1) \cdots dp_n(x_n) = \prod_{i=1}^n \sup_{p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}} u_i(x_i) dp_i(x_i).$$

For univariate distributions, strong duality holds [Karlin and Studden, 1966]. Then,

$$\begin{aligned} \rho_n^U &= \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \{\alpha_i + \beta_i^\top \mu_i\} \text{ such that } \forall\ x_i \in \mathcal{X}, u_i(x_i) \leqslant \alpha_i + \beta_i^\top g_i(x_i), \\ &= \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \sup_{x_i \in \mathcal{X}} \{u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)\}. \end{aligned} \tag{4.7}$$

Problem (4.7) is finite-dimensional and convex (as it is the pointwise supremum of affine functions [Boyd and Vandenberghe, 2004, Section 3.2.3]). In the case of independent and identically distributed random variables, the problem simplifies significantly: it reduces to a single optimization problem with $\rho_n^U = (\rho_1^{\exp})^n$.

However, computing $\sup_{x_i \in \mathcal{X}} \{u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)\}$ often remains numerically intractable. Under strong assumptions, such as the finiteness of the support or the convexity of the objective function on a compact set, it reduces to a finite number of constraints. Proposition 12 outlines a useful tractable setting that we will use in Section 4.2 for improving Hoeffding's inequality.

Proposition 12. *Let $\mathcal{X}$ be compact and $x_i \mapsto u_i(x_i) - \beta^\top (g_i(x_i) - \mu_i)$ be convex. Then, Problem (4.7) formulates as a tractable convex optimization problem with a finite number of constraints:*

$$\rho_n^U = \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} \{\alpha_i + \beta_i^\top \mu_i\} \text{ such that } \forall x \in \text{Extremal}(\mathcal{X}),\ \forall i,\ u_i(x_i) \leqslant \alpha_i + \beta_i^\top g_i(x_i).$$

Proof. Under the assumptions of Proposition 12, the maximization of the convex function $u_i(x_i) - \beta_i^\top (g_i(x_i) - \mu_i)$ over the compact set $\mathcal{X}$ is achieved at extremal points of $\mathcal{X}$ [Boyd and Vandenberghe, 2004, Section 3.2.3]. □

This technique faces two major challenges: first, constructing a valid upper bound $U \in \mathcal{U}$ can be difficult; second, even with a suitable upper bound, the resulting optimization Problem (4.1) may be numerically intractable. The next section focuses on more accurate approximations to the generalized problem of moments by optimizing over the family of upper bounds (instead of keeping one such upper bound fixed).

4.1.2 Optimizing over $\mathcal{U}$.

This section explores Problem (4.5) for the class of product-functions (4.6). Combined with the dual formulation (4.7), we define:

$$\rho_n^{\text{var}} = \inf_{u_i \geqslant 0} \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \prod_{i=1}^n (\alpha_i + \beta_i^\top \mu_i) \text{ such that } \forall i, \forall x_i \in \mathcal{X}, u_i(x_i) \leqslant \alpha_i + \beta_i^\top g_i(x_i), \\ \forall x \in \mathcal{X}^n, F(x_1, \dots, x_n) \leqslant \prod_{i=1}^n u_i(x_i). \tag{4.8}$$

Optimizing with respect to $u_1, \dots, u_n \geqslant 0$, it holds that:

$$\rho_n^{\text{var}} = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \prod_{i=1}^n (\alpha_i + \beta_i^\top \mu_i) \text{ such that } \forall x_i \in \mathcal{X}, \forall i, \alpha_i + \beta_i^\top g(x_i) \geqslant 0, \\ \forall x \in \mathcal{X}^n, F(x_1, \dots, x_n) \leqslant \prod_{i=1}^n (\alpha_i + \beta_i^\top g(x_i)). \tag{4.9}$$

As expected, Problem (4.9) shows no dependence on the u_i's. In Proposition 13, we formally compare ρ_n (4.4) to ρ_n^{var} (4.9) and to ρ_n^U (4.7) for any function $U \in \mathcal{U}$.

Proposition 13. *Let $U \in \mathcal{U}$, ρ_n^U be defined in (4.7), ρ_n in (4.4) and ρ_n^{var} in (4.9). Then it holds that*

$$\rho_n \leqslant \rho_n^{\text{var}} \leqslant \rho_n^U.$$

In addition, the equality $\rho_n^U = \rho_n^{\text{var}}$ holds for optimal values $(u_\star, \alpha_\star, \beta_\star)$ such that $\forall x_i \in \mathcal{X}, u_{i,\star}(x_i) = \alpha_{i,\star} + (\beta_{i,\star})^\top g_i(x_i)$.

Proposition 13 provides optimal product-functions $U \in \mathcal{U}$, as affine functions of the moments. This specific structure indicates that moment-generating functions, used in the Cramér-Chernoff method, are not optimal. However, computing $(\alpha_{i,\star}, \beta_{i,\star})$ is often difficult (that is, solving Problem (4.9) which is finite-dimensional but nonconvex). Proposition 14 ensures a convex reformulation of Problem (4.9).

Proposition 14. *Let ρ_n^{var} be defined in (4.8). Then, it holds that*

$$\log(\rho_n^{\text{var}}) = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} + \sup_{x \in \mathcal{X}^n} \left\{ \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\}. \tag{4.10}$$

In addition, if $x \mapsto \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))$ is convex and $\mathcal{X}$ is compact, then,

$$\sup_{x \in \mathcal{X}^n} \left\{ \log(F(x)) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\} = \sup_{x \in \text{Extremal}(\mathcal{X}^n)} \left\{ \log(F(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i)) \right\}.$$

Proof. First, let us consider the logarithm $\log(\rho_n^{\text{var}}) = \inf_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \log(\alpha_i + \beta_i^\top \mu_i) + \sup_{x \in \mathcal{X}^n} \{\log(F(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))\}$. Noticing that $\inf_{t \geqslant 0} \{ \Big(t\alpha_i + t\beta_i^\top \mu_i - 1\Big) - \log(t\alpha_i + t\beta_i g(x_i))\} = \log(\alpha_i + \beta_i^\top \mu_i) - \log(\alpha_i + \beta_i g(x_i))$, we conclude the reformulation in (4.10). The second assertion follows by maximization of a convex function over a compact set. □

Proposition 14 details a set of assumptions on $\mathcal{X}$, F and g, under which the constraints "$\log(F(x_1, \dots, x_n) \leqslant \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g_i(x_i))$" reduces to a finite number of points. For instance, these assumptions are satisfied for finite first-order moments $g_i(x_i) = x_i$ together with log-convex objectives F (such as exponentials or indicator functions $\mathbf{1}_S$, with compact $S \subset \mathcal{X}$,

see [Boyd and Vandenberghe, 2004, Section 3.5]). An alternative convexification proof is achieved using Donsker-Varadhan in Appendix 4.A, which also includes a formulation of the gap to the generalized problem of moments.

We have established two approaches for deriving upper bounds to the generalized problem of moments for independent random variables (4.4) using a family of product-functions (4.6). First, we introduced a separable approach (4.7) that formulates as a product of n convex optimization problems. However, constructing product functions may not be straightforward. Then, we formulate a variational Problem (4.9) emerging from (4.7) by optimizing with respect to product-functions. It turns out that Problem (4.9) benefits from a convex reformulation which does not require a priori upper bounds to F (but constructs such bounds in the process). Without further assumptions on the probability support $\mathcal{X}$, the objective F or moments g_i, both approaches are intractable. We will see next how they effectively apply in the context of Hoeffding's inequality.

4.2 Revisiting Hoeffding's inequality.

Hoeffding's inequality establishes a subgaussian tail for the sum of independent random variables taking their values in a bounded set with finite means. This section is devoted to refining Hoeffding's inequality, applying the separable and variational frameworks developed in Section 4.1. First, let us recall Hoeffding's inequality as stated in Theorem 26.

Theorem 26. *[Hoeffding, 1963] Let $X_1, ..., X_n$ be independent random variables such that X_i takes it value in $[a_1, b_i]$ almost surely. Then, for every $t \geqslant 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_n - \mathbb{E}[X_n]) \geqslant t\right) \leqslant \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) = \rho_n^{\text{Hoeffding}}. \tag{4.11}$$

Throughout this section, $X_1, \ldots, X_n$ are independent random variables with mean $\mu_1, \ldots, \mu_n$. Without loss of generality, we assume that they take their values in $\mathcal{X} = [0, 1]$ (that is, $b_i = 1$ and $a_i = 0$) and thereby, have finite mean $\mathbb{E}[X_i] = \mu_i \forall i$. Our goal is to approximate the probability $\mathbb{P}(\sum_{i=1}^n X_i \geqslant nt + \sum_{i=1}^n \mu_i)$ for $t \geqslant 0$. In our framework, it translates to functions where for all $x \in [0, 1]^n, F(x) = \mathbf{1}_{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i - nt}$ and $\forall i, \forall x_i \in [0, 1], g_i(x_i) = x_i$.

As a reference, we first consider one random variable, that benefits from an exact analytical bound (4.4). We compare it to the separable technique (4.7) on the moment-generating function. Then, we extend the analysis to n random variables, comparing bounds in the separable and variational approaches to Hoeffding's inequality. More precisely, we examine cases where random variables are i.i.d., and where random variables are divided into two blocks with different means $\mu_1 = \mu_2 = \cdots = \mu_m$ and $\mu_{m+1} = \cdots = \mu_n$. For the case $\mu_1 = \mu_2 = \cdots = \mu_n$, our results asymptotically correspond to large deviations. Finally, we propose a methodology to reconstruct the distribution in the worst-case scenario.

4.2.1 One random variable: comparison of the exact and exponential bounds

Computing optimal bounds for univariate random variables has been extensively studied in past years and is encompassed in the multivariate analyses proposed by Isii [1962], Bertsimas et al. [2000], Vandenberghe et al. [2007]. Let us compute the exact closed-form solution (4.4) and the separable scenario for moment-generating functions (4.7).

Exact optimization problem. Let X_1 be a random variable in $\mathcal{X} = [0,1]$, with $\mathbb{E}[X_1] = \mu$. Recall the exact optimization problem (4.4) for every $t \geqslant 0$,

$$\begin{aligned}\rho_1(t) &= \sup_{p_1 \in \mathcal{P}([0,1])} \int_0^1 \mathbf{1}_{x_1 \geqslant \mu + t} dp_1(x_1) \text{ such that } \int_0^1 x_1 dp_1(x_1) = \mu, \\ &= \inf_{\alpha,\beta} \alpha + \beta\mu \text{ such that } \forall\, x_1 \in [0,1], \mathbf{1}_{x_1 \geqslant \mu+t} \leqslant \alpha + \beta x_1.\end{aligned} \tag{4.12}$$

Problem 4.12 defines a function $\rho_1(\cdot)$ as a solution to a linear program for every $t \geqslant 0$. As stated by Bertsimas and Popescu [2005, Theorem 2.2], strong duality holds for $\mu \in]0,1[$. It turns out that Problem (4.12) is a particular case of both the generalized moment problem (4.2) for univariate distributions and of the generalized problem of moments for independent random variables (4.1). It admits a closed-form solution, detailed in Proposition 15.

Proposition 15. *Let $t \geqslant 0$ such that $\mu + t < 1$. Then, for all $0 \leqslant t \leqslant 1 - (\mu + t)$,*

$$\rho_1(t) = \frac{\mu}{\mu + t}. \tag{4.13}$$

Proof. Function $x_1 \mapsto -(\alpha+\beta x_1)$ (resp. $x_1 \mapsto 1-(\alpha+\beta x_1)$) is convex on $[0, \mu+t]$ (resp. $[\mu+t, 1]$), achieving therefore its extrema in 0 or $\mu + t$ (resp. $\mu + t$, 1). Thus the constraint '$\forall x_1 \in [0,1], \mathbf{1}_{x_1 \geqslant \mu+t} \leqslant \alpha + \beta x_1$' in (4.12) can be reduced to two constraints : $\alpha \geqslant 0$ and $\beta(\mu + t) \geqslant 1$. It follows that $\alpha = 0$ and $\beta = \frac{1}{\mu+t}$. □

Separable approach. Now, let us compute the bound defined in (4.7) considering moment-generating functions $u_\lambda(x) = e^{\lambda(x-(\mu+t))}$. By construction, $\forall x \in \mathbb{R}, \mathbf{1}_{x \geqslant \mu+t} \leqslant e^{\lambda(x-(\mu+t))}$. We define the family of upper bounds $\rho_1^{\exp}$. For every $\lambda \in \mathbb{R}$ and for every $t \geqslant 0$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha,\beta} \alpha + \beta\mu, \text{ such that } \forall x_1 \in [0,1], e^{\lambda(x_1-(\mu+t))} \leqslant \alpha + \beta x_1. \tag{4.14}$$

Problem (4.14) is a convex optimization problem, that is well-defined for every $t \geqslant 0$ and $\lambda \in \mathbb{R}$. For every $t \geqslant 0$, it admits an optimal moment-generating function analytically given in Proposition 16.

Proposition 16. *Let $t \geqslant 0$, and $\rho_1^{\exp}(\lambda, t)$ be defined in (4.14). Then, for every $t \geqslant 0$,*

$$\rho_{1,\star}^{\exp}(t) = \left(\frac{\mu}{\nu}\right)^{\nu} \left(\frac{1-\mu}{1-\nu}\right)^{1-\nu}, \tag{4.15}$$

where $\nu = \mu + t$, and $\rho_{1,\star}^{\exp}$ is the optimal value when optimizing (4.14) with respect to λ.

Proof. . See Appendix 4.B.1.1. □

Proposition 16 yields exactly the Chernoff-Bound for a Bernoulli random variable on the support $\{0,1\}$ with mean μ [Boucheron et al., 2013, Section 2.2]. This result was originally stated in the seminal work of Hoeffding [1963], but our approach ensures tightness for this family of moment-generating functions. It can also be directly derived from Kullback's inequality, as detailed in Appendix 4.B.1.2.

In Figure 4.1, we observe that the exact bound ρ_1 (4.12) significantly improves upon Hoeffding's bound, whereas the bound $\rho_1^{\exp}$ (4.14) in the separable approach only shows improvement for large values of t. Both approaches benefit from a dependence in the first-order moment μ, that does not appear in Hoeffding's inequality (4.11). The next section extends beyond the univariate case.

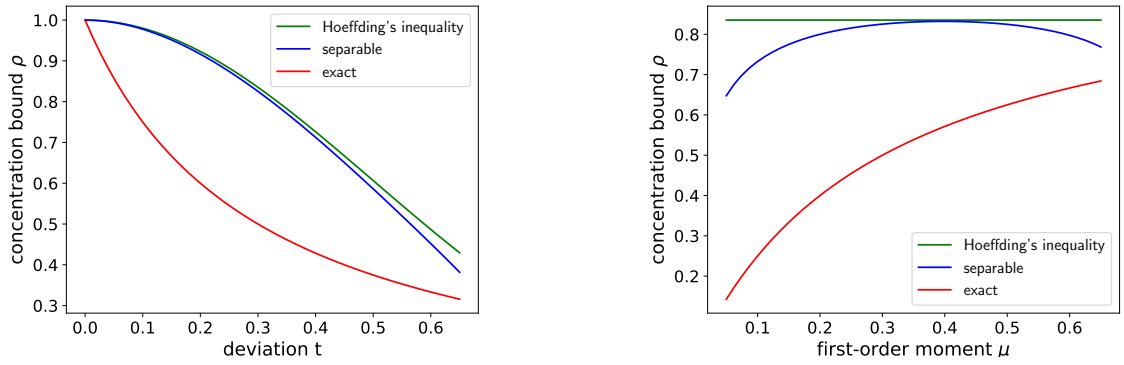


Figure 4.1: Comparison of ρ_1 (4.13) and $\rho_1^{\exp}$ (4.15) to Hoeffding's inequality $\rho_1^{\text{Hoeffding}}$ (4.11). On the left, bounds are plot as a function of t with $\mu = 0.3$, and on the right as a function of μ with $t = 0.3$.

4.2.2 Generalization to n independent random variables.

We consider n independent random variables with $n > 1$. We first examine the case of i.i.d. random variables, where $\mu_1 = \cdots = \mu_n$. The bound obtained in the variational approach (4.9) asymptotically matches the large deviations, and cannot therefore by improved. Next, we consider the case of independent random variables with different means. Specifically, we formulate a tractable upper bound using the separable treatment and discuss the computational limits of the variational approach for large n due to an exponential number of constraints. In the special case of two blocks of variables with different means, the number of constraints involved is $O(n)$.

4.2.2.1 Independent and identically distributed random variables (equal means).

This section focuses on i.i.d. random variables, that is with $\mu_1 = \mu_2 = \cdots = \mu_n$. Thanks to symmetry properties, both the separable and variational approaches yield tractable solutions for any number of variables n.

Separable approach. Let us consider the moment generating function of $X_1 + \cdots + X_n$ for $\lambda \in \mathbb{R}$: $\forall x \in [0,1]^n, u_\lambda(x) = e^{\lambda(\sum_{i=1}^n \{x_i - \mu - t\})} = \prod_{i=1}^n e^{\lambda(x_i-\mu-t)}$. Note that this is a product-function and that for all $x \in [0,1]^n, F(x) = \mathbf{1}_{\sum_{i=1}^n x_i \geqslant n(t+\mu)} \leqslant u_\lambda(x)$. We thus define the separable problem as in (4.7) for $\lambda \in \mathbb{R}$ and $t \geqslant 0$:

$$\rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}} (\alpha_i + \beta_i \mu) \text{ such that } \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu-t)} \leqslant \alpha_i + \beta_i \mu. \tag{4.16}$$

Optimizing with respect to λ, Proposition 17 provides a closed-form as a function of t.

Proposition 17. *Let $\mu_1 = \cdots = \mu_n$, and let us consider the family of moment-generating functions u_λ as defined above. Then, it holds that for all $t \geqslant 0$:*

$$\rho_{n,\star}^{\exp}(t) = \left(\frac{\mu}{\nu}\right)^{n\nu} \left(\frac{1-\mu}{1-\nu}\right)^{n(1-\nu)} = (\rho_{1,\star}^{\exp}(t))^n. \tag{4.17}$$

Proof. Since $\mu_1 = \cdots = \mu_n$, Problem (4.17) benefits from a symmetry property and can be simplified into $\forall t \geqslant 0$, $\forall \lambda \in \mathbb{R}$, $\rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha,\beta \in \mathbb{R}} (\alpha + \beta\mu)$, s.t. $\forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu-t)} \leqslant \alpha + \beta\mu = (\rho_1^{\exp}(\lambda))^n$. We then optimize over λ as in Proposition 16. □

The symmetry properties induced by the i.i.d. assumption allow simplifying (4.16) into a single univariate convex optimization problem, which can be solved efficiently. Again, this is exactly the Chernoff bound for n i.i.d. Bernoulli variables taking their values in $\{0,1\}$ with mean n. We conclude that the separable approach improves Hoeffding's inequality for large deviations t (as for a unique univariate variable). Let us now explore how the variational approach might offer further improvements.

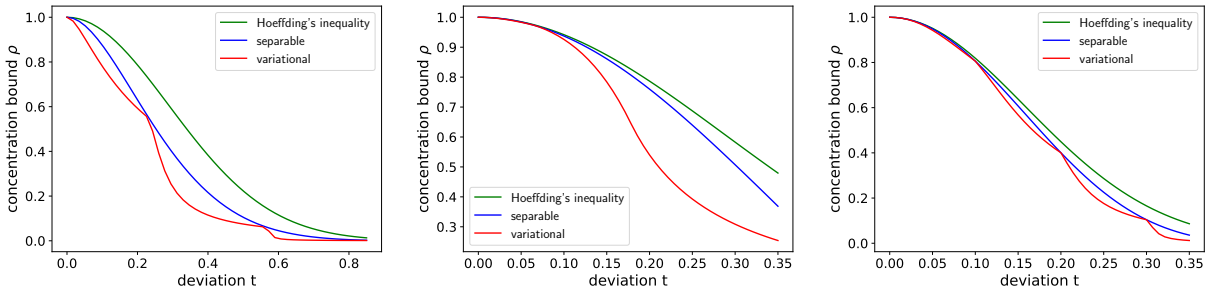
Variational approach. In the context of Hoeffding's inequality and given symmetry properties induced by $\mu_1 = \cdots = \mu_n = \mu$, the bound ρ_n^{var} (4.9) takes the form for all $t \geqslant 0$:

$$\rho_n^{\text{var}}(t) = \inf_{\alpha\in\mathbb{R},\beta\in\mathbb{R}} \prod_{i=1}^n (\alpha + \beta^\top \mu) \;\text{ such that } \forall x \in [0,1]^n,\; \mathbf{1}_{x_1+\cdots+x_n \geqslant n(\mu+t)} \leqslant \prod_{i=1}^n (\alpha + \beta x_i),$$
$$\forall x_i \in [0,1],\; \alpha + \beta x_i \geqslant 0.$$

This problem has a convex reformulation with a finite number of constraints, as proven in Proposition 14,

$$\log \rho_n^{\text{exp}}(t) = \inf_{\alpha,\beta,t\geqslant 0} n(\alpha + \beta\mu - 1) \text{ such that } \; -\sum_{i=1}^n \log(\alpha + \beta x_i) \leqslant 0,\; x \in \text{extremal}(\bar{\mathcal{X}}_n), \tag{4.18}$$

where $\bar{\mathcal{X}}_n = \{(x_1,\ldots,x_n) \in [0,1]^n, x_1 + \cdots + x_n \geqslant n(\mu + t)\}$ and where the constraint "$\forall x \in [0,1], \alpha + \beta x \geqslant 0$" is implied by the logarithm. The set $\bar{\mathcal{X}}_n \subset [0,1]^n$ is compact and symmetric in $(x_1,\ldots,x_n)$. At first sight, the set extremal$(\mathcal{X})_n$ appears to grow exponentially with n. However, thanks to symmetry properties and to the structure of the constraints, it can be reduced to $O(n)$ constraints (see Appendix 4.B.2, via computation of extremal points). Therefore, Problem (4.18) can be efficiently addressed using standard solvers for convex optimization. It is even possible to derive closed-form solutions for $(\alpha_\star, \beta_\star)$, by enumerating all extremal points given $\mu + t$ and solving the KKT condition for Problem (4.18). We provide an example for $n = 2$ in Appendix 4.B.3.



(a) $\mu = 0.1$ and $n = 3$. (b) $\mu = 0.6$ and $n = 3$. (c) $\mu = 0.6$ and $n = 10$.

Figure 4.2: Comparison of bounds derived in the separable approach ρ_n^{exp} (4.16), in the variational approach ρ_n^{var} (4.18) (computed using the solver MOSEK ApS [2022] in cvxpy) to the known Hoeffding's bound $\rho_n^{\text{Hoeffding}}$ (4.11), as a function of the deviation t.

Figure 4.2 illustrates the comparison between the separable and variational approaches to the Hoeffding's bound for different numbers of i.i.d. random variables. It appears that $\rho_{n,\star}^{\text{exp}}$ (4.16) closely tracks Hoeffding's bound. In contrast, ρ_n^{var} (4.18) provides numerical improvements over Hoeffding's inequality by up to a factor of 2 when a small number of random random variables are in play. Furthermore, as the number of variables increases, the variational approach asymptotically matches the separable treatment, as expected from the large deviations theory.

In probability theory, the study of the asymptotic behavior of tails of random variables is known as the large deviations theory, introduced by Varadhan [1988]. In particular, the large deviation principle provides a guarantee on rare events, as outlined in Theorem 27 for the sum of i.i.d. random variables.

Theorem 27 (Cramér [1938]: Large Deviations)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with finite moment-generating functions, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for all $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log(\mathbb{P}(\bar{X}_n \geqslant x)) = -\Gamma^\star(x),$$

where $\Gamma^\star(x) = \sup_{t\geqslant 0}(tx - \Gamma(t))$ and $\Gamma(t) = \log(\mathbb{E}[\exp(tX_1)])$.

The moment-generating function of a univariate random variable formulates as an optimization problem as in (4.14). Corollary 9 provides the large deviations asymptotic for i.i.d. random variables.

Corollary 9. *Let $X_1, \dots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_1] = \mu$, and taking their value in $[0,1]$ almost surely, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. It holds that, for all $t \geqslant 0$,*

$$\lim_{n\mapsto\infty} \frac{1}{n} \log(\mathbb{P}(\bar{X}_n \geqslant \mu + t)) = -(\mu+t)\log\left(\frac{\mu}{\mu+t}\right) - (1-(\mu+t))\log\left(\frac{1-\mu}{1-(\mu+t)}\right).$$

Proof. From the separable approach, $\Gamma(t) = \sup_{p\in\mathcal{P}(\mathcal{X})} \int_0^1 e^{tx} dp(x)$ such that $\int_0^1 x dp(x) = \mu$. Strong duality holds and $\Gamma(t) = \inf_{\alpha,\beta\in\mathbb{R}} \alpha + \beta\mu$, such that $\forall x \in [0,1]$, $e^{tx} \leqslant \alpha + \beta x$. Thus, $\Gamma(t) = 1 + \mu(e^t - 1)$. The desired statement is obtained by computing the Fenchel conjugate. □

Corollary 9 shows that the probability that $\bar{X}_n$ deviates from the mean μ converges exactly to the bound in the separable approach (4.17): for all $t \geqslant 0$, $\mathbb{P}(X_1 + \cdots + X_n \geqslant n(\mu + t)) \approx (\rho_1^{\exp})^n$. Numerical results from Figure 4.2 therefore align with the large deviations for relatively large n.

As a conclusion, when random variables are i.i.d., optimization problems in the separable and variational approach benefit from tractable formulations. The bound in the variational treatment (4.18) shows significant improvements over Hoeffding's inequality when a small number n of random variables are in play but suffers an increasing number of constraints. As n increases, the separable approach (4.16) provides a close estimate of the generalized problem of moments at a lower computational cost.

4.2.2.2 Two blocks of random variables with different means.

Hoeffding's inequality, as presented in Theorem 26, applies generally to independent random variables without specific assumptions on their means. However, formulating the separable and variational approach in a generic setting can be computationally challenging for a large number of variables n. To simplify this, we focus on the scenario where the random variables are divided into two blocks with different means.

Separable approach. Consider the optimization problem (4.7) with different means in the context of Hoeffding's inequality. After solving each subproblem in (α_i, β_i) as in Proposition 16, the bound takes the form, for any $t \geqslant 0$:

$$\rho_{n,\star}^{\exp}(t) = \inf_{\lambda\in\mathbb{R}} e^{-n\lambda(\bar{\mu}_n + t)} \prod_{i=1}^n \left(1 + \mu_i(e^\lambda - 1)\right). \tag{4.19}$$

Optimizing over $\lambda \in \mathbb{R}$ for different means μ_i cannot be achieved in closed-form as in the case of i.i.d. variables. Note that $\log(\rho_{n,\star}^{\exp})$ could be computed as the minimum of a convex objective in λ. We rather explicit an analytical upper bound in Proposition 18.

Proposition 18. *Let $\mu_1, \dots, \mu_n$ be in $]0,1[$ and $\mu_i \neq \frac{1}{2}$. Then it holds for $t \geqslant 0$ that:*

$$\rho_{n,\star}^{\exp}(t) \leqslant \exp\left(\frac{nt^2/2}{\frac{1}{n}\sum_{i=1}^n \log(\frac{\mu_i}{1-\mu_i})}\right).$$

In addition, $\rho_{n,\star}^{\exp}(t) \leqslant \rho_n^{\text{Hoeffding}}$.

Proof. Let us consider $f_i(\lambda) = \log(1+\mu_i(e^\lambda - 1))$. A quadratic upper bound for f was derived by [Jaakkola and Jordan, 2000, Section 2.2], such that $\forall \lambda \in \mathbb{R}, \log(1+\mu_i(e^\lambda-1)) \leqslant \lambda\mu_i + \frac{\lambda^2}{4}\frac{2\mu_i - 1}{\log(\frac{\mu_i}{1-\mu_i})}$. In addition, $\lambda\mu_i + \frac{\lambda^2}{4}\frac{2\mu_i-1}{\log(\frac{\mu_i}{1-\mu_i})} \leqslant \lambda\mu_i + \frac{\lambda^2}{8}$, leading to the final assertion. □

Remark 10. *In the proof for Proposition 18, considering the naive upper function $\log(1 + \mu_i(e^\lambda - 1)) \leqslant \lambda\mu_i + \frac{\lambda^2}{8}$ would have led to Hoeffding's inequality.*

Given random variables with different means, Proposition 18 shows a control of $\rho_{n,\star}^{\exp}(t)$ by an upper bound depending on $(\mu_i)_{i=1,\dots,n}$ and improving Hoeffding's inequality. Therefore, it appears to be suited to different means as well as to two blocks of random variables.

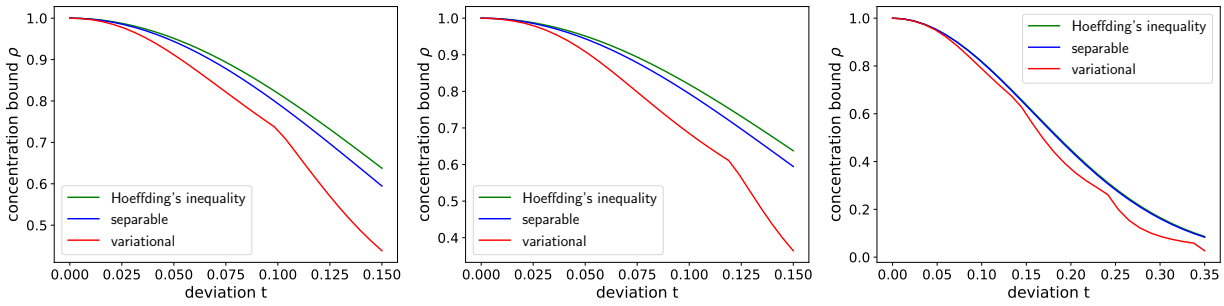
Variational approach. Recall the optimization problem defining the variational approach for random variables in $[0,1]$ having different means $\mu_1, \dots, \mu_n$. For $t \geqslant 0$, we define:

$$\log(\rho_n^{\text{var}}(t)) = \inf_{\alpha\in\mathbb{R}^n, \beta\in\mathbb{R}^n} \sum_{i=1}^n \{\alpha_i + \beta_i\mu_i - 1\}, \text{ such that } -\sum_{i=1}^n \log(\alpha_i + \beta_i x_i) \leqslant 0, \forall x \in \text{ extremal}(\bar{\mathcal{X}}_n), \tag{4.20}$$

where $\bar{\mathcal{X}}_n = \{(x_1, \dots, x_n) \in [0,1]^n, \sum_{i=1}^n x_i \geqslant nt + \sum_{i=1}^n \mu_i\}$. Without further assumptions on the μ_i's, this optimization problem may have up to $O(n!)$ constraints. To simplify the computations, we consider a first group of variables $X_1, \dots X_m$ with mean μ_1, and a second group $X_{m+1}, \dots, X_n$ with mean μ_2, with $1 \leqslant m \leqslant n$. Then, the number of constraints under consideration can be reduced to $O(n)$, as shown in Lemma 17.

Lemma 17. *Let $\mu_1 = \dots = \mu_m$ and $\mu_{m+1} = \dots = \mu_n$ in (4.20). Then, the number of constraints in Problem (4.20) reduces to $O(n^2)$.*

Proof. See Appendix 4.B.4. □



(a) $n = 10$, $m = 5$, $\mu_1 = 0.2$, $\mu_2 = 0.8$ (b) $n = 10$, $m = 2$, $\mu_1 = 0.2$, $\mu_2 = 0.8$ (c) $n = 10$, $m = 2$, $\mu_1 = 0.4$ and $\mu_2 = 0.6$

Figure 4.3: Comparison of the variational approach (4.20) in the context of random variables divided into two blocks of size m and $n-m$, with mean μ_1 and μ_2, to the separable approach (4.19) and to Hoeffding's inequality (4.11).

In Figure 4.3, we observe that the variational approach (4.20) improves largely upon the separable scenario (4.19), as soon as the means of the two blocks differs significantly.

Given n random variables divided into two subgroups with different means, we have formulated tractable formulations in the variational (4.20) and separable (4.19) approaches. On the one hand, we provide an upper bound to the separable approach in Proposition 18 using a quadratic upper bound, which still offers improvements over Hoeffding's inequality. On the other hand, the variational approach can be expressed as a manageable convex optimization problem with $O(n)$ constraints, as stated in Lemma 17. Numerical comparisons show that both approaches behave similarly for a large number of variables n, but tend to differ for small n. In the following, we explore the reconstruction of extremal distributions involved at the optimum.

4.2.3 Reconstruction extremal distributions.

A bound is said to be tight if a distribution satisfies the bound with equality, as defined by Bertsimas and Popescu [2005, Section 2.]. Such a distribution is referred to as an *extremal distribution*. We first restrict our attention to one random variable, for which the exact optimization formulation (4.12) and the separable approach (4.14) admit analytical solutions. Building on these results, we propose a strategy to construct an extremal distribution for n random variables in the variational (4.20) and separable approaches (4.19).

4.2.3.1 Dirac distributions for one univariate random variable.

When it comes to one univariate random variables, we derived in Section 4.2.1 closed form of the exact (4.12) and separable (4.14) approaches. In the following, we construct their corresponding extremal distributions.

Exact optimization problem. Recall the exact optimization problem (4.12) for one variable, for all $t \geqslant 0$:

$$\rho_1(t) = \sup_{p_1 \in \mathcal{P}([0,1])} \int_0^1 \mathbf{1}_{x_1 \geqslant \mu + t} dp_1(x_1) \text{ such that } \int_0^1 x_1 dp_1(x_1) = \mu. \tag{4.21}$$

Its exact value is provided in Proposition 15: for all $t \geqslant 0, \rho_1(t) = \frac{\mu}{\mu+t}$. Dualizing twice the optimization problem (4.21), we propose in Proposition 19 a strategy for reconstructing an extremal distribution.

Proposition 19. *Let $t \geqslant 0$. The following distribution is an optimal solution to (4.12):*

$$\forall x \in [0,1], p(x) = \frac{t}{\mu + t}\delta_{x=0} + \frac{\mu}{\mu + t}\delta_{x=\mu+t}. \tag{4.22}$$

Proof. . The result is obtained by computing directly $\int_0^1 \mathbf{1}_{x \geqslant \mu+t} dp(x) = \frac{\mu}{\mu+t}$. We rather propose a constructive approach. First, recall the dual of Problem (4.21): for all $t \geqslant 0, \rho_1(t) = \inf_{\alpha,\beta \in \mathbb{R}} \alpha + \beta\mu$, such that $\forall x \in [0,1], \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$. By convexity of $x \in [0,1] \mapsto \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$, the problem reduces for all $t \geqslant 0$ to $\rho_1(t) = \inf_{\alpha,\beta} \alpha + \beta\mu$, such that $\alpha \geqslant 0, 1 \geqslant \alpha + \beta(\mu + t)$. We compute again the dual, which reformulates as a problem in the probability space by strong duality: for all $t \geqslant 0$ $\rho_1(t) = \sup_{\lambda_1,\lambda_2 \geqslant 0} \lambda_2$, such that $1 = \lambda_1 + \lambda_2, \mu = \lambda_2(\mu + t)$. The constraint "$1 = \lambda_1 + \lambda_2$" corresponds to the initial constraint $\int_0^1 dp(x) = 1$, and the second one to the first-order moment condition $\int_0^1 x dp(x) = \mu$. At optimality, $(\lambda_{1,\star}, \lambda_{2,\star}) = (\frac{t}{\mu+t}, \frac{\mu}{\mu+t})$ and $(\alpha_\star, \beta_\star) = (0, \frac{1}{\mu+t})$ meaning that the points $x = 0$ and $x = \mu + t$ are active. We conclude that $\lambda_{1,\star}$ and $\lambda_{2,\star}$ are respectively associated to points $x = 0$ and $x = \mu + t$. □

Remark 11. *Other distributions achieve the optimal bound, such as $p(x) = (1 - \mu)\delta_0 + \mu\delta_1$.*

Proposition 19 offers a strategy for reconstructing an extremal distribution, mostly by reducing the constraints involved at the optimum to some active points and by dualizing twice.

Separable treatment. We derive the same technique described in the proof for Proposition 19 to the separable approach. Recall the separable approach applied to moment-generating functions (4.14) is defined for $\lambda \in \mathbb{R}$ and $t \in \geqslant 0$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha,\beta} \alpha + \beta\mu \;\text{ such that }\; \forall\, x_1 \in [0,1], e^{\lambda(x_1-(\mu+t))} \leqslant \alpha + \beta x_1.$$

In the proof for Proposition 16, we showed that $\rho_1^{\exp}(\lambda, t) = e^{-\lambda(\mu+t)}\Big((1+\mu(e^\lambda - 1)\Big)$. We compute an example of an extremal distribution in Proposition 20, that is independent of t and λ.

Proposition 20. *Let $t, \lambda \in \mathbb{R}$. The following distribution is an optimal solution to (4.14):*

$$p(x) = (1-\mu)\delta_{x=0} + \mu\delta_{x=1}.$$

Proof. Following the same approach as in Proposition 19 for determining the extremal distribution of the exact optimization problem. We prove in Appendix 4.B.1.1 that $(\alpha_\star, \beta_\star) = (e^{-\lambda(\mu+t)}, e^{-\lambda(\mu+t)}(e^\lambda - 1))$. After redualizing, it leads to the problem $\sup_{\lambda_1,\lambda_2 \geqslant 0} \lambda_1 e^{-\lambda(\mu+t)} + \lambda_2 e^{\lambda(1-\mu-t)}$, such that $1 = \lambda_1 + \lambda_2, \mu = \lambda_2$. We conclude that $\lambda_{1,\star} = 1-\mu$, $\lambda_{2,\star} = \mu$ are associated to the points $x = 0$ and $x = 1$. □

When studying concentration inequalities applied to a one univariate random variable, we derived examples of extremal distribution. This strategy can be decomposed into two steps. First, the convexity properties of the constraints $\forall x \in [0,1], \mathbf{1}_{x \geqslant \mu+t} \leqslant \alpha + \beta x$ and $\forall x \in [0,1], e^{\lambda(x-\mu-t)} \leqslant \alpha + \beta x$ allow determining active constraints and thus identifying the Dirac delta functions involved at optimality. Second, leveraging strong duality, we dualize the dual, leading to an optimization problem with respect to a simplified space of distributions. We then conclude by identification. In what follows, we show that this technique extends well to n univariate random variables.

4.2.3.2 Generalization to n random variables.

We now turn to the problem of deriving extremal distribution for n independent random variables. The strategy developed for one random variable extends well to multiple random variables in the separable approach. In the variational approach however, this process often involves computing analytically the solution in the dual before obtaining the extremal distribution.

Separable approach. The separable approach (4.7) benefits from a decoupling into n independent optimization problems on one random variables. In the context of Hoeffding's inequality, recall its dual formulation (4.19) below, for $t \geqslant 0$:

$$\rho_n^{\exp}(\lambda, t) = \prod_{i=1}^n \inf_{\alpha_i,\beta_i} (\alpha_i + \beta_i\mu_i) \;\text{ such that }\; \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu_i-t)} \leqslant \alpha_i + \beta_i\mu_i.$$

Proposition 20 applies on each subproblem i, leading to Corollary 10.

Corollary 10. *The following distribution is an optimal solution to (4.19) $p(x) = \prod_{i=1}^n p_i(x_i)$, with $\forall x_i \in [0,1], p_i(x_i) = (1-\mu_i)\delta_{x_i=0} + \mu_i\delta_{x_i=1}$.*

Variational approach. At first glance, the variational approach reintroduces coupling between variables in the optimization problem. Let us recall its dual formulation (4.20) in the context of Hoeffding's inequality, for $t \geqslant 0$,

$$\rho_n^{\text{var}}(t) = \inf_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^n} \prod_{i=1}^n(\alpha_i+\beta_i\mu_i) \text{ such that } \forall\, x\in[0,1]^n, \mathbf{1}_{\sum_{i=1}^n x_i \geqslant \sum_{i=1}^n \mu_i + nt} \leqslant \prod_{i=1}^n(\alpha_i+\beta_i x_i), \\ \forall\, i, \forall\, x_i\in[0,1], 0\leqslant \alpha_i+\beta_i x_i. \tag{4.23}$$

The Lagrangian dual of Problem (4.23) cannot be formulated in closed form due to the product form in the constraints. Let us revisit the original optimization problem from which this problem is derived:

$$\rho_n^{\text{var}} = \inf_{\forall i,u_i\geqslant 0} \inf_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^n} \prod_{i=1}^n(\alpha_i+\beta_i\mu_i), \quad \text{such that } \forall x\in[0,1]^n, \mathbf{1}_{\sum_{i=1}^n x_i \geqslant \sum_{i=1}^n \mu_i + nt} \leqslant \prod_{i=1}^n u_i(x_i), \\ \forall i, \forall x_i\in[0,1], u_i(x_i)\leqslant \alpha_i+\beta_i x_i. \tag{4.24}$$

Using this formulation, Proposition 21 computes explicitly an extremal distribution.

Proposition 21. *Let $(u_\star,\alpha_\star,\beta_\star)$ be a solution to (4.24). Then, $\forall i, \forall x_i\in[0,1], u_{i,\star}(x_i) = \alpha_{i,\star}+\beta_{i,\star}x_i$. In addition, an extremal distribution is given by:*

$$p(x) = \prod_{i=1}^n \left((1-\mu_i)\delta_{x_i=0} + \mu_i\delta_{x_i=1}\right).$$

Proof. Let $(u_\star,\alpha_\star,\beta_\star)$ be optimal solutions in (4.24). Then, Proposition 13 provides the form of the optimal product-function upper bounding F, that we recall: $\forall i, \forall x_i\in[0,1], u_{i,\star}(x_i) = \alpha_{i,\star}+\beta_{i,\star}x_i$. Thus, considering this specific product-function, we have

$$\begin{aligned}\rho_{relax}^n &= \prod_{i=1}^n \sup_{p_i\in\mathcal{P}(\mathcal{X})} \int_0^1 u(x_i)dp_i(x_i) \text{ such that } \int_0^1 x_i dp_i(x_i) = \mu_i, \\ &= \prod_{i=1}^n \inf_{\lambda_i,\nu_i} (\lambda_i\mu_i+\nu_i) \text{ such that } \alpha_{i,\star}-\nu_i+(\beta_{i,\star}-\lambda_i)x_i\leqslant 0, \forall x_i\in[0,1],\end{aligned}$$

where the constraint "$\alpha_{i,\star}-\nu_i+(\beta_{i,\star}-\lambda_i)x_i\leqslant 0, \forall x_i\in[0,1]$" reduces to $\alpha_{i,\star}-\nu_i\leqslant 0$ for$x_i=0$ and $\alpha_{i,\star}-\nu_i+\beta_{i,\star}-\lambda_i\leqslant 0$ for$x_i=1$. We conclude by identification that $p(x) = (1-\mu)\delta_0+\mu\delta_1$. □

The extremal distribution derived in Proposition 21 is exactly equal to the extremal distribution in the separable treatment approach moment-generating functions in Proposition 10. Again, it is independent of the deviation t. In both cases, it turns out that the dependence in t is only supported in the upper function U, either in the moment generating function or in the linear function parametrized by $(\alpha_\star,\beta_\star)$, as detailed in Appendix 4.B.3 for $n=2$.

In this section, we have thus refined Hoeffding's inequality through two different approaches. When the random variables are i.i.d., the separable approach applied to moment-generating functions aligns with the large deviation principle and slightly improves the traditional Hoeffding's inequality. In contrast, the variational approach yields significantly smaller bounds for a small number of variables but requires $O(n)$ constraints. The case of distinct means $\mu_1,\mu_2,\ldots,\mu_n$ is more advanced and we only explicitly attacked the scenario of two blocks with distinct means. Finally, we proposed a strategy for reconstructing an extremal distribution. As a natural extension, these strategies could potentially be applied to Bennett or Bernstein's inequalities, which assume first and second-order conditions. However, extending to higher-order assumptions reveals challenges where both the separable and variational approaches fail to provide computable solutions. In what follows, we propose a new family of upper bounds adapted to such scenarios.

4.3 A polynomial approach based on sum-of-square decomposition.

The separable and variational approaches are constructed from the family of product-functions (4.6) that upper bound F on $\mathcal{X}$. They turn out to be effective for finite first-order moments, but remain computationally out of reach when assuming fixed higher-order moments. Even a simple second-order assumption in Hoeffding's inequality (or an assumption on the variance) results in challenging constraints in the separable approach:

$$\rho_n^{\exp}(\lambda) = \prod_{i=1}^n \inf_{\alpha_i,\beta_i} (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)}) \text{ s.t. } \forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu_i-t)} \leqslant \alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2,$$

where $\mu_i^{(1)}$ (resp. $\mu_i^{(2)}$) represents the first (resp.second) order moment condition. The constraint "$\forall x_i \in [0,1]^n, e^{\lambda(x_i-\mu_i-t)} - (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2)$ admits no closed-form solution. This issue also prevents from computing the large deviations for i.i.d. random variables, which involves computing $\Gamma(t) = \log(\mathbb{E}[\exp(tX_1)])$ as presented in Theorem 27. Similarly, the variational approach involves polynomial constraints with a product structure, making the optimization problem more complex to solve:

$$\rho_n^{\text{var}} = \inf_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^{n\times m}} \prod_{i=1}^n (\alpha_i + \beta_i^{(1)}\mu^{(1)} + \beta_i^{(2)}\mu^{(2)}) \text{ such that } \forall x_i \in \mathcal{X}, \forall i, \alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2 \geqslant 0,$$
$$\forall x \in \mathcal{X}^n, F(x_1,\ldots,x_n) \leqslant \prod_{i=1}^n (\alpha_i + \beta_i^{(1)} x_i + \beta_i^{(2)} x_i^2).$$

In this section, we explore polynomial families of upper bounds adapted to such scenarios. First, we propose to analyze Problem (4.5) using the family of linear upper bounds, for which we derive closed-form upper bounds. This *linear approach* offers already an improvement to Hoeffding's inequality in comparison with the variational approach for small values of t. We then extend this approach to a polynomial family of upper bounds, whose degree equals the number of variables. This results in an optimization problem with an infinite number of polynomial constraints. This so-called *polynomial approach* is closely related to the work of Bertsimas and Popescu [2005], who introduced a series of SDPs to approximate the generalized problem of moments for multivariate random variables. This approach numerically improves Bernstein's and partially Bennett's inequality. Finally, we introduce a *feature-based approach* generalizing the polynomial approach to a broader family of upper bounds. This allows in particular analyzing Hoeffding's inequality using second-order polynomials.

4.3.1 A simple linear upper bound for Hoeffding's concentration.

Before studying polynomial upper bounds, we start by exploring the simpler family of linear upper bounds and applying it in the context of Hoeffding's inequality. This scenario outlines the key concepts that inspire the polynomial approach and results in an optimization problem that can be solved in closed-form.

Let $X_1,\ldots,X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and taking their values in $[0,1]$ and let us introduce the family of linear functions $\mathcal{L} = \{a^\top x + b, a \in \mathbb{R}^n, b \in \mathbb{R}\}$.

Remark 12. *For $n \geqslant 2$, notice that linear functions cannot be formulated as product-functions. Indeed, for two i.i.d. random variables, product-functions takes the form $U(x) = (\alpha+\beta x_1)(\alpha+\beta x_2) = \alpha^2 + \alpha\beta(x_1+x_2) + \beta^2 x_1 x_2$.*

The associated optimization problem in the linear approach takes the form, for $t \geqslant 0$:

$$\rho_n^{\text{lin}}(t) = \inf_{a\in\mathbb{R}^n, b\in\mathbb{R}} \sup_{\forall i, p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} \left(\sum_{i=1}^n a_i x_i + b\right) dp_1(x_1)\cdots dp_n(x_n), \\ \text{such that } \forall x \in \mathcal{X}^n, F(x) \leqslant \sum_{i=1}^n a_i x_i + b. \tag{4.25}$$

By construction, $\rho_n \leqslant \rho_n^{\text{lin}}$. We solve Problem (4.25) analytically in Proposition 22.

Proposition 22. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking their values in $[0,1]$ and with finite mean $\mathbb{E}[X_i] = \mu$. Then, it holds for all $t \in [0, 1-\mu]$ that:*

$$\rho_n^{\text{lin}}(t) = \frac{\mu}{\mu + t}.$$

Proof. Under the assumptions of Hoeffding's inequality and by symmetry,it holds for $t \geqslant 0$:

$$\rho_n^{\text{lin}}(t) = \inf_{a\in\mathbb{R}, b\in\mathbb{R}} an\mu + b \text{ such that } \forall x \in [0,1]^n, \mathbf{1}_{x_1+\cdots+x_n \geqslant n(\mu+t)} \leqslant a\sum_{i=1}^n x_i + b.$$

It follows that $b = 0$ by considering $x = 0$ and minimizing with respect to b. By construction, $\sum_{i=1}^n x_i \geqslant n(\mu + t)$ implies $1 \leqslant a\sum_{i=1}^n x_i \leqslant an(\mu + t)$. We conclude that $a = \frac{1}{n(\mu+t)}$ and the desired result. □

In Proposition 22, we show that ρ_n^{lin} is exactly equal to the exact bound for one univariate random variable as defined in (4.12), that is $\rho_n^{\text{lin}} = \rho_1$. In addition, ρ_n^{lin} shows a dependence on the mean μ, but no dependence on the number of variables under consideration.

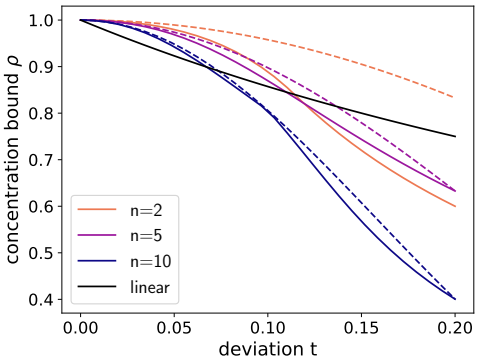


Figure 4.4: Comparison of the bound obtained linear approach (4.25), to the variational (4.18) and exponential (4.16) (in dashed lines) approach, for $\mu = 0.6$ and several values for n.

Surprisingly in Figure 4.4, it happens to improve the variational (4.18) and separable (4.16) approach for small values of the deviation t, even for a large number of variables n. However, for larger values of the deviation, the variational and separable approaches significantly outperform the linear approach. In the following, we introduce a family of polynomials encompassing linear functions, that comes at the cost of computationally more expensive approximations.

4.3.2 A polynomial upper bound: a SoS approximation for polynomial moment assumptions.

This section studies an alternate approach to (4.1) using a family of polynomial upper bounds. Under some assumptions on the degree of these polynomials, it turns out that this family contains

the linear and variational approaches, as feasible points. Its associated Problem (4.5) is therefore guaranteed to provide better approximations to the generalized problem of moments.

We assume $X_1, \dots, X_n$ to be independent random variables with monomials moments of degree at most $a \in \mathbb{N}^\star$, that is $\forall i, \mathbb{E}[g_i(X_i)] = (\mathbb{E}[X_i^k])_{k=1,\dots,a} = (\mu_i^{(k)})_{k=1,\dots,a}$. Denote $J_d = \{(k_1, \dots k_n) \in \mathbb{N}^n,\ k_i \in \mathbb{N},\ k_1 + \dots + k_n \leqslant d\}$ and $\forall \kappa \in J_d$, $\bar{x}^\kappa = \prod_{i=1}^n x_i^{k_i}$ monomials of degree d. The family of (multivariate) polynomials under consideration is

$$\mathcal{Q}_d^n = \left\{ Q(x) = \sum_{\kappa \in J_d} q_\kappa \bar{x}^\kappa, q_\kappa \in \mathbb{R}^{|J_d|} \right\}, \tag{4.26}$$

with a certain degree $d \in \mathbb{N}^+$. In particular, linear functions belongs to this set $\mathcal{L} \subset \mathcal{Q}_d^n$ along with product functions for $d \geqslant n$. Finally, recall that $F(x) = \mathbf{1}_{x \in S}$, where S has a structure specified in Assumption 3.

Assumption 3. *Let $S = \{x \in \mathbb{R}^n, h_1(x) \geqslant 0, \dots, h_m(x) \geqslant 0\}$ and assume that there exists $s(x)$ a polynomial such that $s(x) = s_0(x) + \sum_{j=1}^m s_j(x) h_j(x)$, with $\{x \in \mathbb{R}^n, h(x) \geqslant 0\}$ a compact set and where $h_i(x)$ are polynomial that admits a sum-of-square decomposition.*

In particular, any compact polyhedron verifies Assumption 3 [Bertsimas and Popescu, 2005, Theorem 4.1]. Then, Problem (4.5) associated with polynomials in $\mathcal{Q}_d^n$ takes the form,

$$\begin{aligned} \rho_n^{\text{polynomial},d} = \inf_{Q \in \mathcal{Q}_d^n} \sup_{\forall i, p_i \in \mathcal{P}(\mathcal{X})} & \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} Q(x_1, \dots, x_n) p_1(x_1) \cdots dp_n(x_n), \\ \text{such that } \forall i, \forall k \in 1, \dots, a, & \int_{\mathcal{X}} x_i^k dp_i(x_i) = \mu_i^{(k)}, \text{ and } \forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant Q(x). \end{aligned} \tag{4.27}$$

By construction, $\rho_n \leqslant \rho_n^{\text{polynomial,a}}$.

Proposition 23. *Let $d \leqslant a$, and let the covariance matrix take the form $\sigma_\kappa = \prod_{i=1}^n \mu_i^{(k_i)}$ for $\kappa \in J_d$. Then Problem (4.27) reformulates as:*

$$\rho_n^{\text{polynomial},d} = \inf_{q_\kappa \in \mathbb{R}^{|J_d|}} \sum_{\kappa \in J_d} q_\kappa \sigma_\kappa \text{ such that } \forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant \sum_{\kappa \in J_d} \bar{x}_\kappa q_\kappa. \tag{4.28}$$

Proof. By definition $Q \in \mathcal{Q}_d^n$, $\forall x \in \mathcal{X}^d, Q(x) = \sum_{\kappa \in J_d} q_\kappa \bar{x}^\kappa$. Since $d \leqslant a$ and $\forall i, \forall k \in 1, \dots, a, \int_{\mathcal{X}} x_i^k dp_i(x_i) = \mu_i^{(k)}$, the objective function $\int_{x \in \mathcal{X}^n} Q(x) dp_1(x_1) \cdots dp_n(x_n)$ for $Q \in \mathcal{Q}_d^n$ can be decomposed as the weighted sum of moment μ_i^k. □

By construction, Proposition 23 reveals a condition on the degree d, so that $\mathbb{E}[Q]$ for $Q \in \mathcal{Q}_d$ formulates as a product combination of moments $(\mu_i^{(k)})_{i,k}$. In addition, Problem (4.28) happens to be exactly the Lagrangian dual of the generalized problem of moments applied to multivariate random variable in $\mathbb{R}^n$ [Bertsimas and Popescu, 2005, Equation 2.2]. The major difference lies in the structure of the matrix $\forall \kappa \in J_d, \sigma_\kappa = \mathbb{E}[\prod_{i=1}^n X_i^{k_i}] = \prod_{i=1}^n E[X_i^{k_i}]$ due to independence.

Problem (4.28) suffers from an infinite number of constraints "$\forall x \in \mathcal{X}^n, \mathbf{1}_{x \in S} \leqslant \sum_{\kappa \in J_d} \bar{x}_\kappa q_\kappa$". Connections between such nonnegative polynomial and sum-of-square decomposition have been hightlighted by several authors [Lasserre, 2008, de Klerk and Laurent, 2019], and were formulated as SDPs by Parrilo [2003] and Lasserre [2001]. In their study of optimal bounds for the generalized problem of moments, Bertsimas and Popescu [2005] constructed a sequence of SDPs approximating $\rho_n^{\text{polynomial},d}$, that is recalled below.

Theorem 28. *[Bertsimas and Popescu, 2005, Theorem 4.3] Let $S = \{h_j(x) \geqslant 0, j = 1, \dots, m\}$ and $\mathcal{X} = \{\omega_j(x) \geqslant 0, j = 1, \dots, l\}$ verify Assumption 3, that is such that there exist polynomials*

s and p such that:

$$\forall x \in S, s(x) = s_0(x) + \sum_{i=1}^{m} s_i(x) h_i(x),$$
$$\forall x \in \mathcal{X}, p(x) = p_0(x) + \sum_{i=1}^{l} p_i(x) \omega_i(x),$$

where $(s_i)_{i=1,\dots,m}, (p_i)_{i=1^l}$ are polynomials admitting a sum-of-square decomposition. Then, for every $\epsilon > 0$, there exists a nonnegative integer $r \in \mathbb{N}$ such that $|\rho_n^{\text{polynomial},d} - \tilde{\rho}_n^{\text{polynomial,d}}(r)| \leqslant \epsilon$, where $\tilde{\rho}_n^{\text{polynomial,d}}(r)$ is the value of the following SDP:

$$\begin{aligned}
\tilde{\rho}_n^{\text{polynomial,d}}(r) = \inf_{q_\kappa, s_\kappa, p_\kappa, V \succcurlyeq 0, P \succcurlyeq 0} & \sum_{\kappa \in J_d} q_\kappa \sigma_\kappa, \\
\text{such that } q_\kappa - \delta_{\kappa=0} &= s_\kappa^0 + \sum_{i=1}^{m} \sum_{\eta,\theta \in J_r, \eta+\theta=\kappa} s_\eta^i h_\theta^i, \ \forall \kappa \in J_d, \\
0 &= \sum_{i=1}^{m} \sum_{\eta,\theta \in J_r, \eta+\theta=\kappa} s_\eta^i h_\theta^i, \ \forall \kappa \in J_r \backslash J_d, \\
q_\kappa &= p_\kappa^0 + \sum_{i=1}^{l} \sum_{\eta,\theta \in J_r, \eta+\theta=\kappa} p_\eta^i \omega_\theta^i, \ \forall \kappa \in J_d, \\
0 &= p_\kappa^0 + \sum_{i=1}^{l} \sum_{\eta,\theta \in J_d, \eta+\theta=\kappa} p_\eta^i \omega_\theta^i, \ \forall \kappa \in J_r \backslash J_d, \\
s_\kappa^i &= \sum_{\eta,\theta \in J_r, \eta+\theta=\kappa} v_{\eta,\theta}^i, \ V^i = [v_{\eta,\theta}]_{\eta,\theta \in J_r} \succcurlyeq 0, \ \forall \kappa \in J_r, \ i = 1, \dots, m, \\
p_\kappa^i &= \sum_{\eta,\theta \in J_r, \eta+\theta=\kappa} p_{\eta,\theta}^i, \ P^i = [p_{\eta,\theta}]_{\eta,\theta \in J_r} \succcurlyeq 0, \ \forall \kappa \in J_r, \ i = 1, \dots, l,
\end{aligned} \tag{4.29}$$

where $\sigma_\kappa = \prod_{i=1}^{n} \mu_i^{k_i}$ corresponds to the moments.

Theorem 28 provides a sequence of SDPs approximation Problem (4.28), but does not specify at which degree r a certain precision level ϵ is attained. By construction, it is clear that their degree r must be larger than the degree d of the polynomials under consideration. Bertsimas and Popescu [2005] highlighted a hierarchy between these SDP approximations, which is explicited in Corollary 11.

Corollary 11. *[Bertsimas and Popescu, 2005, Corollary 4.4] Let $\rho_n^{\text{polynomial},d}$ be defined as in (4.28) and $\tilde{\rho}_n^{\text{polynomial,d}}(r)$ as in (4.29). Then, it follows that:*

$$\rho_n \leqslant \rho_n^{\text{polynomial},d} \leqslant \tilde{\rho}_n^{\text{polynomial,d}}(r) \leqslant \cdots \leqslant \tilde{\rho}_n^{\text{polynomial,d}}(1). \tag{4.30}$$

From a numerical perspective, the size of SDPs in (4.29) grows exponentially with degrees r and d, limiting both the precision of the approximation problem (4.28). In the next section however, we apply these approximations to two basic concentration inequalities and improve the existing bounds for a small number of variables n and a small degree r.

4.3.3 Applications to Bernstein's and Bennett's inequalities

The variational and separable approach failed to provide tractable optimization problem for second-order conditions, that appear for example in Bennett's and Bernstein's inequalities. For two random variables, we compute refined bounds using the polynomial approach, and more precisely the SDP aproximations defined in (4.29).

Bernstein's inequality. Bernstein's inequality controls the deviation of the sum of independent random variables to their mean, given an appropriate control of moments. There exists different versions for Bernstein's inequality, and we consider a convenient version requiring finite second-order moments, as provided for example in [Boucheron et al., 2013, Corollary 2.11].

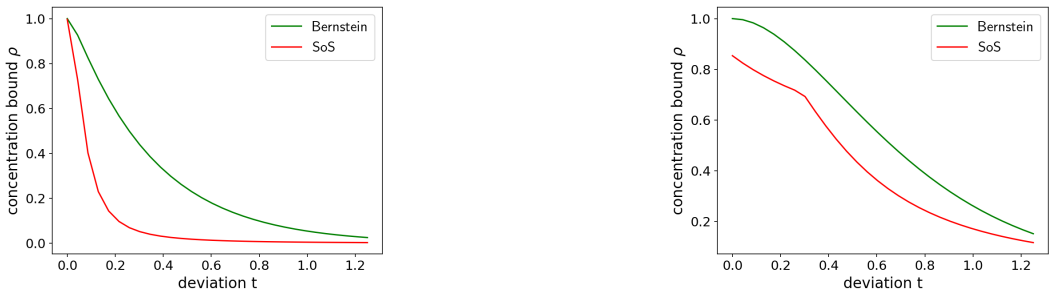
Corollary 12. *Let $X_1, \dots, X_n$ be independent random variables in $[-c, c]$ with means $\mathbb{E}[X_i] = \mu_i$, and variance $v = \sum_{i=1}^n \mathbb{E}[X_i^2]$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geqslant nt + \sum_{i=1}^n \mu_i\right) \leqslant \exp\left(-\frac{n^2 \frac{t^2}{2}}{v + c\frac{nt}{3}}\right). \tag{4.31}$$

Let X_1, X_2 be two i.i.d. random variables taking their values almost surely in $[-1, 1]$ with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu^{(1)}$, and $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = \mu^{(2)}$. We consider the polynomial family $\mathcal{Q}_2 = \{x \in \mathbb{R}^2 \mapsto x_{(2)}^\top Q x_{(2)}, Q \in \mathbb{R}^{(6\times 6)}\}$ where $x_{(2)} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$. Then, Problem (4.27) takes the form:

$$\begin{aligned}\forall t \in \geqslant 0,\ \rho_2^{\text{polynomial},2}(t) = \inf_{Q\in\mathbb{R}^{(6\times 6)}} \operatorname{Tr}(Q\Sigma) \text{ such that } &\forall x \in \mathcal{X}^2,\ 1 \leqslant x_{(2)}^\top Q x_{(2)}, \\ &\forall x \in [-1,1]^2, 0 \leqslant x_{(2)}^\top Q x_{(2)},\end{aligned} \tag{4.32}$$

where $\bar{\mathcal{X}}_2 = \{(x_1, x_2) \in [-1, 1], x_1 + x_2 \geqslant 2(\mu^{(1)} + t)\}$, $\Sigma = \sigma\sigma^\top$ and $\sigma = (1, \mu^{(1)}, \mu^{(1)}, (\mu^{(1)})^2, \mu^{(2)}, \mu^{(2)})$. It results from Theorem 28 that Problem (4.32) can be approximated by a hierarchy of sum-of-square (4.29).



(a) $\mu_1 = -0.3$, $\mu_2 = 0.1$ (b) $\mu_1 = -0.3$, $\mu_2 = 0.5$

Figure 4.5: Comparison of the Bernstein inequality (4.31) as a function of t for $n = 2$, to an SDP approximation in the polynomial approach (4.32) with $d = 2$ and $r = 2$.

Figure 4.5 shows that the SDP approximation of the polynomial approach (4.32) improves upon Bernstein's inequality (4.31) for a small degree in the SoS hierarchy (here $r = 2$). This improvement is not straightforward, since proofs for Bernstein's inequality often require a specific control of moments $\mathbb{E}[X_i^k]$ for all $k \in \mathbb{N}$ (see, e.g.,[Boucheron et al., 2013, proof for Theorem 2.9]). Therefore, considering higher-degree d in the polynomial approximations would probably produce a lower value for $\rho_2^{\text{polynomial,d}}$. However, increasing d requires to increase the degree $r \in \mathbb{R}$ of the sum-of-square approximations (4.29) and thereby, leads to very large SDPs that cannot be handled by our solvers.

Bennett's inequality. Bennett's inequality is similar with Bernstein's inequality, but applies to upper bounded random variables, as recalled in Theorem 29.

Theorem 29. *Bennett [1968] Let $X_1, \cdots, X_n$ be i.i.d. random variables such that $X_i \leqslant a$ almost surely and $\sigma^2 = \sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}X_i)^2]$, then for any $t \geqslant 0$,*

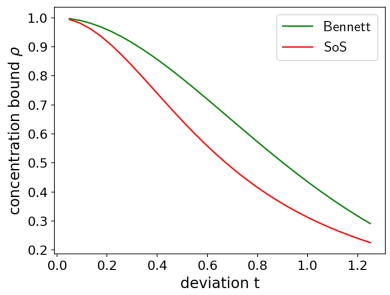
$$\mathbb{P}\left(\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\} \geqslant nt\right) \leqslant \exp\left(-\frac{\sigma^2}{a^2} h\left(\frac{atn}{\sigma^2}\right)\right), \tag{4.33}$$

where $h(t) = (1+t)\log(1+t) - t$. *It implies that* $\mathbb{P}(\sum_{i=1}^n \{X_i - \mathbb{E}[X_i]\} \geqslant nt) \leqslant \exp\left(-\frac{t^2 n^2}{2(\sigma^2 + atn/3)}\right)$.

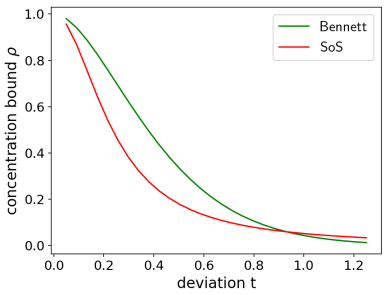
Let X_1, X_2 be two i.i.d. random variables taking their values almost surely in $\mathcal{Z} = [-\infty, 1]$ with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu^{(1)}$, and $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = \mu^{(2)}$. Then, Problem (4.28) takes the form:

$$\rho_2^{\text{polynomial},2}(t) = \inf_{Q \in \mathbb{R}^{6\times 6}} \operatorname{Tr}(Q\Sigma) \text{ such that } \begin{array}{l} \forall x \in \bar{\mathcal{Z}}_2,\ 1 \leqslant x_{(2)}^\top Q x_{(2)}, \\ \forall x \in [-\infty, 1]^2, 0 \leqslant x_{(2)}^\top Q x_{(2)}, \end{array} \tag{4.34}$$

where $\bar{\mathcal{Z}}_2 = \{(x_1, x_2) \in [-\infty, 1], x_1 + x_2 \geqslant 2(\mu^{(1)} + t)\}$, $\nu = (1, \mu^{(1)}, \mu^{(1)}, (\mu^{(1)})^2, \mu^{(2)}, \mu^{(2)})$ and $\Sigma = \nu\nu^\top$.



(a) $\mu_1 = -0.3$, $\mu_2 = 1$



(b) $\mu_1 = -0.3$, $\mu_2 = 0.2$

Figure 4.6: Comparison of Bennett's inequality (4.33) as a function of t for $n = 2$, to the polynomial approach (4.34) with $d = 2$ and $r = 2$ (see the optimization problem in (4.29)).

Figure 4.6 reveals the limit of the polynomial approach, which fails to improve Bennett's inequality for all values of $t \geqslant 0$. The proof relies indeed on additional arguments, such as Jensen's inequality, or Taylor approximations such as $u \mapsto \log(1+u) \leqslant u$, which are not exploited in this approach.

To conclude, the polynomial approach (4.28) effectively refines Bernstein and Bennett's inequalities for some range of values $t \geqslant 0$. However it is hindered by the necessity of large SDP approximations: handling higher-order polynomials and a large number of random variables increases the size of the SDPs under consideration, making them difficult to solve. A review of the complexity of semidefinite optimization and interior point methods can be found in Wolkowicz et al. [2000], Vandenberghe and Boyd [1996].

4.3.4 Higher-order polynomial approximations : a variational reformulation.

The variational, separable and linear approaches allowed refining Hoeffding's inequality. In the polynomial approach developed above, Proposition 23 entails that the degree of the polynomials under consideration is controlled by the highest-order moment. In the context of Hoeffding's inequality, where the first-order moment is finite, it implies that the best polynomial representation is actually linear ($d \leqslant 1$).

In the following, we introduce a numerical procedure to incorporate higher-order polynomials for studying Hoeffding's inequality. We then lay the foundation for a feature-based approach that generalizes the polynomial approach to broader families of upper functions. Finally, we derive an SDP relaxation for the case of two independent random variables.

4.3.4.1 Studying Hoeffding's inequality with second-order polynomials.

Hoeffding's inequality requires finite first-order moments, together with almost surely bounded variables. In what follows, we show how it affects second-order moments. Based on this, we

integrate second-order moment conditions into a polynomial approach for studying Hoeffding's inequality.

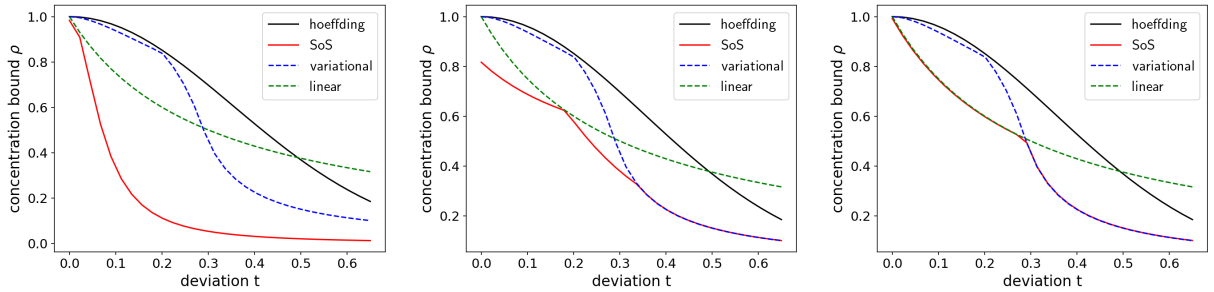
Let $X_1, \dots, X_n$ be independent random variables taking their values in $[0,1]$ almost surely with $\mathbb{E}[X_i] = \mu_i^{(1)}$ for all $i = 1, \dots, n$. Lemma (4.C) ensures the existence and a control on the second-order moment (see proof in Appendix 4.C).

Lemma 18. *Let X be a random variable almost surely on $[0,1]$ with mean $\mathbb{E}[X] = \mu^{(1)}$. Then X admits a finite second-order moment, denoted $\mathbb{E}[X^2] = \mu^{(2)}$. In addition, it holds that $(\mu^{(1)})^2 \leqslant \mu^{(2)} \leqslant \mu^{(1)}$.*

Lemma 18 provides a control of the second-order moment by the first-order moment for bounded random variables. From that, we define an optimization problem based on the polynomial approach (4.28):

$$\begin{aligned}\tilde{\rho}_n^{\text{polynomial},2} = \inf_{\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]} \inf_{Q \in \mathcal{Q}_2^n} \sup_{p_1, \dots, p_n \in \mathcal{P}(\mathcal{X})} & \int_{\mathcal{X}} \dots \int_{\mathcal{X}} Q(x_1, \dots, x_n) dp_1(x_1) \cdots dp_n(x_n), \\ \text{such that } \forall i, & \int_{\mathcal{X}} x_i dp_i(x_i) = \mu_i^{(1)}, \\ \forall i, & \int_{\mathcal{X}} x_i^2 dp_i(x_i) = \mu_i^{(2)}, \\ \forall x \in \mathcal{X}^n, & \mathbf{1}_{x \in S} \leqslant Q(x), \\ \tilde{\rho}_n^{\text{polynomial},2} = \inf_{\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]} & \rho_n^{\text{polynomial},2}(\mu^{(2)}).\end{aligned} \tag{4.35}$$

Given a value of $\mu^{(2)} \in [(\mu^{(1)})^2, \mu^{(1)}]$, the inner optimization problem defining $\rho_n^{\text{polynomial},2}$ can be approached with a hierarchy of sum-of-square defining $\rho_n^{\text{polynomial},2}(r, \mu^{(2)})$, as stated in Proposition 23 and Theorem 28. In Figure 4.7, we use a gridsearch procedure on $\mu^{(2)}$ for approximating $\rho_n^{\text{polynomial},2}(\mu^{(2)})$. It turns out that optimizing $\rho_n^{\text{polynomial},2}(\mu^{(2)})$ with respect to $\mu^{(2)}$ aligns exactly with the minimum bound achieved by the linear (4.25) and variational (4.18) approaches.



(a) $\mu_1 = 0.3$ and $\mu_2 = 0.1$ (b) $\mu_1 = 0.3$ and $\mu_2 = 0.2$ (c) Optimized version for $\mu_1 = 0.3$.

Figure 4.7: Comparison for $n = 2$ of the bound in polynomial approach (4.28) (with an SDP approximation of degree $r = 3$), to the variational (4.18), the linear approach (4.25), and to Hoeffding's inequality (4.11), as a function of the deviation parameter t.

In conclusion, linear functions and product-functions appear to be the minimal polynomial families suitable for studying Hoeffding's inequality. Although the polynomial approach (4.35) incurs high computational costs, the linear approach offers a closed-form solution and the variational approach provides tractable reformulations for i.i.d. random variables and structured independent variables.

4.3.4.2 Generalization to tighter bounds for F: a variational feature-based formulation

We have previously explored the family of product-functions, linear functions as well as polynomials. These approaches were limited by the power of representation allowed by moments (see Proposition 23). In the context of Hoeffding's inequality, we addressed higher-order polynomials using finite lower-order moments through a numerical procedure. We now extend this approach by laying the foundations for a feature-based approach.

Let us start by recalling the variational formulation (4.5):

$$\inf_{H\in\mathcal{H}} \sup_{\forall i, p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} H(x)dp_1(x_1)\cdots dp_n(x_n) \text{ such that } \forall x\in\mathcal{X}^n, F(x)\leqslant H(x),$$

where $\mathcal{H}$ is a family of upper bounds. We introduce a feature vector $\phi : \mathcal{X}\mapsto\mathbb{R}^l, l\in\mathrm{N}$ such that functions have the following representation:

$$F(x) = \left\langle \bar{F}, \bigotimes_{i=1}^n \phi(x_i)\right\rangle = \sum_{i_1,\ldots,i_n=1}^{l} \bar{F}_{i_1,\ldots,i_n}\prod_{k=1}^n \phi(x_k)_{i_k}.$$

In the polynomial (resp. variational) approach for examples, features ϕ were monomials (resp. linear functions). Let us consider a family of functions H decomposing with respect to these features:

$$\begin{aligned}\sup_{\forall i, p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} H(x)dp_1(x_1)\cdots dp_n(x_n) &= \sup_{\forall i, p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})} \left\langle \bar{H}, \bigotimes_{i=1}^n \int_{\mathcal{X}} \phi(x_i)dp_i(x_i)\right\rangle,\\ &= \sup_{\forall i,\sigma_i\in\mathcal{K}(\mu_i)} \left\langle \bar{H}, \bigotimes_{i=1}^n \sigma_i\right\rangle,\end{aligned}$$

where $\mathcal{K}(\mu_i)$ is the set of achievable moments $\mathbb{E}_{p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})}[\phi(X_i)] = \int_{\mathcal{X}}\phi(x_i)dp_i(x_i)$ such that $p_i\in\mathcal{P}_{\mu_i}(\mathcal{X})$. In the context of Hoeffding's inequality, the set of achievable moments $\mathcal{K}(\mu_i)$ corresponds to the intuitive idea that there is no assumption on second-order moments, but that they are related to lower-order moments. Then, the overall optimization problem takes the form:

$$\inf_{H\in\mathcal{H}} \sup_{\forall i,\sigma_i\in\mathcal{K}(\mu_i)} \left\langle \bar{H}, \bigotimes_{i=1}^n \sigma_i\right\rangle \text{ such that } \forall x\in\mathcal{X}^n, \left\langle \bar{F}-\bar{H}, \bigotimes_{i=1}^n \phi(x_i)\right\rangle \leqslant 0. \tag{4.36}$$

Compared to the polynomial approach developed above (4.28), Problem (4.36) provides a more generic formulation for any family of features. The family of features must satisfy two key components : a (tighter) relaxation of the constraint $\forall x\in\mathcal{X}^n, \langle\bar{F}-\bar{H}, \bigotimes_{i=1}^n\phi(x_i)\rangle\leqslant 0$, which can be managed using sum-of-square for polynomial features, and a relaxation for $\sup_{\forall i,\sigma_i\in\mathcal{K}(\mu_i)}\langle\bar{H}, \bigotimes_{i=1}^n\sigma_i\rangle$. In the following proposition, we analyze a simple relaxation of $\sup_{\forall i,\sigma_i\in\mathcal{K}(\mu_i)}\langle\bar{H}, \bigotimes_{i=1}^n\sigma_i\rangle$ for two i.i.d. random variables in the context of Hoeffding's inequality (i.e., $n=2$). For this case, the tensor formulation simplifies into matrices and admits an SDP relaxation.

Proposition 24. *Let X_1, X_2 be i.i.d. random variables taking their values almost surely in $[0,1]$ with finite mean $\mu^{(1)}$ and let $\phi(x) = (1, x, x^2)$ be the feature vector. Then,*

$$\sup_{\sigma\in\mathcal{K}(\mu^{(1)})} \mathrm{Tr}\left(\tilde{H}\sigma\sigma^\top\right) \leqslant \sup_{M\succcurlyeq 0} \mathrm{Tr}\left(\tilde{H}M\right) \text{ such that } 0\leqslant \mathrm{Tr}(ME_{1,1})\leqslant \mathrm{Tr}(ME_{0,2})\leqslant \mathrm{Tr}(ME_{0,1})\leqslant 1,$$
$$\mathrm{Tr}(ME_{0,0}) = 1.$$

Proof. Let $\tilde{H} \in \mathbb{R}^{3\times 3}$ be a symmetric matrix representation of vector $\bar{H} \in \mathbb{R}^6$, such that $\sup_{\sigma\in\mathcal{K}(\mu^{(1)})} \mathrm{Tr}(\tilde{H}\sigma\sigma^\top) = \sup_{M\succcurlyeq 0} \mathrm{Tr}(\tilde{H}M)$ such that $M \in \mathrm{hull}\{\sigma\sigma^\top, \sigma \in \mathcal{K}(\mu)\}$. By definition, $\sigma = (1, \mu^{(1)}, \mu^{(2)})$ holds for the first and second-order moments. We relax this problem by optimizing over $M \succcurlyeq 0$ and incorporating additional constraints on M to ensure it accurately incorporates relationships between the first and second-order moments (Lemma 18, namely $(\mu^{(1)})^2 \leqslant \mu^{(2)} \leqslant \mu^{(1)}$ and $\int_0^1 dp(x) = 1.$). □

Proposition 24 provides an SDP relaxation for approximating Hoeffding's inequality for two random variables and given a quadratic features. Despite the simple relaxation, solving this problem requires to solve efficiently a saddle-point problem and to quantify how far such a relaxation is from the generalized problem of moments (4.1), which are left to future work.

4.4 Conclusion and future works.

Conclusion. In this work, we have introduced two families of upper bounds for approximating the generalized problem of moments for independent random variables. First, we studied a separable approach, leveraging specific upper functions tailored to finite first-order moments. This approach is complemented by a convex variational method, optimizing over the entire family of product-functions. When studying Hoeffding's inequality, these formulations are particularly effective for both i.i.d. random variables and cases where variables are divided into groups with different means, facilitating the reconstruction of associated extremal distributions. Due to the computational limitations of the product-functions based approaches, we broaden our scope by introducing a polynomial family of upper bounds. Here, carefully selected polynomials were employed, resulting in non-negative polynomials that can be approximated using sum-of-square techniques, at a higher computational cost. This framework enabled us to explore concentration properties concerning Bennett's and Bernstein's inequalities, although it failed to improve theoretical bounds universally. We finally extended the polynomial approach into a feature-based approach, thus using polynomials independently of the order of moment assumptions. Without focusing on computational efficiency, it offers a more flexible and comprehensive method for studying concentration inequalities. In summary, our methodologies were each tailored to address different complexities inherent in the problem of moments and the independence.

Future works. Throughout this work, we have highlighted several limitations related to these approaches. The separable approach could benefit from exploring and constructing new product-functions that lead to closed-form solutions, going beyond moment-generating functions. Meanwhile, the polynomial approach, when approximated by SDPs of increasing sizes, converges slowly even for a reasonable number of variables and low-degree polynomials. A deeper study of the connection between the family of polynomials to the linear and product-functions would help simplify the underlying optimization problems. In addition, numerical experiments have shown the limitations of the polynomial approach, which does not take into account some analytical arguments (such as convexity or inequalities derived from Taylor expansion). Introducing some key components of these analyses would probably help improve (or match) known bounds. Furthermore, the feature-based approach could still benefit from exploring efficient relaxations to improve known bounds.

Finally, the generalized problem of moments requires few assumptions on the distribution under consideration, such as bounded moments or random variables lying in bounded sets (e.g., intervals). Other paths to improvements of concentration bounds without independence were explored, such as exploiting additional possible structural properties of the distributions to be characterized. For instance, Popescu [2005] extended the generalized problem of moments to convex classes of distributions, such as unimodal or bimodal distributions. Given unimodal distributions, Van Parys et al. [2015] reconstructed non-discrete extremal distributions, which

are thereby more representative of the effective behaviors of random variables. Extending their work to the case of independent variables would probably help in refining inequalities that can be adapted to different problem structures. In particular, it would be of interest to introduce subgaussian assumptions, that appears in many probability proofs in machine learning, thus improving probabilistic bounds to such problems.

Appendices

4.A Exact formulation of generalized problem of moments for independent random variables using optimal transport:

This proof is an alternative to the convexification of the variational approach from Proposition 14. Using optimal transport, we compute an exact formulation for the generalized problem of moments (4.1), and recover the convex reformulation for the variational approach as result of weak duality.

Let us introduce for $x \in \mathcal{X}^n$, $G(x) = \log(F(x_1, ..., x_n))$. Then, the exact formulation takes the form for the KL divergence $D(q \,||\, p) = \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) dq(x)$:

$$\begin{aligned}\log(\rho_n) &= \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \log\left(\int_{\mathcal{X}^n} e^{G(x)} dp_1(x_1) \cdots dp_n(x_n)\right), \\ &= \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \sup_{\forall i, p_i \in \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q \,||\, \prod_{i=1}^n p_i\right),\end{aligned}$$

By Donsker-Varadhan's inequality. By Pythagorean theorem for the KL divergence and mutual information, we have:

$$\log(\rho_n) = \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \sup_{\forall i, p_i ß \mathcal{P}_{\mu_i}(\mathcal{X})} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q \,||\, \prod_{i=1}^n q_i\right) - \sum_{i=1}^n D(p_i \,||\, q_i).$$

The variational relaxation corresponds in fact to using the fact that $D(q \,||\, \prod_{i=1}^n q_i) \geqslant 0$. In addition, we recover the convex formulation from the Lagrangian relaxation of:

$$\begin{aligned}\log(\rho_n) = \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \inf_{\forall i, \alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} &\int_{\mathcal{X}^n} G(x) dq(x) - D\left(q \,||\, \prod_{i=1}^n q_i\right) + \sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} \\ &+ \sup_{x \in \mathcal{X}^n} \{G(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g(x_i))\}, \\ \leqslant \inf_{\forall i, \alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^m} &\sum_{i=1}^n \{\alpha_i + \beta_i^\top \mu_i - 1\} + \sup_{x \in \mathcal{X}^n} \{G(x) - \sum_{i=1}^n \log(\alpha_i + \beta_i^\top g(x_i))\} \\ &+ \sup_{q \in \mathcal{P}(\mathcal{X}^n)} \int_{\mathcal{X}^n} G(x) dq(x) - D\left(q \,||\, \prod_{i=1}^n q_i\right),\end{aligned}$$

by weak duality. Ignoring the term, $D\left(q \,||\, \prod_{i=1}^n q_i\right)$, the maximum of $\sup_{\mathcal{X}^n} G(x) dq(x) = \max_{x \in \mathcal{X}^n} G(x)$ with respect to $q \in \mathcal{P}(\mathcal{X}^n)$ is attained at a Dirac, at which $D\left(q \,||\, \prod_{i=1}^n q_i\right) = 0$. Thus, the fact that $\rho_n \leqslant \rho_n^{\text{var}}$ holds by weak duality.

4.B Proofs for Hoeffding's inequality

4.B.1 Separable approach for univariate distributions

4.B.1.1 Proof for Proposition 16

Let us introduce a family of functions $u_\lambda(x) = e^{\lambda(x-(\mu+t))}$. Then, for every $t \geqslant 0, \lambda \in \mathbb{R}$,

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha\in R, \beta\in\mathbb{R}} \alpha + \beta\mu \text{ such that } \forall x \in X, e^{\lambda(x-(\mu+t))} \leqslant \alpha + \beta x.$$

The function $x \mapsto e^{\lambda(x-(\mu+t))} - (\alpha + \beta x)$ is convex on $[0, 1]$. Applying Proposition 12, we have:

$$\rho_1^{\exp}(\lambda, t) = \inf_{\alpha\in R, \beta\in\mathbb{R}} \alpha + \beta\mu \text{ such that } e^{-\lambda(\mu+t)} \leqslant \alpha \text{ and } e^{s(1-(\mu+t))} \leqslant \alpha + \beta.$$

The solution is given by $\alpha = e^{-\lambda(\mu+t)}$ and $\beta = e^{-\lambda(\mu+t)}(e^\lambda - 1)$, and thus we have $\rho_1^{\exp}(\lambda, t) = e^{-\lambda(\mu+t)}\Big(1 + \mu(e^\lambda - 1)\Big)$. By optimizing over λ, we have $\lambda_\star = \log\left(\frac{(\mu+t)(1-\mu)}{\mu(1-(\mu+t))}\right)$ and it holds for all $t \in [0, 1 - (\mu)]$ with $\nu = \mu + t$:

$$\rho_{1,\star}^{\exp}(t) = \left(\frac{\mu}{\nu}\right)^\nu \left(\frac{1-\mu}{1-\nu}\right)^{1-\nu}.$$

4.B.1.2 Alternative proof for Proposition 16

Let X be a random variable taking its value almost surely in $[0, 1]$, with mean $\mathbb{E}[X] = \mu$ and associated with the distribution $p \in \mathcal{P}([0, 1])$. Then, for $KL(p, q) = \int_X \log(\frac{p(x)}{q(x)})dp(x)$ and $kl(\mu, \nu) = \log(\frac{\mu}{\nu})\mu$, the Kullback-Leibler divergence, it holds that

$$\begin{aligned}
\log \mathbb{P}(X \geqslant t) &\leqslant \inf_s -st + \log(\mathbb{E}[e^{sX}]) \\
&\leqslant \inf_s -st + \sup_q s\mathbb{E}_q[X] - KL(q, p), \\
&= \inf_s -st + \sup_\nu s\nu - kl(\nu, \mu), \\
&= -kl(t, \mu).
\end{aligned}$$

4.B.2 Variational approach with equal means: computing the extremal points for $\bar{X}_n$

Recall the optimization problem in the variational approach:

$$\rho_n^{\exp} = \inf_{\alpha,\beta,t\geqslant 0} n(\alpha + \beta\mu - 1) \text{ such that } \begin{array}{r} -\sum_{i=1}^n \log(\alpha + \beta x_i^j) \leqslant 0,\ x^j \in \text{extremal}(\bar{\mathcal{X}}_n), \\ \alpha + \beta x \geqslant 0,\ x \in [0, 1], \end{array}$$

where $\bar{\mathcal{X}}_n = \{(x_1, \ldots, x_n) \in [0, 1]^n, x_1 + \cdots + x_n \geqslant n(\mu + t)\}$. Constraints $- - \sum_{i=1}^n \log(\alpha + \beta x_i^j) \leqslant 0$ and $x_1 + \cdots + x_n \geqslant n(\mu + t)$ are symmetric in the coordinates, meaning that they have the same value when permuting $x_1^j, \ldots, x_n^j$. Therefore, the optimization problem requires to formulate only $O(n)$ extremal points, depending on the value for $n(\mu + t) \in [0, 1]$:

- If $n \leqslant n(\mu + t) > n - 1$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (n(\mu + t) - (n - 1), 1, \ldots, 1)\}$,
- If $n - 2 < n(\mu + t) \leqslant n - 1$: Extremal$(\mathcal{X}^n) = \{(1, \ldots, 1); (0, 1, \ldots, 1); (0, n(\mu + t) - (n - 2), 1, \ldots, 1)\}$,

- If $n-3 < n(\mu+t) \leqslant n-2$: Extremal$(\mathcal{X}^n) = \{(1,\dots,1);(0,1,\dots,1);(0,0,1,\dots,1);(0,0,n(\mu+t)-(n-3),1,\dots,1)\}$,
- $\dots$,
- If $n(\mu+t) \leqslant 1$: Extremal$(\mathcal{X}^n) = \{(1,\dots,1);(0,1,\dots,1);\dots;(0,\dots,0,1);(0,\dots,0,n(\mu+t))\}$.

Thus, there are at most $\lfloor n(\mu+t)\rfloor + 1$ extremal points to consider.

4.B.3 Closed-form solution to the variational reformulation for $n = 2$

Recall the optimization problem under consideration for $n = 2$, for two i.i.d. random variables taking their values in $[0,1]$ with mean $\mu \in \mathbb{R}$,

$$\log \rho_n^{\exp}(t) = \inf_{\alpha,\beta,t\geqslant 0} 2(\alpha+\beta\mu-1) \text{ such that } \quad -\log(\alpha+\beta x_1^j) - \log(\alpha+\beta x_2^j) \leqslant 0,\ x^j \in \text{extremal}(\bar{\mathcal{X}}_2), \tag{4.37}$$

where $\bar{\mathcal{X}}_2 = \{(x_1,x_2)\in[0,1]^2, x_1+x_2 \geqslant 2(\mu+t)\}$. Extremal points of $\bar{\mathcal{X}}_2$ depends on the value of $\mu+t$:

- If $1/2 \leqslant \mu+t \leqslant 1$, extremal$(\bar{\mathcal{X}}_2) = \{(1,1),(1,2(\mu+t)1)\}$;
- If $0 \leqslant \mu+t \leqslant 1/2$, extremal$(\bar{\mathcal{X}}_2) = \{(1,1),(0,1),(0,2(\mu+t)\}$.

We derive an optimal solution to problem (4.37) together with optimal values $(\alpha_\star,\beta_\star)$ in the following proposition.

Proposition 25. *Let $\mu \in [0,1]$, and $t \in [0,1-\mu]$. Then, optimal solutions $(\alpha_\star,\beta_\star)$ to (4.37) verify:*

$$\alpha_\star = \begin{cases} \sqrt{\frac{\mu}{\mu+t}} & \text{if } 0 \leqslant t \leqslant \frac{1}{2}-\mu, \\ \sqrt{\frac{1-2t-\mu}{1-\mu}} - \frac{t(2(\mu+t)-1)}{\sqrt{(1-\mu)(1-2t-\mu)(1-(\mu+t))}} & \text{if } \frac{1}{2}-\mu \leqslant t \leqslant \frac{(1-\mu)^2}{2-\mu}, \\ 0 & \text{if } t \geqslant \frac{(1-\mu)^2}{2-\mu}, \end{cases}$$

$$\beta_\star = \begin{cases} \frac{t}{(\mu+t)\sqrt{\mu(\mu+2t)}} & \text{if } 0 \leqslant t \leqslant \frac{1}{2}-\mu, \\ \frac{t}{(1-(\mu+t))\sqrt{(1-\mu)(1-\mu-2t)}} & \text{if } \frac{1}{2}-\mu \leqslant t \leqslant \frac{(1-\mu)^2}{2-\mu}, \\ \frac{\mu}{\sqrt{2(\mu+t)-1}} & \text{if } t \geqslant \frac{(1-\mu)^2}{2-\mu}. \end{cases}$$

Proof. The proof is divided into two parts, depending on the value for $\mu+t$.

1. Let us first assume that $\mu+t \in]1/2,1]$. Then, the optimization problem (4.37) takes the form:

$$\begin{aligned} \inf_{\alpha,\beta}\ 2(\alpha+\beta\mu-1), \text{ such that } \alpha \geqslant 0, & \qquad (\times\lambda_1) \\ -2\log(\alpha+\beta) \leqslant 0, & \qquad (\times\lambda_2) \\ -\log(\alpha+\beta) - \log(\alpha+(2(\mu+t)-1)\beta) \leqslant 0 & \qquad (\times\lambda_3), \end{aligned}$$

where $\lambda_1,\lambda_2,\lambda_3 \geqslant 0$ are dual variables. Reformulating the second constraint into "$\alpha+\beta \geqslant 1$",

it still holds that this problem is a convex. We compute the KKT conditions:

$$\begin{aligned}
2 &= \lambda_1 + \lambda_2 + \lambda_3 \left(\frac{1}{\alpha+\beta} + \frac{1}{\alpha + (2(\mu+t)-1)\beta} \right), \\
2\mu &= \lambda_2 + \lambda_3 \left(\frac{1}{\alpha+\beta} + \frac{(2(\mu+t)-1)}{\alpha + (2(\mu+t)-1)\beta} \right), \\
0 &= \lambda_1 \alpha, \\
0 &= \lambda_2 (1 - (\alpha+\beta)), \\
0 &= \lambda_3 (-\log(\alpha+\beta) - \log(\alpha + (2(\mu+t)-1)\beta)), \\
0 &\leqslant \lambda_1, \lambda_2, \lambda_3.
\end{aligned}$$

We proceed by a distinction of cases:

- Assume first that $\lambda_1 \neq 0$, then $\alpha = 0$. If $\lambda_2 \neq 0$, then $\beta = 1$ and $\lambda_3 = 0$, $\lambda_2 = 2 = 2\mu$, which is false. We conclude that $\lambda_2 = 0$, and that $\lambda_3 \neq 0$. This fixes the value for β, by solving $\log(\beta) + \log(\beta(2(\mu+t)-1)) = 0$ and injecting this relationship into the two first KKT conditions above.

- Assume now that $\alpha \neq 0$. Then, $\lambda_1 = 0$. At least λ_2 or λ_3 must be nonzero, and both cannot be nonzero if $t + \mu \neq 1/2$. Assuming $\log(\alpha+\beta) + \log(\alpha + (2(\mu+t)-1)\beta) \neq 0$ entails $\lambda_3 = 0$ and $\lambda_2 \neq 0$, that is $\alpha + \beta = 1$. Then, $\log(\alpha + (2(\mu+t)-1)\beta) < 0$ which is false. We conclude that $\log(\alpha+\beta) + \log(\alpha + (2(\mu+t)-1)\beta) = 0$, and thus, that $\lambda_3 \neq 0$ and $\lambda_2 = 0$. This leads to the solutions :

$$\begin{aligned}
\alpha_\star &= \sqrt{\frac{1-2t-\mu}{1-\mu}} - \frac{t(2(\mu+t)-1)}{\sqrt{(1-\mu)(1-2t-\mu)}(1-(\mu+t))}, \\
\beta_\star &= \frac{t}{(1-(\mu+t))\sqrt{(1-\mu)(1-\mu-2t)}},
\end{aligned}$$

  for which $\alpha_\star > 0$ for $t \leqslant \frac{(1-\mu)^2}{2-\mu}$.

2. We now consider the case where $\mu + t \in [0, 1/2]$. The optimization problem under consideration takes the form:

$$\begin{aligned}
\inf_{\alpha,\beta} \; & 2(\alpha + \beta\mu - 1), \text{ such that } \alpha \geqslant 0, && (\times\lambda_1) \\
& \alpha + \beta \geqslant 1, && (\times\lambda_2) \\
& -\log(\alpha) - \log(\alpha+\beta) \leqslant 0 && (\times\lambda_3), \\
& -\log(\alpha) - \log(\alpha + 2(\mu+t)\beta) \leqslant 0 && (\times\lambda_4).
\end{aligned}$$

We compute the KKT conditions, leading to

$$\begin{aligned}
2 &= \lambda_1 + \lambda_2 + \lambda_3 \left(\frac{1}{\alpha} + \frac{1}{\alpha+\beta} \right) + \lambda_4 \left(\frac{1}{\alpha} + \frac{1}{\alpha + \beta 2(\mu+t)} \right), \\
2\mu &= \lambda_2 + \lambda_3 \frac{1}{\alpha+\beta} + \lambda_4 \frac{2(\mu+t)}{\alpha + \beta 2(\mu+t)}, \\
0 &= \lambda_1 \alpha, \\
0 &= \lambda_2 (1 - (\alpha+\beta)), \\
0 &= \lambda_3 (-\log(\alpha) - \log(\alpha+\beta)), \\
0 &= \lambda_4 (-\log(\alpha) - \log(\alpha + \beta 2(\mu+t))), \\
0 &\leqslant \lambda_1, \lambda_2, \lambda_3, \lambda_4.
\end{aligned}$$

First, note that $\alpha > 0$, and thus, $\lambda_1 = 0$. In addition, $\lambda_2 = 0$, otherwise, it implies that $\lambda_3 = \lambda_4 = 0$ and $\mu = 1$, which is impossible. In addition, if $\lambda_3 \neq 0$, then $\alpha(\alpha + \beta) = 1$. Yet, since $\alpha > 0$, $\beta \geqslant 0$. The function $t \mapsto \alpha(\alpha + t\beta)$ is nondecreasing, and thus $\alpha(\alpha + 2(\mu + t)\beta) > 1$, which is impossible. We conclude that $\lambda_3 = 0$. By construction, $\lambda_4 \neq 0$, which implies that $\alpha(\alpha + 2(\mu + t)\beta) = 1$, and thus,

$$2 = \lambda_4(\frac{1}{\alpha} + \alpha),$$
$$2\mu = \lambda_4 \alpha 2(\mu + t),$$

from which we conclude the solutions $\alpha = \sqrt{\frac{\mu}{\mu+t}}$ and $\beta = \frac{t}{(\mu+t)\sqrt{\mu(\mu+2t)}}$. □

4.B.4 Proof for Lemma 17.

Problem (4.20) simplifies into

$$\inf_{\alpha \in \mathbb{R}^2, \beta \in \mathbb{R}^2} m(\alpha_1 + \beta_1 \mu_1) + (n-m)(\alpha_2 + \beta_2 \mu_n),$$
$$\text{such that } \forall x \in \text{extremal}(\bar{\mathcal{X}}_n),\ -\sum_{i=1}^{m} \log(\alpha_1 + \beta_1 x_i) - \sum_{i=m+1}^{n} \log(\alpha_2 + \beta_2 x_i) \leqslant 0,$$

where $\bar{\mathcal{X}}_n = \{(x_1, \ldots, x_n) \in [0,1]^n, x_1 + \ldots + x_n \geqslant nt + m\mu_1 + (n-m)\mu_n\}$. We define $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ and denote $q = n(\bar{\mu}_n + t) - \lfloor n(\bar{\mu}_n + t) \rfloor$. We denote the points with the notation $(x_1, \ldots x_m | x_{m+1}, \ldots, x_n)$. Let $n(\bar{\mu}_n + t) \in [n-k+1, n-k[$. The following assertions are true:

- $(1, \ldots, 1) \in \text{extremal}(\bar{\mathcal{X}}_n)$
- If $(1, \ldots, 1 | 0, 1, \ldots, 1) \in \text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in extremal$(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1, \ldots, 1, 0 | 1, \ldots, 1)$ is in extremal$(\bar{\mathcal{X}}_n)$.
- If $(1, \ldots, 1 | 0, 0, 1, \ldots, 1) \in \text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in extremal$(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1, \ldots, 1, 0 | 0, 1, \ldots, 1)$ and $(1, \ldots, 1, 0, 0 | 1, \ldots, 1)$ are in extremal$(\bar{\mathcal{X}}_n)$.
- If $(1, \ldots, 1 | 0, \ldots 0, 1, \ldots, 1) \in \text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in extremal$(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that the point $(1, \ldots, 1, 0 | 0, \ldots, 0, 1, \ldots, 1), \ldots, (1, \ldots, 1, 0, \ldots, 0 | 1, \ldots, 1)$ are in extremal$(\bar{\mathcal{X}}_n)$ (as long as $k \geqslant m$). That is about $O(|k-m|)$ points.
- If $(1, \ldots, 1 | 0, \ldots, 0, q, 1, \ldots, 1) \in \text{extremal}(\bar{\mathcal{X}}_n)$, all its permutations are in extremal$(\bar{\mathcal{X}}_n)$. By symmetry, it is sufficient to notice that points $(1, \ldots, 1, q | 0, \ldots, 0, 1, \ldots, 1)$, $(1, \ldots, 1, q, 0 | 0, \ldots, 0, 1, \ldots, 1)$, …, $(1, \ldots, 1, q, 0, \ldots, 0 | 1, \ldots, 1)$ are in extremal$(\bar{\mathcal{X}}_n)$. That is, about $O(2|k-m|)$ points.

If they are k zeros elements, there are $O(\sum_{i=1}^{k-m} i) = O(k^2) \leqslant O(n^2)$ points.

4.C Connecting the second-order to the first-order moment: proof for Lemma 18

Let X be a random variable almost surely in $[0,1]$ with mean $\mu^{(1)}$. First, its second-order moment exists. In addition, we have,

$$\begin{aligned}
\mu^{(2)} &\leqslant \sup_{p\in\mathcal{P}([0,1])} \int_0^1 x^2 dp(x), \text{ such that } \int_0^1 x dp(x) = \mu^{(1)}, \\
&= \inf_{\alpha,\beta} \alpha + \beta\mu^{(1)} \text{ such that } \forall x \in [0,1], x^2 \leqslant \alpha + \beta x, \\
&= \inf_{\alpha,\beta} \alpha + \beta\mu^{(1)} \text{ such that } \alpha \geqslant 0, 1 \leqslant \alpha + \beta \text{ (by convexity)}, \\
&= \mu^{(1)},
\end{aligned}$$

with $\alpha = 0, \beta = 1$. In addition, by Cauchy-Schwarz, $\int_0^1 x dp(x) \leqslant \left(\int_0^1 x^2 dp(x)\right)$ that is $\mu^{(2)} \geqslant \mu^{(1)}$.

Conclusion and perspectives

5.1 Summary of the thesis

In this thesis, we have investigated three topics, covering worst-case analysis of first-order methods (and continuous-time models), matching pursuit algorithms and concentration inequalities from the perspective of their proof structure.

More precisely, in Chapter 2, we develop a systematic approach for finding quadratic Lyapunov functions for families of ODEs and SDEs approximating SGD, casting the verification as a small-sized LMI feasibility problem. This formulation allows efficient searching for quadratic Lyapunov functions to derive convergence bounds, requiring fewer assumptions than discrete-time systems and enabling shorter proofs. While obtaining guarantees for stochastic optimization can be complex, the SDE approach simplifies the analysis of the trade-off between variance and initial condition terms. However, this analysis does not include approximation guarantees between optimization methods and their continuous-time counterparts, which are often developed using stability techniques in deterministic settings and stochastic modified equations in stochastic settings, though these theorems usually impose many assumptions that could be simplified with computer-assisted proofs.

Chapter 3 focuses on a principled view for generating optimization algorithms by minimizing a smoothness upper bound with respect to a well-chosen norm. For non-regularized models, this leads to coordinate descent with GS-rule, which can be seen as a matching pursuit algorithm in different geometries for underparametrized and overparametrized models. Building on these results, we introduce a new regularized matching pursuit algorithm based on ℓ_1-norm smoothness minimization, analogous to proximal gradient descent in ℓ_2-geometry. Although closely related to proximal coordinate descent with GS-rule, this regularized algorithm does not fit the matching pursuit framework in gauge geometry for overparametrized models and lacks linear convergence in this regime. We finally formulate an ultimate method adapted to overparametrized geometries, which admits no closed-form formulation and is thereby approximated using an inner-loop strategy. Our approach provides refined convergence guarantees for matching pursuit and coordinate descent with GS-rule, adapted to the problem's geometry. As a byproduct, for linear regression and LASSO, we estimate convergence bounds using SDP relaxations and high probability bounds under statistical assumptions. For the linear regression model, these estimates reveals a transition phase in convergence guarantees from underparametrized to overparametrized regimes. For ℓ_1-regularized models, this transition phase interprets λ as a measure of solution sparsity.

Finally, Chapter 4 is devoted to constructing worst-case approaches for concentration inequalities with independent random variables. We first introduce a separable approach using specific upper product-functions, adapted to finite first-order moments, complemented by a

variational method optimizing over these product-functions. Using this methodology, we refine Hoeffding's inequality for both i.i.d. random variables and random variables divided into two blocks having different means. Regarding higher-order finite moments, we introduce polynomial upper bounds, formulating a sequence of increasing approximating SDPs. However, this polynomial approach poses computational challenges in universally improving theoretical bounds. Finally, we extend the polynomial approach to a feature-based method for greater flexibility in studying concentration inequalities but lacking a tractable formulation. To conclude, each approach addresses different complexities inherent in moment problems and the independence of random variables.

5.2 Perspectives

As mentioned in each of the chapters, this study raises new questions and natural extensions.

1. Chapter 2 studied a well-specified family of quadratic Lyapunov functions provided by (2.14) and (2.11). We mentioned natural extensions by taking terms of the forms $\int_0^t f(X_s)ds$ or integrals of the quadratic terms that are used in our Lyapunov functions. Additionally, we believe the framework can be extended without difficulty to differential [Bolte et al., 2010, Section 3.2] and monotone inclusions [Boţ and Csetnek, 2016, 2017] problems, as Ryu et al. [2020] already found interpolation conditions for monotone inclusions. Furthermore, it is not clear how to use PEP when applied to higher-order methods and assumptions (as it already appears in Ito's Lemma for stochastic differential equations).

2. Chapter 2 also raises the issue of how to related continuous-time models and optimization algorithms, as already explored by Scieur et al. [2017] for quadratics. While integration theory [Gautschi, 2012] produces methods approximating an ODE, it still lacks in clear connections between convergence guarantees in discrete and continuous-time. Wilson et al. [2021] introduced the idea of discretizing Lyapunov functions, which could be approached via the performance estimation framework.

3. Building on the results of Chapter 3, we believe it could be of interest to extend this principled approach to accelerated matching pursuit algorithms (and thus, to accelerated coordinate descent algorithms), and to randomly selected coordinates, such as those of Nesterov and Stich [2017] for nonregularized minimization and Fercoq and Richtárik [2015] or Locatello et al. [2018, Section 3] for composite minimization problems. Furthermore, we have numerically observed how the regularization parameter affects the sparsity of the trajectory and thereby, the convergence of regularized matching pursuit (and proximal gradient descent). It could be of interest to compute its active-set complexity (the number of iterations at which the algorithm has identified the sparsity pattern) as defined by Nutini et al. [2018a], and develop screening rules [Sun and Bach, 2022, Ndiaye et al., 2021].

4. Chapter 4 introduced different approaches to approximate concentration inequalities with independent random variables. An in-depth study of the feature-based model and its possible relaxation would probably allow studying higher-order conditions. Finally, the generalized problem of moments applies to distributions with bounded moments or variables within bounded intervals. Recent research explores improving concentration bounds by leveraging structural properties of convex classes of continuous distributions [Popescu, 2005, Van Parys et al., 2015]. Incorporating such assumptions, as well as subgaussian assumptions that are prevalent in machine learning, could further refine probabilistic bounds in these contexts.

Bibliography

A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 10.0.*, 2022. URL http://docs.mosek.com/9.0/toolbox/index.html.

H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016.

H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1): 115–129, 2015.

F. Bach. High-dimensional analysis of double descent for linear regression with random projections. Technical report, arXiv:2303.01372, 2023a.

F. Bach. *Learning Theory from First Principles*. MIT Press - Draft, 2023b.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Z. Bai and J. W. Silverstein. Spectral analysis of large dimensional random matrices. *Springer Series in Statistics*, (2nde Edition), 2010.

N. Bansal and A. Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2002.

H. H. Bauschke and P. L. Combettes. Convex analysis and monotone operator theory in hilbert spaces. *CMS Books in Mathematics*, 2017.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 32(10):15849–15854, 2019.

G. Bennett. A one-sided probability inequality for the sum of independent, bounded random variables. *Biometrika*, 55(3):565–569, 1968.

S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing References 437 House, Moscow, 1927.

R. Berthier, F. Bach, and P. Gaillard. Accelerated gossip in networks of given dimension using Jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020.

Q. Bertrand, Q. Klopfenstein, P.-A. Bannier, G. Gidel, and M. Massias. Beyond l1: Faster and better sparse models with skglm. In *Advances in Neural Information Processing Systems*, volume 35, pages 38950–38965, 2022.

D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

D. Bertsimas, I. Popescu, and J. Sethuraman. Moment problems and semidefinite optimization. *Handbook of Semidefinite Programming*, pages 311–339, 2000.

R. I. Boţ and E. R. Csetnek. Second order forward-backward dynamical systems for monotone inclusion problems. *SIAM Journal on Control and Optimization*, 54(3):1423–1443, 2016.

R. I. Boţ and E. R. Csetnek. A dynamical system associated with the fixed points set of a nonexpansive operator. *J Dyn Diff Equat*, 29:155–168, 2017.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

K. Borgwardt. The average number of pivot steps required by the simplex-method is polynomial. *Zeitschrift für Operations Research*, 26:157–177, 1987.

R. I. Bot, E. R. Csetnek, and D.-K. Nguyen. Fast OGDA in continuous and discrete time. *Foundations of computational mathematics*, 2023.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities : A non Asymptotic Theory of Independence*. Oxford University Press, 2013.

O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*. Springer Berlin Heidelberg, 2004.

N. Bousselmi, J. M. Hendrickx, and F. Glineur. Interpolation conditions for linear operators and applications to performance estimation problems, 2024a. arXiv preprint arXiv:2302.08781, to appear in SIAM journal on Optimization (2024).

N. Bousselmi, N. Pustelnik, J. M. Hendrickx, and F. Glineur. Comparison of proximal first-order primal and primal-dual algorithms via performance estimation, 2024b. arXiv preprint arXiv:2403.10209.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

K. Bredies and D. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12): 4203–4215, 2005.

A. C. Cauchy. Mécanique Céleste. – Méthodes générales pour la détermination des mouvements des planètes et de leurs satellites. In *Oeuvres complètes*. 1836.

T. Chavdarova, M. I. Jordan, and M. Zampetakis. Last-iterate convergence of saddle point optimizers via high-resolution differential equations. OPT21: 13th Annual Workshop for Optimization in Machine Learning, 2021.

K. Comanor, L. Vandenberghe, and S. Boyd. Semidefinite programming and multivariate Chebyshev bounds. *IFAC Proceedings Volumes*, 39(9):597–601, 2006.

P. L. Combettes and J.-C. Pesquet. *Proximal Splitting Methods in Signal Processing*, pages 185–212. Springer New York, 2011.

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

H. Cramér. Sur un nouveau théorème limite de la théorie des probabilités. *Actualites Scientifiques et Industrielles*, 736:2–23, 1938.

A. d'Aspremont, C. Guzmán, and M. Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

A. d'Aspremont, D. Scieur, and A. Taylor. Acceleration Methods. *Foundations and Trends in Optimization*, 5 (1-2):1–245, 2021.

E. de Klerk and M. Laurent. A Survey of Semidefinite Programming Approaches to the Generalized Problem of Moments and Their Error Analysis. In *World Women in Mathematics 2018*, pages 17–56. 2019.

E. De Klerk, F. Glineur, and A. B. Taylor. Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, 30(3): 2053–2082, 2020.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

J. Diakonikolas and M. Jordan. Generalized momentumbased methods: A hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021.

J. Diakonikolas and L. Orecchia. Alternating randomized block coordinate descent. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1224–1232, 2018.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research (JMLR)*, 17(1):2909–2913, 2016.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Y. Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39: 1–16, 2017.

Y. Drori and A. Taylor. On the oracle complexity of smooth strongly convex minimization. *Journal of Complexity*, 68, 2022.

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 327–336, 2012.

H. Fang, Z. Fan, Y. Sun, and M. P. Friedlander. Greed meets sparsity: Understanding and improving greedy coordinate descent for sparse optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108, pages 434–444, 2020.

M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28:2654–2689, 2018.

O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2): 95–110, 1956.

Y. Freund and R. E. Schapire. A short introduction to boosting. *Japanese Society For Artifical Intelligence*, 14: 771–780, 1999.

M. P. Friedlander, I. Macêdo, and T. K. Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

W. Gautschi. *Numerical Analysis*. Birkhäuser Basel, 2012.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming Series A*, 156:59–99, 2015.

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

T. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, and A. Dieuleveut. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *Mathematical Programming Computation (to appear)*, 2024.

C. Guille-Escuret, B. Goujaud, M. Girotti, and I. Mitliagkas. A study of condition numbers for first-order optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1261–1269, 2021.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

A. J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal Research of the National Bureau of Standards*, 49:263–264, 1957.

B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *International Conference on Machine Learning (ICML)*, volume 70, pages 1549 – 1557, 2017.

K. Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14: 185–197, 1962.

K. Isii. Inequalities of the types of Chebyshev and Cramér-rao and mathematical programming. *Annals of the Institute of Statistical Mathematics*, 16:277–293, 1964.

F. Iutzeler and J. Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28:661–678, 2020.

T. S. Jaakkola and M. I. Jordan. Variational probabilistic inference and the qmr-dt network. *Journal of Artificial Intelligence Research*, 10(1):291–322, 1999.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, volume 1, pages 427–435, 2013.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

R. E. Kalman and J. E. Bertram. Control system analysis and design via the "second method" of Lyapunov: I-continuous-time systems. *Journal of Basic Engineering*, 82(2):371–393, 1960.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

S. P. Karimireddy, A. Koloskova, S. U. Stich, and M. Jaggi. Efficient greedy coordinate descent for composite problems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 161, pages 547–557, 2019.

S. Karlin and W. J. Studden. Tchebycheff systems: With applications in analysis and statistics. *Pure and Applied Mathematics, A Series of Texts and Monographs*, 1966.

D. Kim and K. Jeffrey. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159, 06 2016.

W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, volume 1, 2015.

H. J. Landau. *Moments in Mathematics*. American Mathematical Society, 1998.

P. Langevin. Sur la théorie du mouvement brownien [on the theory of brownian motion]. *C. R. Acad. Sci. Paris.*, 146:530–533, 1908.

J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

J.-B. Lasserre. A semidefinite programming approach to the generalized problem of moments. *Mathematical Programming*, 112:65–92, 2008.

N. Le Roux. Anytime tail averaging. *arXiv preprint arXiv:1902.05083*, 2019.

M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes*, volume 23. Ergebnisse der Mathematik und ihrer Grenzgebiete, 1991.

Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *54th Annual IEEE Symposium on Foundations of Computer Science*, pages 147–156. IEEE Computer Society, 2013.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning (ICML)*, volume 70, pages 2101–2110, 2017.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

F. Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 54, pages 860–868, 2017.

F. Locatello, A. Raj, S. P. Karimireddy, G. Raetsch, B. Schölkopf, S. Stich, and M. Jaggi. On matching pursuit and coordinate descent. In *Proceedings of the International Conference on Machine Learning*, volume 80, pages 3198–3207, 2018.

S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels, in: Les equations aux dérivées partielles. *Editions du centre National de la Recherche Scientifique*, pages 87–89, 1963.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 1(4):457 – 483, 1967.

A. W. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4): 1001–1014, 1960.

M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster lasso-type solvers. *NIPS-OPT*, 2017.

M. Massias, A. Gramfort, and J. Salmon. Celer: a fast solver for the lasso with dual extrapolation. In *Proceedings of the International Conference on Machine Learning*, volume 80, pages 3315–3324, 2018.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(75):667–766, 2022.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine learning*. MIT Press, 2018.

C. Moucer, A. Taylor, and F. Bach. A systematic approach to Lyapunov analyses of continuous-time models in convex optimization. *SIAM Journal on Optimization*, 33(3):1558–1586, 2023.

C. Moucer, A. Taylor, and F. Bach. Constructive approaches to concentration inequalities with independent random variables, 2024a.

C. Moucer, A. Taylor, and F. Bach. Geometry-dependent matching pursuit: a transition phase for convergence on linear regression and lasso, 2024b. arXiv preprint arXiv:2403.03072.

E. Ndiaye, O. Fercoq, and J. Salmon. Screening rules and its complexity for active set identification. *Journal of Convex Analysis*, 28:1053–1072, 2021.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. *Wiley Interscience*, 1983.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 2004.

Y. Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88): 10–11, 2012a.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012b.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125 – 161, 2013.

Y. Nesterov and S. U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

J. Nutini. *Greed is good: greedy optimization methods for large-scale structured problems*. PhD thesis, University of British Columbia, 2018.

J. Nutini, M. Schmidt, and W. Hare. "Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optimization Letters*, 13:645–655, 2018a.

J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the International Conference on Machine Learning*, volume 50, pages 1632–1641, 2018b.

A. Orvieto and A. Lucchi. Continuous-time models for stochastic optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2020.

B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

C. Paquette, B. van Merriënboer, E. Courtney, and F. Pegregosa. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, 23:597–673, 2023.

N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends Optimization*, 1:127–239, 2013.

P. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming, Series B*, 96:293–320, 2003.

F. Pedregosa and D. Scieur. Acceleration through spectral density estimation. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 7553–7562, 2020.

F. Pedregosa, G. Negiar, A. Askari, and M. Jaggi. Linearly convergent frank-wolfe with backtracking line-search. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1–9, 2020.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. Polyak. Introduction to optimization. *Optimization Software New York*, 1987.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

I. Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operation Researche*, 30(3):632–657, 2005.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144, 2014.

H. E. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

W. W. Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 245(1240):1–27, 1958.

D. Ruppert. Efficient estimations from a slowly convergent Robbins Monroe process. technical report., 1988.

E. Ryu and S. Boyd. A primer on monotone operator methods survey. *Applied and computational mathematics*, 15:3–43, 2016.

E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

J. M. Sanz Serna and K. C. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3):1542–1565, 2021.

M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition, 2013. arXiv preprint arXiv:1308.6370.

D. Scieur and F. Pedregosa. Universal asymptotic optimality of Polyak momentum. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 8565–8572, 2020.

D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

S. A. B. Sheng Chen and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

B. Shi, W. J. Su, and M. I. Jordan. On learning rates and Schrödinger operators. *arXiv preprint arXiv:2004.06977*, 2020.

B. Shi, S. Du, M. Jordan, and W. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148, 01 2022.

C. Song, S. Cui, Y. Jiang, and S.-T. Xia. Accelerated stochastic greedy coordinate descent by soft thresholding projection onto simplex. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

D. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, page 296–305, 2001.

W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

J. J. Suh, G. Roh, and E. K. Ryu. Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems. In *Proceedings of the International Conference on Machine Learning*, volume 162, pages 20640–20667, 2022.

Y. Sun and F. Bach. Safe screening for the generalized conditional gradient method. *Open Journal of Mathematical Optimization*, 3:1–35, 2022.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

T. Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

W. Tao, Z. Pan, G. Wu, and Q. Tao. Primal averaging: A new gradient evaluation step to attain the optimal individual convergence. *IEEE transactions on cybernetics*, 50(2):835–845, 2018.

A. Taylor and F. Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory (COLT)*, 2019.

A. Taylor and Y. Drori. An optimal gradient method for smooth strongly convex minimization. *Math. Program.*, 199(1–2):557–594, 2022.

A. Taylor, B. V. Scoy, and L. Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning (ICML)*, volume 80, pages 4897–4906, 2018a.

A. B. Taylor. Interpolation and performance estimation of first-order methods for convex optimization, PhD Thesis, Chapter 3 : Convex interpolation, 2017.

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Performance estimation toolbox (pesto): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283, 2017a.

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017b.

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017c.

A. B. Taylor, J. Hendrickx, and F. Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018b.

A. Tewari, P. Ravikumar, and I. Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

P. Tseng. On accelerated proximal gradient methods for convex-concave algorithms. *submitted to SIAM Journal on Optimization*, 2008.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, B(117):387–423, 2009.

B. Van Parys, P. Goulart, and D. Kuhn. Generalized gauss inequalities via semidefinite programming. *Mathematical Programming*, 156:1–32, 2015.

B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. *SIAM Review*, 49(1):52–64, 2007.

S. R. S. Varadhan. *Large deviations and applications*. École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87, 1988.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. In *Proceedings of the National Academy of Sciences*, 2016.

A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *The Journal of Machine Learning Research (JMLR)*, 22(113):1–34, 2021.

H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of Semidefinite Programming*. Kluwer Academic Publishers, 2000.

P. Xu, T. Wang, and Q. Gu. Accelerated stochastic mirror descent: From continuous-time dynamics to discrete-time algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 1087–1096, 2018a.

P. Xu, T. Wang, and Q. Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning (ICML)*, volume 80, pages 5492–5501, 2018b.

T. Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011a.

T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011b.

X. Zhang, D. Schuurmans, and Y.-l. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

RÉSUMÉ

L'optimisation convexe est centrale en science des données, et fournit des outils efficaces pour optimiser des fonctions convexes sur des ensembles convexes. Elle est à la base de nombreux algorithmes d'apprentissage, assurant sufisamment de robustesse pour des tâches telles que la régression, la classification ou le clustering. Par ailleurs, l'optimisation convexe permet de dériver des inégalités de concentration, qui fournissent des bornes probabilistes caractérisant comportement et les performances des algorithmes sur de grands ensembles de données. Cette thèse offre des méthodologies systématiques, utilisant des techniques d'optimisation convexe, pour étudier les modèles d'optimisation en temps continu, aborder les inégalités de concentration, ainsi qu'améliorer la compréhension et le développement des algorithmes de matching pursuit.
La première partie du manuscrit se concentre sur l'étude des garanties de convergence des équations différentielles issues des méthodes de premier ordre, en utilisant des fonctions de Lyapunov. Nous proposons une approche systématique pour trouver et vérifier des fonctions de Lyapunov pour diverses équations différentielles ordinaire et stochastiques, produisant des résultats de convergence avec moins d'hypothèses. Cette fonctionnalité est intégrée dans le package python PEPit. La deuxième partie examine les méthodes de premier ordre gloutonnes, telles que la descente de coordonnées avec la règle de Gauss-Southwell et le matching pursuit, devenues populaires pour leur tendance à proposer des solutions parcimonieuses. Nous proposons une méthode pour générer des algorithmes de matching pursuit (potentiellement régularisés) adaptés à la géométrie du problème en question, ainsi que leurs garanties de convergence. Dans la dernière partie, nous établissons une approche constructive des inégalités de concentration pour des variables aléatoires indépendantes. D'une part, nous introduisons une approche variationnelle qui généralise l'approche par fonction génératrice de moments. Cette approche est adaptée lorsque le premier moment est fixé. D'autre part, nous proposons une approche polynomiale, basée sur une série d'approximations en sommes de carrés, qui est adaptée aux conditions sur les moments d'ordre supérieur.

MOTS CLÉS

Optimisation convexe, méthodes du premier ordre, techniques assistées par ordinateur, analyse par pire-case, matching pursuit, analyse par fonctions de Lyapunov, inégalités de concentration, problème généralisé des moments.

ABSTRACT

Convex optimization is a cornerstone of data science, providing efficient tools for optimizing convex functions over convex sets. It serves as the foundation for numerous machine learning algorithms, enabling robust solutions for tasks such as regression, classification, and clustering. It is also crucial for deriving (approximation to) concentration inequalities, which provide probabilistic bounds that help understand the behavior and performance of algorithms on large datasets. This thesis offers comprehensive and principled methodologies for studying continuous-time optimization models, approaching concentration inequalities, as well as enhancing the understanding and development of greedy algorithms.
The first part of the manuscript focuses on studying convergence guarantees of ordinary differential equations arising from first-order methods, using Lyapunov functions. We provide a systematic approach to finding and verifying Lyapunov functions for various ordinary and stochastic differential equations, yielding convergence results with fewer assumptions. This functionality is integrated into the PEPit python package. The second part investigates greedy first-order methods, such as steepest coordinate descent and matching pursuit, which has become popular due to their natural tendency to propose sparse solutions. We propose a principled approach to generating (regularized) matching pursuit adapted to problem geometry at hand, along with their refined convergence guarantees. In the last part, we introduce a constructive approach to concentration inequalities for independent random variables. We propose on the one hand a variational approach that extends the classical approach via moment generating function and is particularly suited to first-order moment assumptions. On the other hand, we propose a polynomial approach, based on a hierarchy of sums-of-squares approximations, suited to higher-order moment conditions.

KEYWORDS

Convex optimization, first-order methods, computer-assisted techniques, worst-case analysis, matching pursuit, Lyapunov analysis, concentration inequalities, generalized problem of moments.