université de BORDEAUX

PROGRAMME DE RECHERCHE FRANCE 2030 INTELLIGENCE ARTIFICIELLE

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

# DOCTEUR
# DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

MATHÉMATIQUES APPLIQUÉES ET CALCUL SCIENTIFIQUE

Par **Julien HERMANT**

## Étude d'algorithmes inertiels type Nesterov pour l'optimisation accélérée non convexe et stochastique

Soutenue le 7 Octobre 2026

Membres du jury :

| | | | |
|---|---|---|---|
| M. Jean-François Aujol | Professeur | Université de Bordeaux | Directeur de thèse |
| Mme. Aude Rondepierre | Professeure | INSA Toulouse | Co-directrice de thèse |
| M. Aris Danilidis | Professeur | Université technique de Vienne | Rapporteur |
| M. Adrien Taylor | Chargé de recherche | INRIA Paris | Rapporteur |
| M. Antonin Chambolle | Directeur de Recherche | Université Paris Dauphine-PSL | Examinateur |
| M. Jalal Fadili | Professeur | Université Caen Normandie | Examinateur |
| M. Édouard Pauwels | Professeur | École d'économie de Toulouse | Président |

Membre invité :

| | | | |
|---|---|---|---|
| M. Charles Dossal | Professeur | INSA Toulouse | Co-encadrant |

**Étude d'algorithmes inertiels type Nesterov pour l'optimisation accélérée non convexe et stochastique**

**Résumé :** Les algorithmes d'optimisation jouent un rôle important dans beaucoup d'applications. L'un des critères essentiels d'un bon algorithme est de converger rapidement vers une solution, ce qui peut s'avérer difficile en raison de la grande échelle des problèmes modernes. Parmi les mécanismes qui permettent de concevoir des algorithmes rapides, un d'importance est l'inertie. Il est largement observé qu'ajouter à la descente de gradient un tel mécanisme d'inertie accélère sa convergence. Si ce phénomène d'accélération est bien compris dans le cadre de la minimisation convexe lisse avec des gradients exacts, de nombreux problèmes sont non convexes et reposent sur des estimateurs stochastiques du gradient. Cette thèse propose une étude théorique de l'accélération induite par l'inertie de Nesterov dans de tels cadres. Notre première contribution est motivée par l'observation que, dans plusieurs cadres naturels non convexes ou stochastiques, de nombreuses analyses existantes n'établissent aucune propriété d'accélération de l'inertie. Nous identifions des conditions géométriques sous lesquelles l'inertie produit effectivement une accélération démontrable. Pour les fonctions non convexes satisfaisant une condition de Polyak–Łojasiewicz, nous montrons que la possibilité d'accélération dépend de l'angle entre la direction de descente et la direction vers les minimiseurs. Dans le cadre convexe stochastique en somme finie, nous montrons que l'accélération est gouvernée par la corrélation moyenne entre les gradients qui composent la somme. Notre seconde contribution s'appuie sur le système Nesterov continuisé récemment proposé. Celui-ci peut être compris comme un algorithme de Nesterov analysable par une approche de Lyapunov stochastique en temps continu. Nous étendons son cadre d'application, et l'utilisons pour améliorer les garanties de convergence existantes pour l'optimisation fortement quasar-convexe, en affinant le taux de convergence et en affaiblissant les hypothèses sur l'ensemble des minimiseurs. On montre aussi que l'inertie de Nesterov, sous une paramétrisation stochastique appropriée, atteint le meilleur taux connu parmi les méthodes de premier ordre pour approximer un point critique de fonctions à gradient et Hessienne Lipschitz, sans recourir aux mécanismes de réinitialisation d'inertie nécessaires dans les travaux existants.
**Mots-clés :** Optimisation, Accélération, Non Convexe, Stochastique

---

**A Study of Nesterov Momentum Algorithm for Accelerated Nonconvex and Stochastic Optimization**

**Abstract:** Optimization algorithms play an important role in many applications. One of the key criteria for a good algorithm is fast convergence, which can be challenging to achieve due to the large scale of modern problems. Among the mechanisms enabling efficient algorithms, an important one is momentum. In particular, it is widely observed in a large range of learning applications that adding such a mechanism to gradient descent accelerates its convergence. While this acceleration phenomenon is well understood in the setting of smooth convex minimization using full gradients, many machine learning problems are non-convex and rely on stochastic mini-batch estimators of the gradient. This thesis provides a theoretical study of the acceleration induced by Nesterov's momentum mechanisms in such settings. Our first contribution is motivated by the observation that in several natural non-convex or stochastic settings, many existing analyses do not establish any acceleration property with momentum. We identify geometric conditions under which momentum provably yields acceleration. For non-convex functions satisfying a Polyak–Łojasiewicz condition, we show that the possibility of acceleration depends on the angle between the descent direction and the direction toward minimizers. In the convex stochastic finite-sum setting, if it verifies the interpolation assumption (typical of over-parameterized models), we show that acceleration is governed by the average correlation between the gradients composing the sum. Our second contribution builds on the recently proposed continuized Nesterov method. It can be understood as a Nesterov algorithm, which can be analyzed using continuous-time stochastic Lyapunov approaches. We extend this framework and use it to derive new results in non-convex optimization. In particular, we prove that Nesterov's Momentum, under a suitable stochastic parametrization, achieves the best known rate among gradient-based methods for approximating a critical point of functions with Lipschitz-continuous gradients and Hessians, without the momentum-reset mechanisms needed by existing works. We also improve existing convergence guarantees for strongly quasar-convex optimization by refining the rate and weakening assumptions on the set of minimizers.
**Keywords:** Optimization, Acceleration, Non Convex, Stochastic

---

**Unité de recherche**
UMR 5251 Université, 33000 Bordeaux, France.

# Remerciements

Je remercie mes encadrants, Aude, Charles et Jean-François, pour leur disponibilité, leur confiance et pour m'avoir guidé dans mes débuts en recherche. Leurs conseils, leurs remarques et leur rigueur ont été appréciables. Je remercie mes rapporteurs, Aris Daniilidis et Adrien Taylor, ainsi que mes examinateurs, Antonin Chambolle, Jalal Fadili et Édouard Pauwels, d'avoir accepté d'examiner ma thèse. Je remercie particulièrement mes rapporteurs pour leurs retours, qui ont aidé à améliorer ce document.

Je remercie les personnes avec qui j'ai eu la chance de travailler et d'écrire des articles, qu'elles viennent de l'IMB — Émeric, Marien et Nicolas —, d'ailleurs en France — Irène et Lorick — ou d'encore ailleurs — Deliang et Yu. Je remercie les équipes administratives et informatiques pour leur disponibilité, particulièrement Nadia Mechalakh, que j'ai souvent embêtée pour qu'elle m'aide à organiser mes missions.

Je tiens à remercier Benjamin Dupuy, mon professeur de maths de terminale au lycée Max Linder. Il m'a fait aimer *réellement* les mathématiques et a joué un rôle important dans la naissance de ma vocation pour la recherche. Je veux aussi remercier Christèle Etchegaray de m'avoir permis de faire de la recherche pour la première fois en me prenant en stage de M1, et plus généralement l'ensemble des enseignants que j'ai pu croiser et qui m'ont aidé et inspiré.

Je remercie tous les camarades doctorants et autres avec qui j'ai passé tout ou partie de ces trois années, en particulier les personnes qui ont su apprécier mon humour subtil.

Enfin, je remercie ma famille, en particulier Eve pour son soutien considérable, et peut-être aussi pour réussir à me supporter. Finalement, je souhaite au petit Trajan de trouver une meilleure manière d'occuper son temps que d'améliorer d'un facteur constant les bornes de convergence des fonctions quasar-convexes.

11000011 10000000 00100000
01101101 01101111 01101110
00100000 01110000 11000011
10101000 01110010 01100101

# Résumé Substanciel en Français

Beaucoup de problèmes peuvent se formuler de la façon suivante

$$\arg\min_{x\in\mathbb{R}^d} f(x), \text{ où } f:\mathbb{R}^d\to\mathbb{R} \text{ est continûment différentiable,}$$

pour un $d\in\mathbb{N}$. C'est le cas de l'apprentissage de certains modèles paramétrés, où leur entraînement consiste souvent à définir une fonction $f$ adaptée, dont la minimisation revient à choisir des paramètres qui optimisent le modèle sur un ensemble de données d'exemples. Plusieurs difficultés peuvent alors arriver. D'abord si le modèle est complexe, par exemple si c'est un réseau de neurones (Rosenblatt, 1958; Rumelhart et al., 1986; LeCun et al., 1998), la fonction $f$ peut présenter une géométrie difficile, ce qui complique son optimisation. Ensuite, si un grand nombre de paramètres permet d'accroître l'expressivité du modèle, il accroît également la dimension $d$ du problème d'optimisation associé. Ce phénomène est particulièrement marqué pour les grands modèles de langage. Par exemple, le modèle GPT-3 comporte 175 milliards de paramètres (Brown et al., 2020). Enfin, utiliser une base de données d'exemples de très grande taille complique encore plus le processus d'optimisation. Des problèmes d'une telle échelle peuvent être particulièrement difficiles à optimiser.

**Algorithmes** Des méthodes d'optimisation efficaces sont une composante fondamentale dans le processus d'apprentissage de tels modèles. Aujourd'hui, les algorithmes d'optimisation les plus utilisés, comme Adam (Kingma and Ba, 2015), AdamW (Loshchilov and Hutter, 2017) ou encore Muon (Jordan et al., 2024), incorporent plusieurs mécanismes. L'un de leurs points communs est d'exploiter des évaluations du gradient $\nabla f$ de la fonction à minimiser. L'un des exemples les plus simples d'un tel algorithme est la descente de gradient, qui, à partir d'une initialisation $\tilde{x}_0$ et d'une suite de pas $\{\gamma_k\}_{k\in\mathbb{N}}$, est définie par la récurrence

$$\tilde{x}_{k+1} = \tilde{x}_k - \gamma_k \nabla f(\tilde{x}_k), \qquad \gamma_k \in \mathbb{R}_+. \tag{GD}$$

Un autre principe fondamental partagé par ces algorithmes est l'utilisation d'un mécanisme d'*inertie*. Dans cette thèse, nous nous intéressons plus spécifiquement à un algorithme qui peut être compris comme une descente de gradient à laquelle un tel mécanisme d'inertie est ajouté. Plus précisément, nous considérons l'algorithme inertiel de Nesterov (Nesterov, 1983), qui, pour une initialisation $\tilde{x}_0, \tilde{z}_0 \in \mathbb{R}^d$ et des suites de paramètres $\{\alpha_k\}_{k\in\mathbb{N}}$, $\{\beta_k\}_{k\in\mathbb{N}}$, $\{\gamma_k\}_{k\in\mathbb{N}}$ et $\{\gamma'_k\}_{k\in\mathbb{N}}$, s'écrit

$$\begin{cases} \tilde{y}_k &= \tilde{x}_k + \alpha_k(\tilde{z}_k - \tilde{x}_k), \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma_k \nabla f(\tilde{y}_k), \\ \tilde{z}_{k+1} &= \tilde{z}_k + \beta_k(\tilde{y}_k - \tilde{z}_k) - \gamma'_k \nabla f(\tilde{y}_k). \end{cases} \tag{NM}$$

La suite $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ de cet algorithme se réduit à la descente de gradient lorsque le paramètre $\alpha_k$ est choisi identiquement nul. Lorsque $\alpha_k > 0$, $\tilde{x}_{k+1}$ peut être compris comme une descente de gradient appliquée à un point $\tilde{y}_k$ qui, par rapport à $\tilde{x}_k$, est tiré vers un autre point $\tilde{z}_k$, créant une inertie. Il est couramment observé empiriquement que l'ajout d'un mécanisme d'inertie accélère la convergence de la descente de gradient (Sutskever et al., 2013). Dans cette thèse, nous étudions la possibilité d'établir théoriquement une telle propriété d'accélération.

**Structure de $f$ et Accélération Théorique** On suit l'approche classique qui consiste à ne pas supposer que la fonction $f$ provient d'un problème spécifique, mais plutôt qu'elle vérifie une certaine structure, généralement exprimée sous la forme d'une inégalité. Dans certains cadres classiques, le phénomène d'accélération est déjà bien établi. C'est le cas des fonctions convexes à gradient Lipschitz, pour lesquelles

l'algorithme (NM), convenablement paramétré, permet d'assurer $f(\tilde{x}_k) - \min_{x\in\mathbb{R}^d} f(x) = \mathcal{O}\left(\frac{1}{k^2}\right)$ (Nesterov, 1983). Cette borne caractérise l'efficacité de l'algorithme. Un résultat fondamental montre que lorsqu'il s'agit de minimiser ces fonctions convexes à gradient Lipschitz, la dépendance en $k$ dans la borne ne peut pas être améliorée par un autre algorithme n'ayant accès qu'au gradient et à la fonction (Nemirovskij and Yudin, 1983). De plus, la descente de gradient n'atteint pas une telle vitesse (Drori and Teboulle, 2014). Cependant, un tel cadre ne correspond pas à de nombreuses applications pratiques. Dans le cas de l'entraînement de modèles paramétrés, la complexité du modèle peut conduire à un problème de minimisation où $f$ n'est pas convexe. Or il existe des structures de non-convexité pour lesquelles aucune propriété d'accélération due à l'inertie ne peut être établie (Carmon et al., 2020). Aussi, la présence d'une base de données de très grande taille rend le calcul du gradient coûteux, car intuitivement, pour chaque donnée il faut calculer un gradient. Cela pousse à sous-échantillonner les données, et à utiliser seulement une moyennisation des gradients de ce sous-ensemble; c'est ce qui est souvent fait en pratique. Une telle moyenne de gradient est un estimateur aléatoire du gradient. Même pour des fonctions $f$ à la géométrie favorable (convexité), certains types d'aléa peuvent entraîner une impossibilité d'accélérer significativement avec de l'inertie (Jain et al., 2018). Ces observations peuvent sembler contredire les bénéfices empiriques observés en pratique. Elles suggèrent en réalité que les fonctions effectivement minimisées possèdent une structure supplémentaire.

**Notre Objectif** Dans cette thèse, nous nous intéressons donc à des cadres non convexes et/ou stochastiques dans lesquels il est possible d'établir une accélération théorique de (NM) par rapport à (GD). Selon les situations, notre objectif sera soit d'obtenir de nouveaux résultats de convergence accélérée, soit de fournir une interprétation de bornes accélérées déjà connues.

# Résumé des Contributions de la Thèse

On donne un résumé des différentes contributions.

## Accélération et Stochasticité: Impact de la corrélation entre gradients

**Cadre** Dans le Chapitre 2, nous considérons un problème convexe possédant une structure de somme finie, $f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$. On considère un estimateur du gradient de la forme $\tilde{\nabla}_K(x) := \frac{1}{K}\sum_{i\in B} \nabla f_i(x)$, où $B$ est un sous-ensemble d'indices de taille $K$ échantillonné uniformément sur $\{B \subset \{1,\dots,N\} \mid \mathrm{Card}(B) = K\}$. Notre point de départ est que, pour des fonctions (fortement) convexes, une hypothèse sur le bruit de l'estimateur, appelée *strong growth condition*, permet d'obtenir les mêmes bornes que dans le cas déterministe, à un facteur de dégradation près, plus ou moins impactant selon l'intensité du bruit. Ce facteur joue un rôle important, car il peut déterminer si une accélération est effectivement obtenue ou non. Cependant, il existe très peu de situations dans lesquelles sa valeur est explicitement connue.

**Contribution** Nous montrons alors que cette valeur est fortement liée à la corrélation moyenne entre les gradients composant la somme,

$$\sum_{1\le i<j\le N} \langle \nabla f_i(x), \nabla f_j(x)\rangle.$$

Plus cette corrélation moyenne est élevée, plus le facteur de dégradation sera favorable, et ainsi, plus importante est la propriété d'accélération avec inertie. C'est une caractérisation mathématique d'un fait intuitif: si les gradients sont très corrélés, on peut dire qu'ils sont "d'accord" sur la direction à suivre, auquel cas on peut amplifier leur direction avec de l'inertie. S'ils ne sont pas corrélés, l'information d'un seul gradient peut fortement diverger par rapport à l'information moyenne. Dans un tel cas, amplifier sa direction est dangereux. Empiriquement, on observe alors qu'en créant un problème de régression linéaire où la corrélation moyenne est proche de zéro, l'inertie n'accélère effectivement pas. En revanche, dans le cas de l'optimisation d'un réseau de neurones, on observe que la corrélation moyenne est élevée, ce qui coïncide avec le bénéfice observé de l'inertie dans ce cas. Cette contribution est issue de l'article suivant.

> J. Hermant, M. Renaud, J-F. Aujol, C. Dossal, and A. Rondepierre.
> **Gradient correlation is a key ingredient to accelerate SGD with momentum**.
> In *The Thirteenth International Conference on Learning Representations*, 2025

## Accélération sous hypothèse de Forte Quasar Convexité

**Cadre** Un autre cadre dans lequel (NM) permet d'obtenir une accélération est celui des fonctions $\mu$-fortement convexes à gradient $L$-Lipschitz. Dans ce cas, (NM) assure $f(\tilde{x}_k) - \min_{x\in\mathbb{R}^d} f(x) = \mathcal{O}\big(\big(1-\sqrt{\frac{\mu}{L}}\big)^k\big)$, contre $f(\tilde{x}_k) - \min_{x\in\mathbb{R}^d} f(x) = \mathcal{O}\big(\big(1-\frac{\mu}{L}\big)^k\big)$ avec (GD). Le *conditionnement* $L/\mu$ vérifie toujours $L/\mu \geq 1$, et l'on a même typiquement $L/\mu \gg 1$. Dans ce régime, la borne obtenue avec (NM) est significativement meilleure. Une relaxation classique de la forte convexité est donnée par les fonctions vérifiant l'inégalité de Polyak-Łojasiewicz (PL) (Polyak, 1963; Łojasiewicz, 1963). Ces fonctions non convexes apparaissent dans de nombreuses applications. Cependant, pour leur minimisation, on ne peut en général pas établir d'amélioration significative due à l'inertie (Yue et al., 2023). Une structure intermédiaire entre la forte convexité et PL est celle de la forte quasar convexité. Elle est définie pour un minimiseur $x^* \in \mathbb{R}^d$ et des paramètres $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$ par l'inégalité

$$x \in \mathbb{R}^d,\ \min_{u\in\mathbb{R}^d} f(u) \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2. \tag{1}$$

Pour ces fonctions, sous hypothèse de $L$-Lipschizité du gradient, on a avec la descente de gradient la borne $f(\tilde{x}_k) - f^* = \mathcal{O}\left((1-\tau\frac{\mu}{L})^k\right)$, où avec l'inertie de Nesterov on peut avoir $f(\tilde{x}_k) - f^* = \mathcal{O}\left((1-\tau\sqrt{\frac{\mu}{L}})^k\right)$ (Hinder et al., 2020). Pour ces fonctions, (NM) permet d'obtenir une dépendance au conditionnement en racine carrée. Cette classe semble donc permettre de relâcher la forte convexité tout en conservant ses propriétés d'accélération.

**Contributions** Nous apportons plusieurs contributions autour de la borne $f(\tilde{x}_k) - f^* = \mathcal{O}\left((1-\tau\sqrt{\frac{\mu}{L}})^k\right)$ obtenue avec (NM). Les résultats existants reposaient soit sur des sous-routines de recherche de paramètres nécessitant des évaluations du gradient (Hinder et al., 2020), soit sur une paramétrisation aléatoire (Wang and Wibisono, 2023). Nous montrons qu'il est possible d'atteindre cette borne sans sous-routines, à l'aide d'une paramétrisation déterministe plus classique, au prix d'hypothèses supplémentaires sur la courbure. Nous améliorons également légèrement cette borne, en obtenant un taux de la forme $\mathcal{O}\left((1-c\tau\sqrt{\frac{\mu}{L}})^k\right)$, où $c > 1$. Dans le cas $\tau = 1$, nous relâchons l'hypothèse d'unicité du minimiseur imposée par (1). Alors qu'un résultat précédent obtenait une telle extension sous l'hypothèse que $\arg\min f$ est un espace affine, nous gagnons en généralité en supposant à la place que cet ensemble est convexe. Enfin, nous mettons en évidence certaines limites de l'hypothèse (1). En particulier, nous montrons qu'elle ne bénéficie pas des mêmes propriétés d'accélération que la forte convexité, contrairement à ce que la forme des bornes pourrait suggérer. Nous proposons alors une alternative, consistant à étudier les fonctions PL vérifiant une condition d'angle :

$$\forall x \in \mathbb{R}^d, \quad 1 \geqslant \frac{\langle \nabla f(x), x - x^*\rangle}{\|\nabla f(x)\|\|x - x^*\|} \geqslant a > 0.$$

La valeur de $a$ caractérise la corrélation entre la direction de descente $-\nabla f$ et la direction qui pointe vers un minimiseur $x^*$. Une valeur de $a$ proche de 1 indique que le gradient nous emmène dans une "bonne" direction. Il est intuitif qu'une telle propriété soit favorable à un algorithme inertiel, car cela traduit que l'inertie amplifie des déplacements qui emmènent dans la bonne direction. On observe alors que l'accélération pourra alors être caractérisée par la valeur de $a$. Les contributions mentionnées se trouvent dans les articles suivants:

J. Hermant, J-F. Aujol, C. Dossal, and A. Rondepierre.
**Study of the behaviour of Nesterov Accelerated Gradient in a non convex setting: the strongly quasar convex case**. In *Optimization*, 2026

J. Hermant, J-F. Aujol, C. Dossal, L. Huang and A. Rondepierre.
**Continuized Nesterov Acceleration for Non-Convex Optimization**. *arXiv*, 2025

J. Hermant.
**Acceleration for Polyak-Łojasiewicz Functions with a Gradient Aiming Condition**. *arXiv*, 2026

## Utilisation du Nesterov Continuisé pour l'accélération Fortement Quasar Convex et sous hypothèse de Lipschizité du gradient et de la Hessienne

**Cadre** Dans (Even et al., 2021) est introduit le processus d'inertie de Nesterov continuisé:

$$\begin{cases} dx_t &= \eta_t(z_t - x_t)dt - \gamma_{t^-}\nabla f(x_{t^-})N(dt) \\ dz_t &= \eta'_t(x_t - z_t)dt - \gamma'_{t^-}\nabla f(x_{t^-})N(dt) \end{cases} \tag{2}$$

où $N(dt) = \sum_{k\geq 0}\delta_{T_k}(dt)$ est une mesure de Poisson. Si on remplaçait les $N(dt)$ par des $dt$, le processus deviendrait une équation différentielle ordinaire (EDO) limite de l'algorithme (NM) que l'on obtient en prenant des pas tendant vers zéro (Su et al., 2016; Kim and Yang, 2023). Contrairement à une telle EDO, qui n'est qu'approximable, l'aspect fondamental du processus (2) est que, bien qu'il soit formulé en temps continu, il peut être évalué exactement au moyen d'un algorithme de Nesterov à paramètres aléatoires. Pour une fonction $f$ (fortement) convexe à gradient $L$-Lipschitz, l'algorithme associé à ce processus retrouve, en espérance, les bornes de convergence de l'algorithme de Nesterov classique (Even et al., 2021). Un résultat analogue a été obtenu pour les fonctions fortement quasar convexes (Wang and Wibisono, 2023). De manière intéressante, ce dernier résultat ne nécessite ni sous-routine de recherche de paramètres, ni restriction supplémentaire sur la courbure, contrairement aux approches discutées précédemment.

**Contributions** Motivés par l'observation précédente, ainsi que par les limites de l'approche classique consistant à considérer des limites EDO d'algorithmes, nous généralisons les outils nécessaires à l'analyse de la convergence du processus continuisé. Nous l'appliquons ensuite pour obtenir de nouveaux résultats de convergence. En fait, certaines contributions mentionnées dans le cadre des fonctions fortement quasar convexes ont été obtenues à l'aide de ce processus. Il s'agit de la légère amélioration de la borne de convergence et la relaxation d'hypothèse sur l'ensemble des minimiseurs.

Comme autre application, nous nous intéressons aux fonctions à gradient et Hessienne Lipschitz. Dans ce cadre, des algorithmes fondés sur l'inertie de Nesterov atteignent des bornes accélérées par rapport à la descente de gradient. Cependant, ces algorithmes nécessitent un contrôle renforcé de l'inertie, au sens où celle-ci doit parfois être réinitialisée. En utilisant le processus continuisé, on a montré que quitte à randomiser des paramètres dans (NM), on peut se passer de ces mécanismes de contrôle. Les contributions mentionnées se trouvent dans les articles suivants. D'abord celui-ci **déjà mentionné** dans la section "Accélération sous hypothèse de Forte-Quasar Convexité".

> J. Hermant, J-F. Aujol, C. Dossal, L. Huang and A. Rondepierre.
> **Continuized Nesterov Acceleration for Non-Convex Optimization**. *arXiv*, 2025 (déjà mentionné plus haut)

Puis, pour le résultat dans le cas de fonctions à gradient et Hessienne Lipschitz.

> J. Hermant, J-F. Aujol, C. Dossal, L. Huang, A. Rondepierre and I. Waldspurger.
> **Continuized Nesterov Momentum Achieves the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity in Smooth Nonconvex Optimization**. *arXiv*, 2026

# Contents


**1 Introduction** **14**
- 1.1 Effect of Non-Convexity and Stochastic Gradients on Acceleration with Nesterov's Momentum 17
  - 1.1.1 Relaxation of Convexity: Smoothness Only . . . 17
  - 1.1.2 Relaxation of Strong Convexity . . . 18
  - 1.1.3 Relaxing Access to Exact Gradient: retaining acceleration with Strong Growth Condition 22
- 1.2 A Convergence Analysis Method and its Limitation: limit-ODE of Algorithms . . . 24
- 1.3 Contributions . . . 27
- 1.4 Appendices of Introduction . . . 34
  - 1.4.1 Convexity and Quasar Convexity . . . 34
  - 1.4.2 Comments on Algorithms . . . 38

**2 Nesterov Momentum with Batch Estimator: Effect of Gradient Correlation on Acceleration** **40**
- 2.1 Introduction . . . 40
- 2.2 Background . . . 41
- 2.3 Convergence speed of SNM and comparison with SGD . . . 42
  - 2.3.1 Convergence results for SGD . . . 42
  - 2.3.2 Convergence in expectation for SNM . . . 43
  - 2.3.3 Almost sure convergence for SNM . . . 44
- 2.4 Characterizing convergence with strong growth condition and gradient correlation . . . 45
  - 2.4.1 Average positive correlation condition . . . 45
  - 2.4.2 RACOGA: relaxing the PosCorr condition . . . 45
  - 2.4.3 The strong growth condition with batch size 1 determines how the performance scales with batch size . . . 46
- 2.5 Numerical experiments . . . 47
  - 2.5.1 Linear regression . . . 47
  - 2.5.2 Neural networks . . . 48
- 2.6 Conclusion . . . 49
- 2.A Additional experiments and details . . . 49
  - 2.A.1 Linear regression . . . 50
  - 2.A.2 Neural networks . . . 51
- 2.B Background : convex optimization and the Nesterov algorithm . . . 55
  - 2.B.1 Performance of GD and NM . . . 55
  - 2.B.2 The different forms of Nesterov algorithm . . . 56
- 2.C Related works . . . 58
  - 2.C.1 Gradient correlation conditions . . . 58
  - 2.C.2 Linear regression acceleration results . . . 59
  - 2.C.3 SGC related convergence results . . . 59
- 2.D Some examples with critical strong growth condition constant . . . 61
  - 2.D.1 Large $\rho_1$ with linear regression . . . 61
  - 2.D.2 Non convex functions such that $\rho_1$ does not exist . . . 61
- 2.E Convergence of SGD . . . 62
  - 2.E.1 Proofs of Theorem 2.3.1 and Theorem 2.3.2 . . . 62

2.E.2 Convergence of SGD with strong growth condition . . . 63
2.F Convergence of SNM without strong growth condition . . . 64
2.F.1 Proof of Theorem 2.F.1, convex case . . . 65
2.F.2 Proof of Theorem 2.F.1, strongly convex case . . . 67
2.F.3 Additional Lemma . . . 69
2.G Convergence of SNM with strong growth condition . . . 70
2.G.1 Convex-Expectation . . . 71
2.G.2 Strongly convex - Expectation . . . 72
2.G.3 Convex - almost sure . . . 74
2.G.4 Strongly convex - almost sure . . . 77
2.H Proofs of Section 2.4 . . . 78
2.H.1 Proof of Proposition 2.4.1 . . . 78
2.H.2 Proof of Proposition 2.4.2 . . . 80
2.H.3 Proof of Lemma 2.4.1 . . . 81
2.H.4 Proof of Theorem 2.4.1 . . . 82
2.H.5 Bound on RACOGA . . . 83
2.I RACOGA values in linear regression . . . 84
2.I.1 Two functions in two dimensions . . . 84
2.I.2 RACOGA is high over most of the space . . . 84
2.I.3 The curvature problem: first-order algorithms are attracted by low RACOGA value areas . . . 86
2.I.4 Proof of Proposition 2.I.1 . . . 88

**3 Study of Nesterov Momentum for Strongly Quasar Convex Optimization and the Continuous-Discrete Rupture 89**
3.1 Introduction . . . 89
3.2 Preliminaries . . . 90
3.2.1 Strong convexity and the question of acceleration . . . 91
3.2.2 A relaxation of strong convexity: strong quasar convexity . . . 92
3.3 Acceleration with curvature assumption for strongly quasar-convex functions . . . 94
3.3.1 The gradient descent convergence . . . 94
3.3.2 The Nesterov Momentum convergence . . . 95
3.3.3 Composite non differentiable case . . . 97
3.4 Continuous analysis . . . 98
3.4.1 High-resolution ordinary differential equation . . . 98
3.4.2 The continuous/discrete rupture . . . 100
3.5 Geometrical considerations . . . 101
3.6 Numerical experiments details . . . 102
3.6.1 Algorithm performance . . . 102
3.6.2 Explicit expression of the function displayed in figure 3.1 . . . 104
3.A Some properties on strongly quasar-convex functions . . . 104
3.A.1 Strong convexity on average on segments joining the minimizer . . . 105
3.A.2 Synthetic strongly quasar-convex example proof . . . 105
3.A.3 Local strong convexity around the minimizer for $C^2$ functions . . . 105
3.A.4 No local convexity for non $C^2$ functions: a non-convex pathological construction . . . 106
3.B Proofs of section 3 . . . 109
3.B.1 Differentiable strongly quasar-convex . . . 109
3.C Continuous analysis through High-resolution ODEs . . . 116

**4 Continuized Nesterov Acceleration for Non-Convex Optimization: Application to Strongly Quasar Convex Functions 120**
4.1 Introduction . . . 120
4.1.1 A Promising Alternative: Continuized Nesterov . . . 120
4.1.2 Limits of the Framework and Contributions . . . 121
4.2 Assumptions and Notations . . . 122

4.2.1 Geometrical Assumptions . . . 122
4.2.2 Stochastic Gradients . . . 122
4.3 How to Use the Continuized Nesterov Framework . . . 123
4.3.1 Lyapunov Analysis with Stochastic Calculus . . . 124
4.3.2 Almost Automatic Discretization . . . 124
4.3.3 Application to Smooth, Polyak-Łojasiewicz Functions . . . 125
4.4 Improved Factor in Convergence Rate for Strongly quasar convex Functions and Better Trajectory-Wise Guarantees . . . 128
4.4.1 Improved Factor in Convergence Rate . . . 128
4.4.2 Convergence Results with High Probability and Almost Sure . . . 129
4.5 Non-Smooth Lyapunov Functions for Relaxing the Uniqueness Assumption . . . 131
4.5.1 A Technical Challenge: $\pi(\cdot)$ is not $C^1$ . . . 131
4.5.2 Convergence Result . . . 133
4.6 Convergence Proofs . . . 134
4.6.1 Proof of Theorem 4.4.2 . . . 134
4.6.2 Proof of Theorem 4.5.1 . . . 137
4.A Proof of Proposition 4.3.2. . . . 143
4.B Convergence Results Trajectory-Wise . . . 144
4.B.1 Proof of Proposition 4.4.1 . . . 144
4.B.2 Proof of Proposition 4.4.2 . . . 144
4.C Technical Lemmas . . . 145
4.C.1 Proof of Lemma 4.2.1 . . . 145
4.C.2 Proof of Lemma 4.6.3 . . . 145
4.D Background . . . 146

**5 Reconsideration of Acceleration for Strongly-Quasar Convex functions, Existing Flaws and Alternative 147**
5.1 Introduction . . . 147
5.1.1 Contributions . . . 147
5.1.2 Related Works on Aiming Conditions . . . 148
5.2 Background . . . 148
5.2.1 Algorithms and Equations . . . 148
5.2.2 Classical setting for acceleration . . . 149
5.2.3 Relaxed setting . . . 150
5.3 Acceleration under Strong Quasar Convexity: Hidden Pitfalls . . . 150
5.3.1 Problems of Strong Quasar Convexity . . . 151
5.3.2 Discussion . . . 152
5.4 Aiming condition as an Alternative Characterization of Convergence . . . 152
5.4.1 Link between Aiming Condition and Strong Quasar Convexity . . . 153
5.5 Acceleration for PL Functions with an Aiming Condition . . . 154
5.5.1 $\mathrm{AC}^a$ Gives Weight to the Local Information at $x^*$ . . . 155
5.5.2 Acceleration using Momentum with Large Enough Aiming Condition Constant . . . 156
5.5.3 Decoupling information: more interpretability, but may create sub-optimality. . . . 157
5.6 Relaxed Setting: Aiming Condition on Average . . . 157
5.7 Numerical Experiments . . . 158
5.7.1 An Example with Negative Aiming Condition . . . 158
5.7.2 Heuristic Evaluation of the Acceleration Bound . . . 160
5.A Related Works and Background Details . . . 163
5.A.1 Details for Table 5.1 . . . 163
5.A.2 Choice of Parameters for $\mathrm{SQC}_\tau^\mu$. . . . 164
5.B More Details on Figures and Numerical Experiments . . . 164
5.B.1 Details about Figure 5.1 and Figure 5.2 . . . 165
5.B.2 Details about Figure 5.3 . . . 165
5.B.3 Proof of Proposition 5.3.1 . . . 167
5.B.4 Supplementary Figures . . . 168

5.C Technical Lemmas . . . 168
5.D Proofs of Section 5.5 . . . 172
5.D.1 Proof of Theorem 5.5.1 . . . 172
5.D.2 Proof of Theorem 5.5.2 . . . 173
5.D.3 Proof of Theorem 5.5.3 . . . 175
5.E Trajectory-adapted convergence . . . 178
5.E.1 Proof of Theorem 5.6.1 . . . 178
5.E.2 Discrete case . . . 180
5.E.3 Possible Extensions: Adapt to other Conditions and Exact Rates . . . 182
5.F Continuized Toolbox . . . 184
5.G Proof of Proposition 5.4.1 . . . 185
5.G.1 Proof of Proposition 5.4.2 . . . 185

**6 Application of the Continuized Nesterov Momentum Method to Functions with Lipschitz Gradient and Hessian: achieving the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity without Momentum-Control Mechanisms 187**
6.1 Introduction . . . 187
6.1.1 Related works . . . 188
6.2 Background . . . 189
6.2.1 Continuized Nesterov Algorithm . . . 189
6.3 Convergence Result without Lipschitz Hessian Assumption . . . 190
6.4 Main Result under Lipschitz Hessian Assumption . . . 191
6.4.1 Convergence Result . . . 191
6.5 Overview of the proof of Theorem 6.4.1 . . . 193
6.5.1 Part I - Lyapunov Controls from Section 6.3 . . . 194
6.5.2 Part II - Using the Hessian Lipschitz Property to Obtain a Three Terms Upper-Bound 195
6.5.3 Part III: Control of Term 1 . . . 196
6.5.4 Part IV: Control of Term 2 and Term 3 . . . 197
6.5.5 Part V: Tuning $\alpha$ . . . 199
6.5.6 Part VI: Control of the double Poisson integral factor . . . 200
6.5.7 Part VII: Final Results . . . 201
6.A Complementary details to Section 6.1 . . . 202
6.A.1 Second-order stationary points . . . 202
6.A.2 Detailed formulations of existing security check and alternative mechanisms . . . 203
6.B Proof of Lemma 6.3.1 and Proposition 6.3.1 . . . 203
6.C Proof of the Lemmas stated in Section 6.5 . . . 207
6.C.1 Proof of Lemma 6.5.1 . . . 207
6.C.2 Proof of Lemma 6.5.2 . . . 210
6.C.3 Proof of Lemma 6.5.3 . . . 211
6.C.4 Proof of Lemma 6.5.4 . . . 212
6.C.5 Proof of Lemma 6.5.5 . . . 213
6.C.6 Proof of Lemma 6.5.6 . . . 217
6.C.7 Proof of Lemma 6.5.7 . . . 219
6.C.8 Proofs of Section 6.5.5 . . . 220
6.C.9 Proof of Lemma 6.5.8 . . . 222
6.D Proof of Theorem 6.5.1 . . . 223
6.D.1 Upper-bounding the sums . . . 224
6.D.2 lower-bounding expectations . . . 230
6.D.3 Combining lower and upper-bounds . . . 232
6.E Additional technical Lemmas used to prove Theorem 6.4.1 . . . 234

**Conclusion and Perspectives 239**

# Notations

- By default, $f : \mathbb{R}^d \to \mathbb{R}$, $d \in \mathbb{N}$.
- $\operatorname{argmin} f := \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$.
- $f^* := f(x^*)$, for some $x^* \in \operatorname{argmin} f$. If clear from context, we may also $\min f := f(x^*)$, for some $x^* \in \operatorname{argmin} f$.
- By default, $x^* \in \operatorname{argmin} f$.
- For a function $E : \mathbb{R} \to \mathbb{R}$, we denote $\dot{E}_t := \frac{\partial E}{\partial t}$.
- $\|\cdot\|$ is the Euclidean norm for vectors.
- $\|\cdot\|_p$ is the $\ell^p$ norm for vectors. It reduces to the Euclidean norm for $p = 2$, so with our notations $\|\cdot\| = \|\cdot\|_2$.
- $\|\cdot\|_{\text{spec}}$ is the spectral norm for matrices.
- $\mathcal{E}(1)$ is the exponential distribution with parameter 1.
- $\Gamma(k, 1)$ is the Gamma distribution with shape parameter $k$ and rate 1.
- $\mathcal{P}(\lambda)$ is the Poisson distribution with parameter $\lambda > 0$.

# Chapter 1

# Introduction

Optimization studies the problem of searching for an optimal element from a set of admissible alternatives. In many situations, we can model this problem via a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the set of such admissible strategies. This function assigns a cost value to each strategy: the lower the cost value, the better the strategy. We consider the case where $\mathcal{X} = \mathbb{R}^d$, which is a special case of continuous optimization, where $\mathcal{X}$ is an uncountable set. Given this model, the problem reduces to solving

$$\arg\min_{x\in\mathbb{R}^d} f(x), \text{ where } f : \mathbb{R}^d \to \mathbb{R} \text{ is continuously differentiable.} \tag{1.1}$$

The assumption that $f$ is continuously differentiable does not necessarily hold for some applications, but it already encompasses several interesting problems. An important example of an application that is modeled by (1.1) is supervised learning with some parameterized models.

**Parameterized Models** Let $\mathcal{U}$ and $\mathcal{V}$ be two sets related by an unknown map $g : \mathcal{U} \to \mathcal{V}$. For instance, $\mathcal{U}$ may contain vectors describing a train journey, such as the schedule or the number of days before departure, while $\mathcal{V}$ may represent ticket prices. Given a dataset of examples $\{(u_i, v_i)\}_{i=1}^N \subset \mathcal{U} \times \mathcal{V}$, the goal is to *learn* from them to approximate $g$, by a model $\tilde{g}$ that satisfies

$$\tilde{g}(u_i) \approx v_i, \qquad i = 1, \dots, N.$$

A standard approach is to choose $\tilde{g}$ from a parameterized family of functions $f(\cdot, \theta)$, with $\theta \in \mathbb{R}^d$, by searching for a parameter $\theta$ that makes the model fit the data. Wether the fit is good can be measured by $\ell(f(u_i, \theta) - v_i)$, where $\ell : \mathcal{V} \to \mathbb{R}$ is a *loss function*. If $\mathcal{V} := \mathbb{R}$ and $\ell$ is the one-dimensional squared function, averaging over the data leads to the standard least square loss

$$\operatorname{argmin}_{\theta\in\mathbb{R}^d} F(\theta), \quad F(\theta) := \frac{1}{N}\sum_{i=1}^{N}\big(f(u_i, \theta) - v_i\big)^2. \tag{1.2}$$

Thus, the learning problem reduces to minimizing a loss over the admissible parameter set, here taken to be the unconstrained space $\mathbb{R}^d$, although constraints may also be imposed. In many cases, we do not know how to solve (1.2) analytically. Instead, we use algorithms to approximate a solution.

**Algorithms I: Gradient Descent.** Perhaps one of the simplest algorithms to solve (1.1) is gradient descent, whose idea dates back at least to (Cauchy et al., 1847). After choosing $\tilde{x}_0 \in \mathbb{R}^d$ as our initialization, gradient descent is defined iteratively through the following procedure: given the iterate $\tilde{x}_k$, $k \in \mathbb{N}$, we define

$$\tilde{x}_{k+1} = \tilde{x}_k - \gamma_k \nabla f(\tilde{x}_k), \ \gamma_k \in \mathbb{R}_+. \tag{GD}$$

This algorithm exploits gradient information to compute its iterates. One may consider higher-order information, such as using the Hessian matrix (*e.g.* Newton methods). However, in large-scale problems, *i.e.* when $d \gg 1$, computing such objects may become computationally intractable, while computing the gradient may remain possible. The convergence of gradient descent has been extensively studied. Under the following structural assumption of convexity and $L$-smoothness

$$\forall x, y \in \mathbb{R}^d, \quad f(x) + \langle \nabla f(x), y - x\rangle \le f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2, \tag{1.3}$$

with a proper choice of sequence $\{\gamma_k\}_{k\in\mathbb{N}}$, e.g. constant equal to $1/L$, one can ensure $f(\tilde{x}_k) - \min_{x\in\mathbb{R}^d} f(x) = \mathcal{O}(1/k)$[1] (Nesterov, 2013). Such a result ensures that $f(\tilde{x}_k)$ converges to $f^*$, while giving an upper bound on the speed at which it does so. Gradient descent is a special instance of **first-order algorithms**, namely algorithms that can access evaluations of $f$ and $\nabla f$. It is commonly observed that among first-order algorithms, gradient descent stands as a rather simple yet slow algorithm. The most popular algorithms in deep-learning (Adam (Kingma and Ba, 2015), AdamW (Loshchilov and Hutter, 2017), or Muon (Jordan et al., 2024)) appear in practice to optimize significantly more efficiently than gradient descent. In simple terms, these algorithms are all modifications of gradient descent. Among the additional mechanisms they incorporate, an important one is *momentum*.

[1] This is significantly improvable if allowing a varying fractal-like sequence of stepsizes (Altschuler and Parrilo, 2025a), see also the strongly convex case (Altschuler and Parrilo, 2025b).

**Algorithms II: Nesterov's Momentum.** The following algorithm, which generalizes an algorithm introduced by Nesterov in 1983 (Nesterov, 1983), is the central object of this thesis. Let $\tilde{x}_0, \tilde{z}_0 \in \mathbb{R}^d$ such that $\tilde{x}_0 = \tilde{z}_0$ is the initialization of our algorithm. Let $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ be the momentum sequences, $\{\gamma_k\}_{k\in\mathbb{N}}$ and $\{\gamma'_k\}_{k\in\mathbb{N}}$ be the stepsizes sequences, all being sequences of real numbers. The Nesterov Momentum (NM) algorithm is defined as follows, for all $k \in \mathbb{N}$

$$\left\{\begin{array}{ll} \tilde{y}_k & = \tilde{x}_k + \alpha_k(\tilde{z}_k - \tilde{x}_k), \\ \tilde{x}_{k+1} & = \tilde{y}_k - \gamma_k \nabla f(\tilde{y}_k), \\ \tilde{z}_{k+1} & = \tilde{z}_k + \beta_k(\tilde{y}_k - \tilde{z}_k) - \gamma'_k \nabla f(\tilde{y}_k). \end{array}\right. \tag{NM}$$

This algorithm can be reformulated in different ways, and it is close to the well known Polyak's Heavy Ball algorithm; we refer to Section 1.4.2 for more details. The special choice $\alpha_k = 0$ for all $k \in \mathbb{N}$ reduces $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ in (NM) to (GD) with stepsizes $\{\gamma_k\}_{k\in\mathbb{N}}$. If $\alpha_k > 0$, we observe that $\tilde{x}_{k+1}$ is defined as a gradient step taken at the point $\tilde{y}_k$. This point is typically defined as a convex combination–in most situations $\{\alpha_k\}_{k\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$– between $\tilde{x}_k$ and another point $\tilde{z}_k$. This point $\tilde{z}_k$ can be seen as generated by a gradient step taken at another position, and with a different stepsize $\gamma'_k$. In many situations, this stepsize will be selected larger than $\gamma_k$, such that the sequence $\{\tilde{z}_k\}_{k\in\mathbb{N}}$ will perform a broader exploration compared with $\{\tilde{x}_k\}_{k\in\mathbb{N}}$. In some sense, the algorithm is averaging between a conservative step and a more aggressive step, the latter giving an effect of *momentum* on the first. Assuming again the convex and $L$-smooth structure (1.3), we can ensure with a proper parameterization of (NM) that $f(\tilde{x}_k) - \min_{x\in\mathbb{R}^d} f(x) = \mathcal{O}(1/k^2)$ (Nesterov, 2013). This bound improves[2] significantly over (GD), from $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$. More, for this convex and $L$-smooth structure (1.3), the dependence on $k$ in this bound cannot be improved by any first-order algorithms (Nemirovskij and Yudin, 1983). Subsequent works based on the Performance Estimation Problem (PEP) approach (Drori and Teboulle, 2014) showed that (NM) can achieve a bound that is exactly optimal among first-order algorithms (Drori, 2017), see (d'Aspremont et al., 2021) for a survey on optimal acceleration results in convex optimization. These optimality results illustrate the great potential of Nesterov's momentum. Yet, importantly, the results presented so far are derived in a setting that significantly differs from many practical problems. This is well illustrated by considering the specific form of problem (1.2).

**Challenges: Non-Convexity and Stochasticity** From the formulation (1.2) appears naturally two potential challenges.

- **Non-Convexity.** A possible choice of parameterized model is the linear model $f(u,\theta) = \langle u, \theta\rangle$, which makes $F$ in (1.2) a convex function. However, assuming a linear model may be too restrictive to capture complex relations. In the last 15 years, an old Rosenblatt (1958) class of models called *neural networks* has allowed us to achieve spectacular improvements over all existing methods (Krizhevsky et al., 2012; Vaswani et al., 2017), becoming the state of the art for many applications. The underlying minimization problem is not convex anymore, and solving (1.2) becomes, theoretically, significantly more challenging. Generally speaking, solving (1.1) with $f$ not being convex can be more difficult, with finding a global solution being intractable for high-dimensional problems in some settings (Danilova et al., 2022, Section 2.2).

- **Stochasticity.** In the case of a very large dataset, as it is typical in many applications, we have to solve (1.2) with $N \gg 1$. Here, computing $\nabla F$ then requires computing $\nabla f_i(\theta) := \frac{\partial}{\partial\theta}(f(u_i,\theta) - v_i)^2$, for each $i \in \{1, \cdots, N\}$, which can be computationally too expensive. To reduce the computational cost, instead of running (GD), it is natural to instead sample, at each iteration $k \in \mathbb{N}$, some $i_k$ from $\{1, \cdots, N\}$, and replace $\nabla f(\tilde{x}_k)$ by $\nabla f_{i_k}(\tilde{x}_k)$. This leads to the stochastic gradient descent

$$\tilde{x}_{k+1} = \tilde{x}_k - \gamma_k \nabla f_{i_k}(\tilde{x}_k),\ \gamma_k \in \mathbb{R}_+. \tag{SGD}$$

  Algorithms that exploit such stochastic estimates instead of the gradient are named **stochastic first-order** algorithms. When replacing the gradient by a stochastic estimation, the structure of the noise

[2]The bounds given in this manuscript are *worst-case* type: they upper-bound the convergence speed of an algorithm uniformly with respect to functions in a given class. It should be kept in mind that, even if algorithm A benefits from a better upper-bound compared with algorithm B when it comes to minimizing a class of functions, it does not mean that algorithm A will necessarily be faster than algorithm B for all functions in this class.

greatly influence the convergence results we can obtain. In particular, for some noise settings, even for the convex and $L$-smooth structure (1.3), the algorithm (SGD) with constant stepsizes will be such that $f(\tilde{x}_k) \nrightarrow \min_{x\in\mathbb{R}^d} f(x)$ almost surely.

Perhaps surprisingly, the training of the highly non-convex models that are neural networks appears to be tackled efficiently in practice with stochastic first-order algorithms. In particular, it is observed empirically that there is a benefit of using momentum also in this non-convex, stochastic setting, see for example (Sutskever et al., 2013; He et al., 2016).

**Our General Problem** In this thesis, we are interested in settings where it is possible to prove theoretically an acceleration property of (NM) over (GD) while $f$ is non-convex, and/or when we access stochastic estimates of the gradient. Our goal will be either to obtain new accelerated convergence results or to provide a better understanding of existing acceleration bounds. We note finally that although we do not directly study (1.2) with $f$ given by a neural network, we will often draw connections between the framework we will consider and this problem.

## 1.1 Effect of Non-Convexity and Stochastic Gradients on Acceleration with Nesterov's Momentum

An important observation is that relaxing either the convexity assumption or the deterministic nature of the gradient may prevent to achieve improved theoretical convergence guarantees with (NM). It means that the momentum mechanism is not intrinsically beneficial, which may somewhat contradict practical observations. It is thus of interest to identify non-convex and stochastic settings in which we conserve acceleration properties, and to better understand these structures. We first discuss non-convex aspects (Section 1.1.1 and Section 1.1.2), then the stochastic aspect (Section 1.1.3).

### 1.1.1 Relaxation of Convexity: Smoothness Only

A clear example of loss of acceleration guarantee is given when relaxing the convex and $L$-smooth structure (1.3) by the single $L$-smooth property

$$\forall x, y \in \mathbb{R}^d, \quad f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2}\|x-y\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x-y\|^2. \tag{1.4}$$

Discarding convexity is a severe relaxation. Indeed under convexity (left inequality in (1.3)), we have the property $\nabla f(x^*) = 0 \Leftrightarrow x^* \in \operatorname{argmin} f$. It does not necessarily hold anymore, assuming (1.4). In such a setting, finding a minimizer of $f$ is intractable in general, specifically for high-dimensional problems ($d \gg 1$) (Danilova et al., 2022, Section 2.2). Instead, one can target a point $\tilde{x} \in \mathbb{R}^d$ ensuring $\|\nabla f(\tilde{x})\| \leq \varepsilon$, for some precision $\varepsilon > 0$. Gradient descent (GD) with proper parameterization finds such a point in at most $\mathcal{O}(1/\varepsilon^2)$ iterations (Nesterov, 2013). This dependence in $\varepsilon$ is not improvable by any first-order algorithm (Carmon et al., 2020). In particular, (NM) cannot improve it, only achieving the same bound (Ghadimi and Lan, 2016).

**Acceleration with Lipschitz Gradient and Hessian** Property (1.4) is equivalent to have that the gradient of $f$ is $L$-Lipschitz (Chapter 3, Property 3.2). Since 2017, we know that if together with this property, we also assume that $f$ is twice differentiable with Lipschitz Hessian, the $\mathcal{O}(\varepsilon^{-2})$ bound can be improved to $\mathcal{O}(\varepsilon^{-7/4})$ using first-order methods (Carmon et al., 2017; Jin et al., 2018; Li and Lin, 2023). In this setting, this dependence in $\varepsilon$ is not improvable by any other first-order methods, according to a very recent paper (Zhou, 2026). The algorithms for which this bound is proven are all based on the following special case of (NM):

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \alpha_k(\tilde{x}_k - \tilde{x}_{k-1}), \\ \tilde{x}_{k+1} = \tilde{x}_k - \gamma \nabla f(\tilde{y}_k). \end{cases} \tag{NM'}$$

It is a special case in the sense that (NM) can be rewritten in a form that reduces to (NM') for some choice of parameters (Chapter 3, Proposition 3.B.1). In this Lipschitz gradient and Hessian setting, Nesterov's

momentum thus helps ensuring a better convergence guarantee. However, one cannot obtain this result by simply running (NM'), as we discuss next.

**The $\mathcal{O}(\varepsilon^{-7/4})$ Bound Requires Momentum-Control Mechanisms** At each iteration, the methods that achieve the bound $\mathcal{O}(\varepsilon^{-7/4})$ require performing a *security check*. If it fails at iteration $k$, the **momentum built up so far is erased**, which is practically done by setting $\alpha_k = 0$. At a conceptual level, these procedures can be summarized as follows:

**Algorithm 1** Structure of existing algorithms achieving the $\mathcal{O}(\varepsilon^{-7/4})$ complexity

1: **if** Security check succeeds, **then**
2: Run update rule (NM')
3: **else**
4: Erase accumulated momentum (+ potentially trigger alternative mechanisms)
5: **end if**

In (Carmon et al., 2017; Jin et al., 2018), the security check monitors the level of non-convexity between consecutive iterates. When it exceeds a given threshold:

$$f(\tilde{x}_k) \leq f(\tilde{y}_k) + \langle \nabla f(\tilde{y}_k), \tilde{x}_k - \tilde{y}_k \rangle - \frac{\ell}{2} \|\tilde{x}_k - \tilde{y}_k\|^2, \quad \ell > 0,$$

accumulated momentum is erased and restarted, and a negative-curvature exploitation step is triggered as an alternative mechanism. (Li and Lin, 2023) employs a security check based on the length of the trajectory residual. Exceeding a given threshold

$$k \sum_{t=0}^{k-1} \|\tilde{x}_{k+1} - \tilde{x}_k\|^2 > B^2, \quad B > 0,$$

it activates a restart mechanism, thus erasing the accumulated momentum.

**Discussion** We stated results from the literature that indicate that the use of momentum can provide a theoretical acceleration, which shows that its benefit can be proven outside of the convex world. However, these acceleration results require controlling the momentum phenomenon, at least by using restart mechanisms. Restart mechanisms can be interesting: they have been shown to provide benefits in the convex setting, enhancing practical behavior (O'donoghue and Candes, 2015) or helping design methods that do not need the knowledge of typically unknown constants involved in geometrical assumptions (Aujol et al., 2024a). However, restart mechanisms are not mandatory to achieve theoretical acceleration in this convex setting. More generally, they do not seem to be necessary for momentum methods to perform well in practical non-convex problems. The existing results may somewhat contradict this observation, as they may suggest that enforcing a momentum-control mechanism is necessary to achieve an accelerated convergence in this setting. This raises the question:

**Question: Are momentum-control mechanisms fundamentally necessary to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity?**

This question is of interest to better understand acceleration on a class of function that encompasses highly non-convex problems. As a concrete example, the squared loss for regression with a neural network whose activation functions are with Lipschitz first and second derivatives, *e.g.* if using SoftPlus activations, belongs to this class (Renaud et al., 2026, Proposition 2). More generally, functions that are $C^3$ belong to this class, as long as the domain is restricted to a compact set.

### 1.1.2 Relaxation of Strong Convexity

We will have a special interest in functions relaxing the $\mu$-strongly convex and $L$-smooth structure, which we recall now

$$\forall x, y \in \mathbb{R}^d, \quad f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2, \tag{1.5}$$

for $0 < \mu < L$. Enforcing $\mu = 0$ in the above recovers the convex, $L$-smooth structure. Having $\mu > 0$ describes a setting that is even more favorable for optimization: there exists a unique global minimizer, and one can obtain linear convergence guarantees. In particular, gradient descent (GD) with constant stepsizes equal to $1/L$ ensures $f(\tilde{x}_k) - f^* = \mathcal{O}((1 - \frac{\mu}{L})^k)$ (Nesterov, 2013). The ratio $L/\mu \geq 1$ is called *condition number*, and characterizes how ill-conditioned the minimization problem is. A higher condition number typically reflects a more difficult problem, which is reflected by a worse convergence guarantee.

**Another Acceleration Setting** When (1.5) holds, Nesterov's momentum (NM) can ensure $f(\tilde{x}_k) - f^* = \mathcal{O}((1 - \sqrt{\frac{\mu}{L}})^k)$ (Nesterov, 2013, Section 2.2). Compared with the convex $L$-smooth case (1.3), this is a new instance of acceleration over gradient descent. Indeed, as typically for large-dimensional problems, we have $L/\mu \gg 1$, improving from $\mu/L$ to $\sqrt{\mu/L}$ yields a substantial gain. However, as for the convex case, assuming (1.5) is too restrictive for many applications. We note in the case of deep-learning applications, even the upper-bound in (1.5) is arguably not a relevant assumption (Zhang et al., 2019a; Cohen et al., 2021), but relaxing this aspect is not considered in this manuscript. Instead, we are interested in the possibility of relaxing the lower-bound in (1.5), while keeping convergence bounds that are both **accelerated** and **linear**.

**The Polyak-Łojasiewicz Relaxation** Among relaxations of strong convexity, a class that attracted a significant amount of interest are Polyak-Łojasiewicz (PL) functions (Polyak, 1963; Łojasiewicz, 1963). A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to this class if there exists $\mu > 0$ such that the following inequality holds for all $x \in \mathbb{R}^d$

$$\|\nabla f(x)\|^2 \geq 2\mu\Big(f(x) - f^*\Big). \tag{PL}$$

This inequality has been considered independently by Polyak (Polyak, 1963), and by Łojasiewicz as a special case of the Łojasiewicz inequality (Łojasiewicz, 1963). A function that is $\mu$-strongly convex (*i.e.* satisfies the lower-bound in (1.5)) verifies the above inequality. It is direct to observe that (PL) implies that a critical point is a global minimizer. Importantly, under this inequality and $L$-smoothness (1.4), gradient descent (GD) with constant stepsize equal to $1/L$ ensures $f(\tilde{x}_k) - f^* = \mathcal{O}((1 - \frac{\mu}{L})^k)$, which writes similarly as the convergence bound in the strongly convex case. These functions enable linear convergence bounds, and can be nonconvex: they may exhibit oscillations, as shown by the simple 1-dimensional example $f(t) = t^2 + 3\sin(t)^2$, that is PL for some $\mu > 0$ (Garrigos and Gower, 2023, Lemma A.4). For more practical examples, this assumption can appear naturally when dealing with least squares losses. In particular, the square loss of a sufficiently over-parameterized neural networks, *i.e.* if they have enough parameters, will verify the PL inequality on a subspace of $\mathbb{R}^d$ (Liu et al., 2022). In a recent paper, it was shown that if a PL function is smooth enough (say $C^\infty$), it can be written as a non-linear least square (Boumal et al., 2026).

This class of functions seems then to be a good model of non-convex problems that arise from least-square losses, providing an interesting framework to study the acceleration phenomenon of momentum in a non-convex setting. Unfortunately, apart from potentially improving the convergence bound by a numerical factor, a theoretical acceleration when using momentum cannot be demonstrated for PL, $L$-smooth functions. Precisely, the dependence on $\mu/L$ in the bound cannot be improved (Yue et al., 2023). In particular, without additional assumptions, such as higher-order smoothness of $f$ and initializing the algorithm close enough to a solution (Kassing and Weissmann, 2024; Feng and Yuan, 2026)[3], we cannot obtain the squared root dependence in $\mu/L$ that is obtained with (NM) when minimizing strongly convex functions (1.5). It is thus of interest whether we can identify a relaxation of strong convexity with more structure that yet allows for accelerated convergence bounds with Nesterov's momentum.

**An Intermediate Class: Strong Quasar Convexity** An intermediate non-convex class that is expected to capture acceleration by momentum is *strong quasar convexity* (Hinder et al., 2020), sometimes referred to as *weak-quasi-strong convexity* (Necoara et al., 2019; Bu and Mesbahi, 2020). A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to this class if there exists a minimizer $x^* \in \mathbb{R}^d$ and parameters $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$ such that for all $x \in \mathbb{R}^d$

$$f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2. \tag{1.6}$$

[3](Kassing and Weissmann, 2024) actually considers Polyak's Heavy Ball algorithm (PM), which is arguably a different kind of momentum (Shi et al., 2022). (Feng and Yuan, 2026) considers the specific form of (NM').

Such functions verify the PL inequality with parameter $\tau^2\mu$ (Chapter 3, Proposition 3.2.1). It is immediate that strongly convex functions satisfy (1.6) with $\tau = 1$. With strongly convex functions, strongly quasar convex functions share the regularity that they admit a unique minimizer. This prevents many problems from directly fitting into this class. Still, it can model interesting non-convex behaviors. Indeed, these functions can exhibit important oscillations, see Figure 1.1, and some generalized linear models fit into this class (Wang and Wibisono, 2023). Moreover, from a more empirical perspective, several works (Guille-Escuret et al., 2022, 2024) observed that, when training neural networks, an equivalent formulation of (1.6) holds along the optimization path of certain stochastic first-order algorithms, up to replacing a (unknown) solution $x^*$ in (1.6) with the final iterate of the algorithm. This indicates that, by relaxing the global aspect of the definition of strong quasar convexity to consider the geometry of functions along the optimization path, it becomes a relevant assumption for modeling complicated models.

We will show further that if such functions are also $L$-smooth, gradient descent (GD) ensures $f(\tilde{x}_k) - f^* = \mathcal{O}\left((1 - \tau\frac{\mu}{L})^k\right)$. On the other hand, (NM) can ensure $f(\tilde{x}_k) - f^* = \mathcal{O}\left((1 - \tau\sqrt{\frac{\mu}{L}})^k\right)$ (Hinder et al., 2020). We thus recover the bounds of the smooth strongly convex case up to this $\tau$ factor, indicating we retain its acceleration phenomenon. Then, it seems that strong quasar convexity is a good tradeoff between relaxing convexity and achieving acceleration. Although this may seem sufficiently satisfactory to stop here, there is, in fact, more to be said.

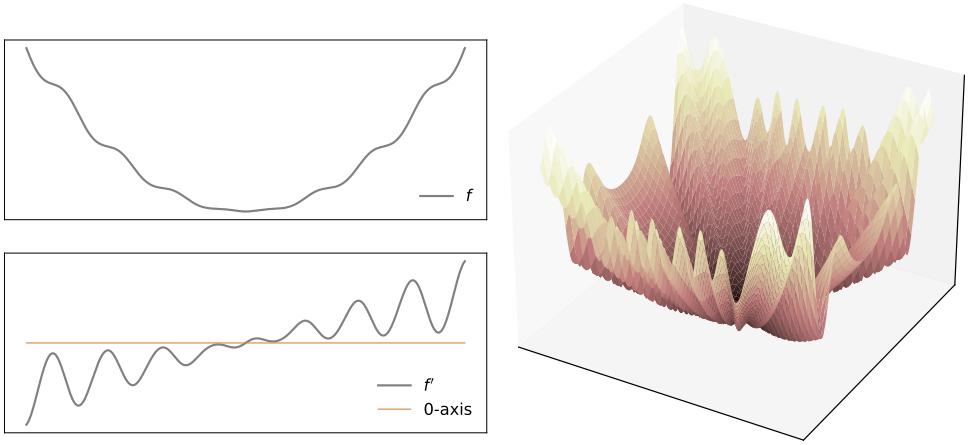



Figure 1.1: Examples of strongly quasar convex functions. 1-dimensional example on the left (its derivative on the bottom), 2-dimensional example on the right.

**Discussion on the Mechanisms to Achieve the Bound $\mathcal{O}((1 - \tau\sqrt{\frac{\mu}{L}})^k)$** There are two types of first-order algorithms that achieve the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$, all based on (NM). A first requires performing a subroutine at each iteration $k$ to compute $\alpha_k$ in (NM) (Hinder et al., 2020), which requires performing evaluations of $f$ and $\nabla f$. A second is to use a stochastic parameterization of (NM) - $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ are stochastic - that achieves the bound in expectation: $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}\left(\left(1 - \tau\sqrt{\frac{\mu}{L}}\right)^k\right)$, without subroutines (Wang and Wibisono, 2023). The first induces increased computational cost, the second a non-deterministic result. Moreover, such mechanisms are not mandatory to achieve acceleration in the strongly convex case, which may suggest that these mechanisms are a cost to pay because of relaxing it with strong quasar convexity. Naturally, a question arises:

**Question: Can we achieve the bound $\mathcal{O}((1 - \tau\sqrt{\frac{\mu}{L}})^k)$ with a fixed in advance, deterministic parameterization of (NM) ?**

**Discussion on Improving the Bound $\mathcal{O}((1 - \tau\sqrt{\frac{\mu}{L}})^k)$** In the smooth strongly convex case, Nesterov's convergence bound can be improved, see Taylor and Drori (2023) for one that reaches optimality. As an example of improved bound, (Aujol et al., 2022) shows that we can obtain a convergence bound that writes as $\mathcal{O}\left(\left(1 + \sqrt{2\frac{\mu}{L} - 6\frac{\mu}{L}}\right)^{-k}\right)$. When $\mu/L$ is small, this is close to a bound of the form $\left(1 - \sqrt{2\frac{\mu}{L}}\right)^k$, improving

by a $\sqrt{2}$ factor over the previously known rate $\left(1-\sqrt{\frac{\mu}{L}}\right)^k$. Obtaining such refinement under strong quasar convexity may not be that trivial, as even achieving the standard bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ requires special care, as discussed in the previous paragraph.

**Question: Can we improve upon the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ ?**

**Discussion on the Uniqueness of the Minimizer** The property of having a unique minimizer, inherited from the definition, is something we would prefer to avoid. In the case of over-parameterized models, it is expected that the set of minimizers is a submanifold of $\mathbb{R}^d$ whose dimension grows as the over-parameterization is important (Cooper, 2021). A strategy to relax this property is to replace a minimizer $x^*$ in (1.6) by $\pi(x)$, where $\pi : \mathbb{R}^d \to \mathbb{R}^d$ is the orthogonal projection onto $\operatorname{argmin} f$, leading to for all $x \in \mathbb{R}^d$,

$$f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), \pi(x) - x\rangle + \frac{\mu}{2}\|\pi(x) - x\|^2 . \tag{1.7}$$

It is done for instance in (Gupta and Wojtowytsch, 2025) in the case $\tau = 1$ and under $L$-smoothness. However, to achieve a rate of the order $\mathcal{O}((1-\sqrt{\frac{\mu}{L}})^k)$, they assume that **$\operatorname{argmin} f$ is an affine space**. This would be a restrictive assumption even if $f$ were assumed to be convex. As it happens, replacing $x^*$ by an orthogonal projection has been used for a convex relaxation of strong convexity (Aujol et al., 2024c), a setting that can be understood as having (1.7) with $\tau = 1$, $L$-smoothness and convexity (1.3). In their work, the set of minimizers is a convex set (necessarily, because of the convexity of $f$), but it is not required to be an affine space. It raises the question:

**Question: Can we achieve the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ while relaxing the uniqueness of minimizer property beyond the affine-set assumption ?**

**Discussion on Interpretation** Conceptually, it may not be straightforward to understand why we can achieve different bounds for (GD) and (NM) under strong quasar convexity, while it is not possible under PL. Specifically, it raises the question of the supplementary structure that benefits strong quasar convexity compared with PL. Acceleration cannot be explained by the property of having a unique minimizer; such a property is also verified by the functions that serve as lower bounds to show the optimality of GD for smooth PL functions (Yue et al., 2023). This leads to the following question:

**Question: Up to the uniqueness of the minimizer property, can we provide an intuitive characterization of what supplementary structure benefits strongly quasar convex functions compared with PL functions ?**

Also, it is not necessarily clear what the role of $\tau$ and $\mu$ is in the definition (1.6). By analogy with the strongly convex case, we may be naturally tempted to interpret $L/\mu$ as a conditioning measurement. However, if a function satisfies (1.6) for some pair $(\tau, \mu)$, it also satisfies it with the pair $(\theta\tau, \mu/\theta)$, for any $\theta \in (0, 1]$ (Hinder et al., 2020, Observation 5). Taking $\theta$ small enough, it implies that one may verify both strong quasar convexity (1.6) and $L$-smoothness while having $\mu > L$. In such a case, the conditioning measurement interpretation loses its meaning, and it questions the convergence rates. In particular, this indicates that stating that an acceleration phenomenon similar to strong convexity holds for these functions might be a misleading shortcut, namely:

**Question: Are strongly quasar convex functions truly benefiting from a similar acceleration phenomenon compared with strongly convex functions ?**

We will see in Chapter 5 that the answer is negative. It does not mean that the results established under this assumption are obsolete, but rather that there is a hidden variable to consider to have a better picture of acceleration.

We summarize in the following table all the convergence results introduced so far.

| | Upper-Bound (GD) | Upper-Bound (NM) | Lower Bound |
|---|---|---|---|
| $L$-smooth (Grad.) | $\frac{L}{\varepsilon^2}$ | $\frac{L}{\varepsilon^2}$ | $\frac{L}{\varepsilon^2}$ |
| + $L_2$-Lipschitz Hessian (Grad.) | $\frac{L}{\varepsilon^2}$ | $\frac{L^{1/2}L_2^{1/4}}{\varepsilon^{7/4}}$ | $\frac{L^{1/2}L_2^{1/4}}{\varepsilon^{7/4}}$ |
| + Conv. ($f$) | $\frac{L}{\varepsilon}$ | $\sqrt{\frac{L}{\varepsilon}}$ | $\sqrt{\frac{L}{\varepsilon}}$ |
| + $\mu$-Strong Conv. ($f$) | $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$ | $\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$ | $\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$ |
| +$(\tau,\mu)$-Strong Quas. Conv. ($f$) | $\frac{L}{\tau\mu}\log\left(\frac{1}{\varepsilon}\right)$ | $\frac{1}{\tau}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$ | $\frac{1}{\sqrt{\tau}}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$ |
| + $\mu$-Polyak-Łojasiewicz ($f$) | $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$ | $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$ | $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$ |

Table 1.1: Summary of the convergence results presented so far. Conv. is for Convexity, Quar. for Quasar. (Grad.) indicates a bound on the number of gradient evaluations to achieve a point that verifies $\|\nabla f\| \leq \varepsilon$. ($f$) indicates a bound on the number of gradient evaluations to achieve a point that verifies $f - f^* \leq \varepsilon$. All stated convergence bounds have been translated in this language: $f(\tilde{x}_k) - f^* = \mathcal{O}((1-\gamma)^k)$ for some $\gamma \in (0,1)$ implies a bound $\mathcal{O}((1/\gamma)\log(1/\varepsilon))$, while $f(\tilde{x}_k) - f^* = \mathcal{O}(K_0/k^\alpha)$ for some $K_0, \alpha > 0$ implies a bound $\mathcal{O}((K_0/\varepsilon)^{1/\alpha})$. The Lower Bound column provides lower bounds that hold for all first-order algorithms. Numerical factors, other logarithmic factors, and dependence on initialization have been omitted. All associated references have been provided before, except for the lower bound $\frac{1}{\sqrt{\tau}}\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)$ (Hinder et al., 2020), and the upper-pound under PL for (GD), which is proven in Proposition 4.3.2 and holds with high probability. The bound for (GD) under Lipschitz Hessian holds by default because of the $L$-smoothness property. To our knowledge, it is not clear wether also having the Hessian Lipschitz property enables to improve the bound Cartis et al. (2010). The bounds in red are those for which we bring new contributions.

### 1.1.3 Relaxing Access to Exact Gradient: retaining acceleration with Strong Growth Condition

We consider the case of accessing a stochastic estimate of $\nabla f$ instead of $\nabla f$ itself. Motivated by the problem (1.2), we consider the following finite sum problem

$$\operatorname{argmin}_{x\in\mathbb{R}^d} f(x), \text{ with } f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x). \tag{FS}$$

We consider a stochastic estimate of the form $\tilde{\nabla}_K(x) := \frac{1}{K}\sum_{i\in B}\nabla f_i(x)$, where $B$ is a batch of indices of size $K$ sampled uniformly from $\{B \subset \{1,\dots,N\} \mid \operatorname{Card}(B) = K\}$. Apart from the geometrical structure of the function, the structure of the noise is now also of great importance. If each $f_i$ is $L_i$-smooth (1.4), a specific noise structure follows, for instance, in the case $K = 1$ (Garrigos and Gower, 2023, Lemma 4.19)

$$\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right] \leq 2L_{\max}(f(x) - f^*) + 2L_{\max}\left(\min f - \mathbb{E}\left[\min f_i\right]\right), \tag{1.8}$$

with $L_{\max} = \max_{1\leq i\leq N} L_i$. The presence of the constant term on the right-hand side illustrates that in general, one may have $\mathbb{E}\left[\left\|\tilde{\nabla}_1(x^*)\right\|^2\right] \neq 0$, for any $x^* \in \operatorname{argmin} f$. In such settings, even upon reaching a solution, the stochastic estimator continues to perturb the iterates. As a consequence, in this case, algorithms such as stochastic gradient descent (SGD) with fixed stepsizes cannot converge. Adjustments are needed to obtain convergence, such as taking a sequence of stepsizes going to zero, or using variance reduction techniques such as SVRG (Johnson and Zhang, 2013) or SAGA (Defazio et al., 2014). In the case where we keep a geometrical structure that allows acceleration in the deterministic setting, namely $f$ is smooth and convex (1.3) or strongly convex (1.5), the first adjustment necessarily deteriorates the convergence bound compared with the deterministic setting, while the second requires periodic computations of the full gradient (SVRG-like methods) or requires storing a table of gradients of size $N \times d$ (SAGA-like methods).

**Limit of the scope of this work: the question of generalization** From a machine learning perspective, where each $f_i$ in (FS) is associated with a training example, minimizing the objective in (FS) is not necessarily the ultimate goal. The objective measures performance on the $N$ training examples, whereas we ultimately want the model to perform well on unseen data. This is the question of *generalization*. Although generalization lies beyond the scope of this manuscript, we note that the choice of algorithm can influence it (Soudry et al., 2018; Jelassi and Li, 2022).

**The special case of interpolation** The constant term in (1.8) is zero as long as

$$\exists x^* \in \arg\min f, \forall 1 \leq i \leq N,\ x^* \in \arg\min f_i \tag{interpolation}$$

Such a setting is called *interpolation*, and it is of interest for some practical problems; overparameterized models are generically able to perfectly fit the learning data (Cooper, 2021; Allen-Zhu et al., 2019; Nakkiran et al., 2021; Zhang et al., 2021), which in turn implies interpolation. From an optimization point of view, such a phenomenon implies that for this $x^*$, $\mathbb{E}\left[\left\|\tilde{\nabla}_K(x^*)\right\|^2\right] = 0$, for any $K \in \{1, \cdots, N\}$. The fact that there exists a solution $x^*$ where the variance of the estimator is zero has important implications: we can use (SGD) with constant stepsizes while ensuring convergence, and we retain the same order in the convergence bound compared with the deterministic case. Precisely, for (SGD) with batch size 1, if each $f_i$ is $L_i$-smooth we obtain $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}(1/k)$ in the convex case (Gower et al., 2021), and $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}((1 - \frac{\mu}{L_{\max}})^k)$ in the $\mu$-strongly convex case (Gower et al., 2019). In light of this observation, considering a stochastic version of (NM), namely the Stochastic Nesterov Momentum algorithm:

$$\begin{cases} \tilde{y}_k & = \tilde{x}_k + \alpha_k(\tilde{z}_k - \tilde{x}_k), \\ \tilde{x}_{k+1} & = \tilde{y}_k - \gamma_k \tilde{\nabla}_K(\tilde{y}_k), \\ \tilde{z}_{k+1} & = \tilde{z}_k + \beta_k(\tilde{y}_k - \tilde{z}_k) - \gamma'_k \tilde{\nabla}_K(\tilde{y}_k), \end{cases} \tag{SNM}$$

it is natural to wonder if it retains convergence rates of similar order than the deterministic setting. If so, it would imply a significant acceleration compared with (SGD).

**Acceleration under** (interpolation) **with** (SNM)**: Negative Results** Regarding the hope of obtaining such deterministic-like acceleration, several negative results are known. One reason is that a stochastic estimator can be viewed as an inexact gradient, that is, as a gradient corrupted by some error. The works (Devolder et al., 2014; Aujol and Dossal, 2015) show that momentum algorithms are more sensitive to gradient errors than gradient descent, because the momentum mechanism tends to accumulate these errors. Moreover, the parameter choices that yield acceleration in the deterministic setting may cause stochastic versions of Nesterov's momentum to diverge, even in the favorable case where each $f_i$ is $\mu$-strongly convex and $L$-smooth, each sharing the same minimizer (Assran and Rabbat, 2020). To our knowledge, existing acceleration results hold in the specific case of linear regression (Jain et al., 2018; Liu and Belkin, 2020; Varre and Flammarion, 2022), and still, depending on the data distribution, acceleration may not hold. The tools used in this linear regression setting seem hardly generalizable outside of this case.

**A Good Noise Control for Acceleration: Strong Growth Condition** The strong growth condition (SGC) (Polyak, 1987) assumes that for some $\rho_K \geq 1$ we have

$$\forall x \in \mathbb{R}^d,\ \mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] \leq \rho_K \|\nabla f(x)\|^2. \tag{SGC}$$

The $\rho_K$ quantifies the noise level: the higher $\rho_K$, the higher the noise, $\rho_K = 1$ corresponding to no noise. In (Vaswani et al., 2019), it is demonstrated that under (SGC), (SNM) can recover the deterministic acceleration with (NM) up to a $\rho_K$ factor. Precisely, we obtain $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}(\rho_K^2/k^2)$ for convex $L$-smooth functions, and $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}((1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}})^k)$ for $\mu$-strongly convex $L$-smooth functions. In contrast, under (SGC), (SGD) achieves $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}(\rho_K/k)$ for convex $L$-smooth functions, and $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}((1 - \frac{1}{\rho_K}\frac{\mu}{L})^k)$, respectively (Gupta et al., 2024). This noise assumption thus seems to describe the suitable noise setting such that the deterministic acceleration phenomenon still holds.

**Strong Growth Condition: Existing Characterizations.** A natural question that arises from the previous paragraph is how relevant this assumption is and how restrictive it may be. Existing examples and characterizations are rather limited. As a concrete example, (Vaswani et al., 2019) show that the squared-hinge loss, used for training a linear classifier for binary classification, verifies (SGC) under some assumptions on the data. More generally, they show that if $f$ verifies (PL), (interpolation), and all $f_i$ being $L_i$-smooth (1.4), then (SGC) with batch size 1 is verified with $\rho_1 = L_{max}/\mu$.

**Discussion: Acceleration ?** In the case of strongly convex, $L$-smooth functions, under (SGC), it appears that the bound $\mathcal{O}((1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}})^k)$ of (SNM) improves over the bound $\mathcal{O}((1 - \frac{1}{\rho_K}\frac{\mu}{L})^k)$ of (SGD), for any parameters (except $\mu = L$). However, one could also consider a bound for (SGD) considering strong convexity, (interpolation), and all $f_i$ being $L_i$-smooth (1.4), but not (SGC). With batch size $K = 1$, such a bound is $\mathcal{O}((1 - \frac{\mu}{L_{\max}})^k)$. In this case, the characterization $\rho_1 = L_{max}/\mu$ stated in the previous paragraph holds. Plugging this value of $\rho_1$ in the bound $\mathcal{O}((1 - \frac{1}{\rho_1}\sqrt{\frac{\mu}{L}})^k)$ of (SNM), it becomes $\mathcal{O}((1 - \frac{\mu}{L_{max}}\sqrt{\frac{\mu}{L}})^k)$. The interplay between $L$ and $L_{\max}$ is not important here; just consider that we have $\mu \leq L$. What is important is that using this characterization of $\rho_1$, the bound of (SNM) derived with (SGC) is actually worse than the one of (SGD) derived without (SGC). Comparing the two bounds, we obtain that the one of (SNM) is better as long as an inequality of the form $\rho_1 \leq \sqrt{L^2_{\max}/(\mu L)}$ holds. There is no reason to think that this bound holds in all situations, which indicates that the acceleration statement under (SGC) could be misleading in general. This raises the following question.

**Question: In the convex and interpolation setting, can we provide a clearer characterization of what can make** (SNM) **accelerate over** (SGD) **?**

In fact, we will see that the characterization $\rho_1 = L_{\max}/\mu$ can be arbitrarily sub-optimal. Our answer to this question will be closely tied to a new characterization of the constant in (SGC).

## 1.2 A Convergence Analysis Method and its Limitation: limit-ODE of Algorithms

The gradient descent algorithm (GD) can be seen as an Euler discretization of the following ordinary differential equation (ODE)

$$\dot{x}_t = -\nabla f(x_t), \quad t \geq 0, \tag{GF}$$

for some $x_0 \in \mathbb{R}^d$. Similarly, depending on the parameter choices, (NM) can be seen as a discretization of the following system, which we call Nesterov Momentum ODE (NMO)

$$\left\{ \begin{array}{ll} \dot{x}_t & = \eta_t(z_t - x_t) - \gamma_t \nabla f(x_t) \\ \dot{z}_t & = \eta'_t(x_t - z_t) - \gamma'_t \nabla f(x_t) \end{array} \right. \tag{NMO}$$

for sequences $(\eta_t)_{t\geq 0}, (\eta'_t)_{t\geq 0}, (\gamma_t)_{t\geq 0}, (\gamma'_t)_{t\geq 0}$ in $\mathbb{R}^{\mathbb{R}_+}$, see (Kim and Yang, 2023, Appendix A.1). A first interest of considering such objects is that algorithms that have the same limit-ODE can be classified as belonging to the same family, such that these limit-ODE can be seen as unifying objects. In the reverse way, one can use various discretizations of an ODE to deduce new algorithms and study their convergence. This has been done, for instance, in (Siegel, 2019), that mix explicit and semi-implicit steps, or (Zhang et al., 2019b), which uses a Runge-Kutta discretization[4]. Another interest is that the study of these object may serve as a guideline to study the corresponding algorithms, as we discuss next.

**Limit-ODE of algorithms: Tool for Convergence Analysis** We can study the convergence of the solution of (GF) or (NMO) using Lyapunov strategies, *i.e.*, introducing a function $E(t, \overline{x}_t)$, $t \geq 0$. The function is typically chosen such that its decrease with time provides interesting information on the convergence. Such a decrease can be proven by showing $\dot{E}(t, \overline{x}_t) \leq 0$. Importantly, the possibility to differentiate can significantly simplify computation compared with computations involving sequences generated by algorithms. Once a satisfying result is obtained for the solution of the ODE, several doors open. When considering algorithms

[4]In both cases, the ODE discretized is a special case of (NMO) (the heavy ball equation).

with decreasing step-sizes, one can use the fact that the algorithm asymptotically behaves like the associated ODE to deduce a convergence result for this algorithm (Castera et al., 2021). Otherwise, the continuous-time analysis can serve as an inspiration to carry out a new analysis for the algorithm, providing a testbed to find a relevant Lyapunov function. As an example, for $f$ convex, one may introduce $(\overline{x}_t)_{t\geq 0} = (x_t, z_t)_{t\geq 0}$ as solutions of (NMO), and consider

$$E(t, \overline{x}_t) = \frac{t^2}{4}(f(x_t) - f^*) + \frac{1}{2}\left\|z_t - x^*\right\|^2.$$

Under proper parameterization of (NMO), we have $\dot{E}(t, \overline{x}_t) \leq 0$, from which we can deduce $f(x_t) - f^* = \mathcal{O}(1/t^2)$ (Su et al., 2016). Now, for $\{\overline{x}_k\}_{k\in\mathbb{N}} = \{\tilde{x}_k, \tilde{z}_k\}_{k\in\mathbb{N}}$ generated by (NM), we can study a strongly inspired discrete-time Lyapunov function

$$E(k, \overline{x}_k) = \frac{k^2}{4L}(f(\tilde{x}_k) - f^*) + \frac{1}{2}\left\|\tilde{z}_k - x^*\right\|^2.$$

Under suitable parameterization of (NM), it is decreasing for $f$ convex, $L$-smooth function, implying in turn the result $f(\tilde{x}_k) - f^* = \mathcal{O}(1/k^2)$ (Su et al., 2016). Here, it was possible to adapt the continuous-time analysis to transfer the bound $\mathcal{O}(1/t^2)$ from the continuous-time setting to a bound $\mathcal{O}(1/k^2)$ in the discrete-time setting. This specific result and Lyapunov function in the discrete setting already existed in the literature Nesterov (1983). However, this point of view can help to guide the intuition, and has been used to obtain several extensions in the convex setting, see *e.g.* (Su et al., 2016; Siegel, 2019; Attouch et al., 2022, 2023; Shi et al., 2022; Aujol et al., 2023a, 2024c; Li et al., 2024). A drawback of this strategy is that the analysis has to be carried twice, both in the continuous and discrete case. Another seems to appear when $f$ is not convex anymore, as we discuss next.

**Limitation of the Strategy in the Non-Convex Setting** When $f$ is not convex, the convergence properties of $(x_t)_{t\geq 0}$ may not transfer to $\{x_n\}_{n\geq 0}$ as well as in the convex case. There exists some evidence in the literature:

- In the case of $L$-smooth functions, a certain kind of Lyapunov function (the total energy) decreases when applied to the solution of the continuous Heavy Ball equation, regardless of the parameterization of the ODE. For the discrete counterpart, an assumption on the stepsize is required, which intuitively becomes more restrictive as we increase the effect of momentum. Violating this assumption can make the discrete version of the Lyapunov function increase, see (Jin et al., 2018, Section A.1).
- For functions with Lipschitz gradient and Hessian, a convergence result exists for the continuous Heavy Ball equation (a special case of (NMO)), achieving a critical point in $\mathcal{O}(\varepsilon^{-7/4})$ computation time (Okamura et al., 2024). This is the continuous counterpart of the results discussed in Section 1.1.1. One can wonder whether a direct discretization of this ODE may lead to a momentum algorithm that achieves the bound $\mathcal{O}(\varepsilon^{-7/4})$. This would solve the question about the necessity of momentum control mechanisms, see our discussion in Section 1.1.1. The authors, however, require a supplementary restrictive assumption to obtain such a result, see the first arXiv version of their work. They interpret this supplementary assumption as inherited from the discretization process.

We will further show that such discrepancies also occur in the strongly quasar convex setting. This leads to the following observation:

**There exist non-convex settings where results exist for the ODE that are not matched with a result for an algorithm, or only under additional restrictions or needing additional mechanisms.**

In particular, it raises the question whether this ODE-based strategy is less relevant when considering non-convex functions.

**Another Method: the Continuized Nesterov Equations** In (Even et al., 2021), the following process is introduced. Consider discrete random times $T_1, T_2, \dots$ where for all $i$, $T_{i+1} - T_i$ are independent and follow an exponential law of parameter 1. Between these random times, define a process $(x_t, z_t)_{t\geq 0}$ as the solution of the following ODE

$$\left\{ \begin{array}{ll} dx_t &= \eta_t(z_t - x_t)dt \\ dz_t &= \eta'_t(x_t - z_t)dt \end{array} \right. \tag{1.9}$$

If $\eta_t, \eta'_t \geq 0$ for all $t > 0$, $(x_t)_{t\geq 0}$ and $(z_t)_{t\geq 0}$ are linearly interpolating on each interval $(T_i, T_{i+1})$. At $t = T_i$, the process *jumps* by performing gradient steps

$$x_{T_i} = x_{T_i^-} - \gamma_{T_i^-}\nabla f(x_{T_i^-}), \quad z_{T_i} = z_{T_i^-} - \gamma'_{T_i^-}\nabla f(x_{T_i^-}), \tag{1.10}$$

where $x_{T_i^-}$ denotes the left-limit $\lim_{t\to 0^-} x_{T_i+t}$. This process writes as a stochastic differential equation that we call the *continuized Nesterov equations* (named Continuized Nesterov acceleration in the original paper (Even et al., 2021)):

$$\left\{ \begin{array}{ll} dx_t &= \eta_t(z_t - x_t)dt - \gamma_{t^-}\nabla f(x_{t^-})N(dt) \\ dz_t &= \eta'_t(x_t - z_t)dt - \gamma'_{t^-}\nabla f(x_{t^-})N(dt) \end{array} \right. \tag{CNE}$$

where $N(dt) = \sum_{k\geq 0}\delta_{T_k}(dt)$ is a Poisson point measure. A core property of (CNE) is that the evaluations of the process $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$, $k \in \mathbb{N}$, can be **computed exactly**. Precisely, one can show that they are solution of a Nesterov momentum algorithm of the form (NM):

$$\left\{ \begin{array}{ll} \tilde{y}_k &= (\tilde{z}_k - \tilde{x}_k)\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + \tilde{x}_k \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma\nabla f(\tilde{y}_k, \xi_{k+1}) \\ \tilde{z}_{k+1} &= \tilde{z}_k + \eta'\frac{\left(1-e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta'+\eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(\tilde{y}_k - \tilde{z}_k) - \gamma'\nabla f(\tilde{y}_k, \xi_{k+1}) \end{array} \right. \tag{CNA}$$

with stochastic parameters involving the random times $\{T_k\}_{k\geq 0}$. We call this variant the Continuized Nesterov Algorithm (CNA). We insist that in contrast with the classic ODE framework, where (NM) is a discrete approximation of (NMO), (CNA) is an exact evaluation of (CNE). This implies that the convergence results obtained for the solution of (CNE) almost **automatically transfer** to the aforementioned discretization. Such convergence results for the solution of (CNE) can be obtained with continuous-time stochastic Lyapunov analysis, using Itô formulas. These analyses share similarities with the continous time Lyapunov analysis of the ODE (NM).

The authors in (Even et al., 2021) apply this system to (strongly) convex, $L$-smooth functions, and show the continuized version of the Nesterov momentum algorithm recovers, in expectation, the same convergence speed as the classical Nesterov algorithm. Their analysis can be extended to the case of strong quasar convexity, deducing a convergence result of the form $\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] = \mathcal{O}((1 - \tau\sqrt{\frac{\mu}{L}})^k)$ (Wang and Wibisono, 2023). This is exactly the result based on a stochastic parameterization of (NM) mentioned in the paragraph "Discussion on the Mechanisms to Achieve the Bound $\mathcal{O}((1 - \tau\sqrt{\frac{\mu}{L}})^k)$" of Section 1.1.2; the stochastic parameterization arises from the Poisson measures. Surprisingly, the continuized version avoids running subroutines at each iteration to compute a parameter, as needed by existing works, see Section 1.1.2.

**Discussion** Apart from the specific field of decentralized asynchronous algorithms (Nabli et al., 2023; Nabli and Oyallon, 2025), there exist no other applications of the continuized process for optimization. It is therefore unclear whether this method can bring genuine new benefits, or whether it should rather be viewed as an original curiosity that either recovers known results in the convex case, or fortuitously yields a new result in the specific setting of strong quasar convexity. Also, for this latter case, it is not immediately clear why the randomization induced by the Poisson process allows to avoid the subroutine procedures of existing works. It raises the question:

**Question: Can the continuized method help achieve new acceleration results in non-convex optimization ?**

As we will see later, one of our contributions is simply to place the observations made in the previous paragraphs side by side and to interpret them from the right perspective. Doing so, it will be clear that the answer to this question is positive. In fact, it will help us address some of the limitations mentioned in previous sections.

## 1.3 Contributions

We summarize our contributions to address the questions raised in the previous sections.

### Chapter 2: Nesterov Momentum with Batch Estimator: Effect of Gradient Correlation on Acceleration

Chapter 2 is devoted to answering the question raised in Section 1.1.3, which we recall to be

**In the convex and interpolation setting, can we provide a clearer characterization of what can make** (SNM) **accelerate over** (SGD) **?**

Following the discussion in Section 1.1.3, we recall that to obtain acceleration with (SNM), the (SGC) value $\rho_K$ should not be too large. Our central contribution is to provide a new characterization of $\rho_K$. The limit of the existing characterization can appear on a simple example: consider the function $f(x) = \frac{1}{2}\Big(\frac{\mu}{2}\langle e_1, x\rangle^2 + \frac{L}{2}\langle e_2, x\rangle^2\Big)$, with $0 < \mu < L$ and $e_1, e_2$ standard basis vectors. This function satisfies interpolation, it is a sum of two $L$-smooth functions with $L_{\max} = L$, and it is $\frac{\mu}{2}$-strongly convex, so it is $\frac{\mu}{2}$-PL. Following the existing characterization of (Vaswani et al., 2019) (discussed in Section 1.1.3), $f$ satisfies (SGC) with $\rho_1 = 2\frac{L}{\mu}$. Playing with $\mu$ and $L$ can make $\rho_1$ be arbitrarily large. However, by developing $\|\nabla f(x)\|^2$, direct computations show that (SGC) is actually verified for $\rho_1 = 2 < 2\frac{L}{\mu}$. In this case, this existing characterization $\rho_1 = 2\frac{L}{\mu}$ can be arbitrarily sub-optimal. This idea of developing $\|\nabla f(x)\|^2$, which amounts to exploiting the finite-sum structure, can be done with an arbitrary function $f = \frac{1}{N}\sum_{i=1}^N f_i$:

$$\|\nabla f(x)\|^2 = \frac{K}{N}\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] + \frac{2}{N^2}\frac{N-K}{N-1}\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle.$$

On the right-hand side, the first term that appears is a variance term of the estimator, which is an important component of (SGC). Whether this assumption is verified or not relies on the second term. This term can be interpreted as an *average correlation* between the gradient of functions that compose $f$. If it is non-negative, then (SGC) is verified for $\rho_K = K/N$. This constant depends on the batch size $K$ and the number of functions $N$, it does not depend on any geometrical constants. In fact, the equality above only uses the differentiability of each $f_i$. We can generalize this observation, introducing a quantification of the average correlation.

**Definition 1.3.1.** *We say that $f$ verifies the Relaxed Averaged COrrelated Gradient Assumption (RACOGA) if there exists $c \in \mathbb{R}$ such that the following inequality holds:*

$$\forall x \in \mathbb{R}^d \backslash \overline{\mathcal{X}},\ \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2} \geq c, \tag{RACOGA}$$

*where $\overline{\mathcal{X}} = \{x \in \mathbb{R}^d, \forall i \in \{1, \dots, N\}, \|\nabla f_i(x)\| = 0\}$.*

The parameter $c$ quantifies a non-negative or non-positive correlation, and the case $c = 0$ reduces to the setting discussed just before. Then, our main result is the following.

**Theorem 1.3.1.** *Assume that* (RACOGA) *is verified for $c > -\frac{1}{2}$. Then, for batch size $K$, $f$ verifies* (SGC) *with $\rho_K$ that verifies*

$$\rho_K \leq \frac{1}{K(N-1)}\left(\frac{N(N-K)}{1+2c} + (K-1)N\right). \tag{1.11}$$

Here, the $\rho_K$ constant depends on the batch-size $K$, the number of functions $N$, and the correlation measure $c$. This result indicates that a high enough correlation between gradients is what allows acceleration of (SNM) over (SGD) in the (strongly) convex setting. We note that in this work, we were interested in the convergence bounds of (strongly) convex functions, yet our characterization could be applied to a broader setting, as it does not require convexity to hold.

As a byproduct of Theorem 1.3.1, we will also be able to deduce a link between correlation (the RACOGA value) and how increasing batch size impacts the convergence bounds. In the case $c < 0$, it appears that theoretically, using mini-batch is not cheaper computationally, as using full batch ($K = N$) leads to fewer gradient evaluations needed to achieve a given precision. These findings are supported by numerical experiments. First, we design a linear regression problem where the (RACOGA) values are close to zero, and indeed observe no acceleration of (SNM) compared with (SGD), and no computational gain by using mini-batch. Then, we observe that when training some neural networks, the (RACOGA) values $c$ along the optimization path are not only non-negative, but also relatively large, which coincides with the observed benefit of momentum. The content of this chapter is based on the following paper.

J. Hermant, M. Renaud, J-F. Aujol, C. Dossal, and A. Rondepierre.
**Gradient correlation is a key ingredient to accelerate SGD with momentum**.
In *The Thirteenth International Conference on Learning Representations*, 2025

## Chapter 3: Study of Nesterov Momentum for Strongly Quasar Convex Optimization and the Continuous-Discrete Rupture

In the next three chapters, we turn on studying non-convex relaxations of strong-convexity, specifically the strongly quasar convex functions. In Chapter 3, we mainly address the following question raised in Section 1.1.2:

**Can we achieve the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ with a fixed in advance deterministic parameterization of** (NM) **?**

Our answer is **partially positive**. We find that it is true as long as we assume a supplementary structure, apart from strong quasar convexity and $L$-smoothness. The key property is the following.

**Definition 1.3.2.** *For some $L > 0$ and $a \leq L$, the function $f$ is said to be $(a, L)$-curvatured if it satisfies for all $x, y \in \mathbb{R}^d$:*

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{a}{2}\|x-y\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x-y\|^2. \tag{1.12}$$

$(a, L)$-curvatured functions also correspond to $a$-weakly convex, $L$-smooth functions. $L$-smoothness (1.4) is the same assumption, with $a = -L$. $(a, L)$-curvatured functions allow an asymmetric characterization of the lower and upper bounds. If $a$ is such that $a > -L$, it can be understood as imposing a more severe restriction on the non-convexity of $f$. In particular, for $a = 0$, it reduces to convex $L$-smooth functions. Our convergence result is the following.

**Theorem 1.3.2.** *Let $f$ be a $(\tau, \mu)$-strongly quasar convex function and $(a, L)$-curvatured function such that $a \geq -\tau\sqrt{\mu L}$. Then, $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ generated by* (NM) *with constant parameters $\gamma_k = \frac{1}{L}$, $\alpha_k = \frac{1}{1+\sqrt{\frac{\mu}{L}}}$, $\beta_k = 1 - \tau\sqrt{\frac{\mu}{L}}$, $\gamma'_k = \frac{1}{\sqrt{\mu L}}$ ensures*

$$f(\tilde{x}_k) - f^* = \mathcal{O}\left(1 - \tau\sqrt{\frac{\mu}{L}}\right)^k. \tag{1.13}$$

Therefore, we obtain the bound $\mathcal{O}(1-\tau\sqrt{\frac{\mu}{L}})^k$ of existing results, deterministically, and without subroutines, but up to a supplementary assumption on the lower curvature.

One can wonder if such a restriction also appears when considering the ODE version of the algorithm. Algorithm (NM), with parameters prescribed as in the previous theorem, has the following high-resolution limit-ODE (Shi et al., 2022):

$$\left(1 + \frac{1-\tau}{2}\sqrt{\mu\gamma}\right)\ddot{x}(t) + (1+\tau)\sqrt{\mu}\dot{x}(t) + \sqrt{\gamma}\nabla^2 f(x(t))\dot{x}(t) + (1+\tau\sqrt{\mu\gamma})\nabla f(x(t)) = 0, \tag{1.14}$$

for some parameter $\gamma > 0$. This is a one-trajectory variant version of (NMO), see Proposition 4.D.1. We will show (Theorem 3.4.2) that for $f$ strongly quasar convex and $L$-smooth, the solution of this system verifies $f(x(t)) - f^* = \mathcal{O}(e^{-\tau\frac{\sqrt{\mu}}{2}t})$. This is to compare with $f(x(t)) - f^* = \mathcal{O}(e^{-\frac{\sqrt{\mu}}{2}t})$ obtained for the high-resolution

ODE associated to the algorithm in the strongly convex, $L$-smooth case (Shi et al., 2022). In this continuous case, similar convergence bounds are achieved in the strongly convex and strongly quasar convex cases, up to the $\tau$ parameter, and strikingly, **without any need of curvature assumptions**. This suggests that something is lost during the discretization process. We provide heuristic evidence that non-convexity is precisely what hurts by comparing side-by-side Lyapunov functions in the continuous and discrete cases, see Section 3.4.2. With this perspective, the subroutine used by (Hinder et al., 2020) can be seen as a way to adaptively discretize the ODE. As we already discussed, using the continuized Nesterov method allows us to achieve the bound of Theorem 1.3.2 without a subroutine or curvature restriction, and with a fixed in advance, stochastic parameterization of (NM). This would somewhat indicate that **the continuized Nesterov method avoids the limitation caused by the discretization process**. Investigating this method is the subject of the next chapter. The content of Chapter 3 is based on the following paper.

> J. Hermant, J-F. Aujol, C. Dossal, and A. Rondepierre.
> **Study of the behavior of Nesterov Accelerated Gradient in a nonconvex setting: the strongly quasar convex case**.
> In *Optimization*, 2026

## Chapter 4: Continuized Nesterov Acceleration for Non-Convex Optimization: Application to Strongly Quasar Convex Functions

Chapter 4 is motivated by the observations stated in the previous paragraph: between the continuous and the discrete case, discrepancies in the result we can obtain seem to be caused by non-convexity. In the strongly quasar convex case, the continuized Nesterov method, which we can view as an alternative, seems to circumvent this limitation at the cost of non-deterministic results. The reason is not necessarily clear. We recall the formulation of the continuized Nesterov equations

$$\begin{cases} dx_t &= \eta_t(z_t - x_t)dt - \gamma_{t^-}\nabla f(x_{t^-})N(dt) \\ dz_t &= \eta'_t(x_t - z_t)dt - \gamma'_{t^-}\nabla f(x_{t^-})N(dt) \end{cases} \tag{1.15}$$

where $N(dt) = \sum_{k\geq 0}\delta_{T_k}(dt)$ is a Poisson point measure. We recall the question raised in Section 1.2:

**Can the continuized method help achieve new non-convex acceleration results ?**

One of the most important aspects of (1.15) is that even though it is a continuous-time process that can be studied with stochastic continuous-time Lyapunov analysis, it can be accessed exactly through an algorithm, see Section 1.2. With this in mind, our main insight is the following.

**Since intuitively $\mathbb{E}[dN_t] = dt$, then the mean process $(\mathbb{E}[x_t], \mathbb{E}[z_t])_{t\geq 0}$ is smooth.**

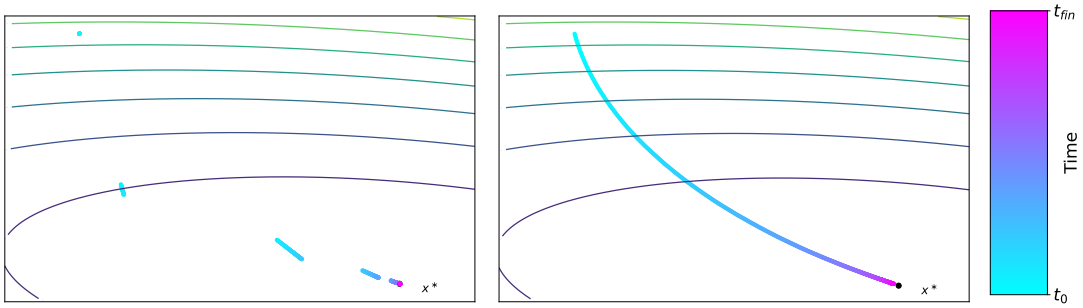



Figure 1.2: Left: Visualization of a single trajectory of $(x_t)_{t\geq 0}$ satisfying (1.15), between $t_0$ and $t_{\text{fin}}$. Right: Visualization of the average over $10^5$ such trajectories, providing an estimate of the mean process $(\mathbb{E}[x_t])_{t\geq 0}$ for which the jumps are smoothed out. This figure is taken from Chapter 4.

This observation captures the promise of a powerful alternative. Indeed, ODEs such as (NMO) are *idealized objects*, in the sense that they access the information given by the gradient of $f$ at any continuous time $t \in \mathbb{R}_+$. This idealized nature is precisely what makes them inaccessible in practice: one can only

approximate them *e.g.* by means of Euler discretization. The dynamics (1.15) lie somewhere in between this idealized world and our real, practical world. They are not *idealized objects* in the same sense as ODEs, since they access gradients only at discrete times. However, they recover such an idealized behavior once expectations are taken. In this sense, and with some abuse of language, the solution of (1.15) can be viewed as an algorithm whose expected dynamics are governed by an ODE. This naturally suggests that, at the cost of non-deterministic results, the continuized process may circumvent some of the limitations caused by discretization. Motivated by this potential, we provide some generalization of the tools required to carry out convergence analyses with this process. Then, we use it to address two questions raised in Section 1.1.2 concerning strongly quasar convex functions. We first focus on the following.

**Can we improve upon the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ ?**

Inspired by a Lyapunov analysis carried in the ODE case (Aujol et al., 2022), we obtain that for some $\varepsilon \in (0,1)$ such that $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$, with probability $1 - \frac{1}{c_0} - e^{-(c_1-1-\log(c_1))k}$, for some $c_0 > 1$, $c_1 \in (0,1)$, we achieve a bound of the form

$$f(\tilde{x}_k) - f^* = \mathcal{O}\big(\frac{c_0}{\varepsilon} e^{-(1-\varepsilon)\tau\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}c_1 k}\big).$$

If the exponent in the exponential is small, this is roughly similar to the bound

$$\mathcal{O}\left(\frac{c_0}{\varepsilon}\left(1-(1-\varepsilon)c_1\tau\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}\right)^k\right).$$

In the case $\tau = 1$, this becomes $\mathcal{O}\left(\frac{c_0}{\varepsilon}\left(1-(1-\varepsilon)\sqrt{2}c_1\sqrt{\frac{\mu}{L}}\right)^k\right)$. This is close to an improvement by a factor $\sqrt{2}$, which is the one that was achieved in the ODE case (Aujol et al., 2022). Our main limitations are the $(1-\varepsilon)$ and $1/\varepsilon$ factors; it is unclear whether they can be avoided. We also note that it is unclear whether this bound can be achieved without the continuized method. Indeed, a careful look at the works achieving the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$, whether the one based on a subroutine (Hinder et al., 2020) or using curvature restriction (Chapter 3), shows that their analysis crucially rely on exploiting the same negative squared norm term of the form $\mathcal{O}(-\|\tilde{x}_k - \tilde{y}_k\|^2)$. This specific term does not appear in the analysis with the refined choice of Lyapunov function that allows the bound to be improved, so these analyses may no longer work.

We turn now to the other question:

**Can we achieve the bound $\mathcal{O}((1-\tau\sqrt{\frac{\mu}{L}})^k)$ while relaxing the uniqueness of minimizer property beyond the affine-set assumption ?**

We recall that to address this question, we replace the strong quasar convexity structure as defined in (1.6) by the following

$$\forall x \in \mathbb{R}^d,\ f^* \geq f(x) + \langle \nabla f(x), \pi(x) - x\rangle + \frac{\mu}{2}\|x - \pi(x)\|^2,$$

where $\mu > 0$ and $\pi$ is the orthogonal projection onto $\operatorname{argmin} f$. The affine-set assumption requires $\operatorname{argmin} f$ to be an affine set. It was used in (Gupta and Wojtowytsch, 2025) to achieve the bound $\mathcal{O}((1-\sqrt{\frac{\mu}{L}})^k)$ with Nesterov Momentum algorithm up to a curvature restriction similar to the one discussed in the previous chapter. Interestingly, **they can achieve the desired result in the ODE case without the affine-set restriction**. This is another instance of discrepancies between the ODE and algorithm settings. We relax the affine-set property by having that $\arg\min f$ is a convex set. This is the most general assumption we can make to have that $\pi$ is globally well defined: convexity is necessary to have that the orthogonal projection is well defined on the whole space $\mathbb{R}^d$, in the sense that it maps a point on $\mathbb{R}^d$ to a unique point in $\operatorname{argmin} f$ (Kritikos, 1938). Under this assumption (and another about the smoothness of the boundary of $\operatorname{argmin} f$), we ensure that with probability $1 - \frac{1}{c_0} - e^{-(c_1-1-\log(c_1))k}$ for some $c_0 > 1$, $c_1 \in (0,1)$, we have

$$f(\tilde{x}_k) - f^* = \mathcal{O}\left(c_0 e^{-\left(1-\frac{1+\sqrt{5}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right)\sqrt{\frac{\mu}{L}}c_1 k}\right).$$

If the exponent of the exponential is small enough, it is roughly similar to

$$\mathcal{O}\left(c_0\left(1-c_1\left(1-\frac{1+\sqrt{5}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right)\sqrt{\frac{\mu}{L}}\right)^k\right).$$

If $\mu/L$ is small enough, we have that $\left(1 - \frac{1+\sqrt{5}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right) \approx 1$, such that we obtain the accelerated rate. This provides a positive answer in the case $\tau = 1$, with high probability. A difficult aspect of the proof is that the Itô formula designed to carry the Lyapunov analysis with (CNE) was originally stated for smooth functions. Our Lyapunov function, however, involves the projection $\pi$, which is not differentiable on the boundary of $\operatorname{argmin} f$. We solved this issue by designing a Lyapunov analysis tool that handles non-smooth Lyapunov functions, only requiring local Lipschitzness. We note finally that the convergence results presented in this paragraph actually hold when using stochastic gradients under the strong growth condition (SGC). The content of this chapter is based on the following paper.

J. Hermant, J-F. Aujol, C. Dossal, L. Huang and A. Rondepierre.
**Continuized Nesterov Acceleration for Non-Convex Optimization**.
*arXiv preprint*, 2025

## Chapter 5: Reconsideration of Acceleration for Strongly-Quasar Convex functions, Existing Flaws and Alternative

In the two previous chapters, we provided new results for optimization of $L$-smooth, strongly quasar convex functions with (NM). Analogies with convergence bounds of the strongly-convex, $L$-smooth setting, naturally lead to interpret these bounds as accelerated. Indeed, (GD) ensures the bound $f(\tilde{x}_k) - f^* = \mathcal{O}\left(\left(1 - \tau\frac{\mu}{L}\right)^k\right)$ while (NM) ensures the bound $f(\tilde{x}_k) - f^* = \mathcal{O}\left(\left(1 - \tau\sqrt{\frac{\mu}{L}}\right)^k\right)$, such that it seems that we improve the dependence on the conditioning from $\mu/L$ to $\sqrt{\mu/L}$. However, in Section 1.1.2, we briefly discussed the fact that interpreting $\mu/L$ as a conditioning measurement, as in the smooth strongly convex case, might be a misleading shortcut. With this observation, we raised the following question.

**Are strongly quasar convex functions truly benefiting from a similar acceleration phenomenon compared with strongly convex functions ?**

We provide a negative answer in Chapter 5. The key is the following observation: for strongly quasar convex functions, it is indeed natural to compare the bounds of (GD) and (NM) obtained for this class, namely $\mathcal{O}\left(\left(1 - \tau\frac{\mu}{L}\right)^k\right)$ and $\mathcal{O}\left(\left(1 - \tau\sqrt{\frac{\mu}{L}}\right)^k\right)$. However, strongly quasar convex functions are also PL functions, for some parameter. In some situations, **the bound derived for** (GD) **under PL may actually be sharper**. A telling piece of evidence to support a negative answer can be derived from the literature of smooth PL functions. In (Yue et al., 2023) are constructed lower-bound functions that are $\mu$-PL and $L$-smooth, such that the convergence bound is at best $\left(1 - c\frac{\mu}{L}\right)^k$ with $c$ a numerical factor. This lower bound discards the possibility of improving the dependence in $\mu/L$ for the class of smooth PL functions in general. These lower bound functions actually benefit from important regularity, such as a unique minimizer, and more, from **strong quasar convexity** (Proposition 5.3.1).

**Proposition 1.3.1.** *The $\mu$-PL, $L$-smooth, lower-bound function from (Yue et al., 2023), is also strongly quasar convex, for some parameters $(\tau, \mu') \in (0, 1] \times \mathbb{R}_+^*$.*

This result shows that an acceleration of the same nature as for the strongly convex case is not possible in general when considering strongly quasar convex functions. Indeed, in the strongly case, the improvement from $\mu/L$ to $\sqrt{\mu/L}$ from (GD) to (NM) is such that its ratio $\sqrt{\mu/L}/(\mu/L) = \sqrt{L/\mu}$ goes to infinity with the conditioning $L/\mu$ going to infinity. There is no such behavior for strongly quasar convex functions.

The previous discussion suggests that our understanding of acceleration under strong quasar convexity remains incomplete. In particular, the fact that bounds derived under strong quasar convexity sometimes fail to improve upon those obtained under the PL condition suggests that this assumption may contain a hidden parameter governing whether acceleration is possible. A way to identify such a hidden structure is to answer the last question raised in Section 1.1.2, namely:

**Up to the uniqueness of the minimizer property, can we provide an intuitive characterization of what supplementary structure benefits strongly quasar convex functions compared with PL functions ?**

Our answer can be simplified as follows.

**Theorem 1.3.3.** *Suppose $f$ is a $\mu$-PL, $L$-smooth function with a unique minimizer. There exist parameters such that $f$ is strongly quasar convex if and only if there exists some $a > 0$ such that $f$ satisfies the following* ***aiming condition****:*

$$\forall x \in \mathbb{R}^d, \quad 1 \geqslant \frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\| \|x - x^*\|} \geqslant a > 0. \tag{AC}$$

The aiming condition is rather intuitive: parameter $a$ measures the alignment between the descent direction $-\nabla f$ and the direction toward the minimizer. A non-negative value indicates that following the direction $-\nabla f$ brings us toward the minimizers, while a non-positive value indicates it brings us in the opposite direction. Having a strong alignment ($a$ close to 1) is particularly favorable to momentum methods, as it ensures that the accumulated momentum reinforces progress toward $x^*$. In contrast, if $a$ is small, momentum may amplify steps taken in poorly aligned directions.

Now, we may use the structure (AC) together with the structure PL and $L$-smooth to study the convergence of (NM). Using the continuized version (CNA), we obtain that for such functions with probability $1 - \frac{1}{c_0} - \exp(-(c_1 - 1 - \log(c_1))k)$, for some $c_0 > 1$ and $c_1 \in (0, 1)$ we have

$$f(\tilde{x}_k) - f^* = \mathcal{O}\big(e^{-a\left(\frac{\mu_0}{L_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}(1-c_1)k}\big),$$

where $\mu_0$ and $L_0$ are quadratic growth constant, namely verify $f(\tilde{x}_k) - f^* \geq \frac{\mu_0}{2}\|x - x^*\|^2$ and $f(\tilde{x}_k) - f^* \leq \frac{L_0}{2}\|x - x^*\|^2$ for any $x \in \mathbb{R}^d$. Such inequalities are implied by $\mu$-PL and $L$-smooth, respectively, and satisfy $\mu_0 \geq \mu$ and $L_0 \leq L$ (Karimi et al., 2016). Because this bound is obtained under the PL assumption, it is directly comparable to the bound of (GD) obtained for PL, $L$-smooth functions. With few transformations and simplifications, we get that the convergence bound is better with (NM) as long as the aiming condition constant $a$ is large enough, namely if

$$a \;\geq\; \left(\frac{L_0}{\mu_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}.$$

This indicates an interplay between the conditioning $\mu/L$ and the aiming condition constant. By running (GD) and (NM) on the lower bound functions from (Yue et al., 2023), we first observe empirically that (NM) indeed does not accelerate over (GD), and secondly we observe heuristically that the left and right-hand side of the above inequality scale similarly on this function, which supports the relevance of the bound. The content of this chapter is based on the following paper.

J. Hermant.
**Acceleration for Polyak-Łojasiewicz Functions with a Gradient Aiming Condition**.
*arXiv preprint*, 2026

## Chapter 6: Application of the Continuized Nesterov Momentum Method to Functions with Lipschitz Gradient and Hessian: Achieving the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity without Momentum-Control Mechanisms

Chapter 6 is devoted to a new example of setting where the continuized method can bring new results. In the setting discussed in Section 1.1.1, we address the question:

**Are momentum-control mechanisms fundamentally necessary to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity?**

Once more, we start our investigation by observing an existing result in the ODE world. Authors of (Okamura et al., 2024) consider the Heavy-Ball equation:

$$\ddot{x}_t + \alpha \dot{x}_t + \nabla f(x_t) = 0, \tag{HB}$$

which can be understood as a special case of (NMO). They show that under the Lipschitz gradient and Hessian structure, if $(x_t)_{t\geq 0}$ is a solution of (HB) with $\alpha = \mathcal{O}(T^{-1/7})$, then defining for $t \geq 0$ the following average

$$\overline{x}_t = \int_0^t w_t(s) x_s ds \tag{1.16}$$

with $w_t(s) := \frac{\alpha e^{\alpha s}}{e^{\alpha t}-1}$, one has that we find a point $\hat{x}$ such that $\|\nabla f(\hat{x})\| \leq \varepsilon$ in at most $\mathcal{O}(\varepsilon^{-7/4})$ running time. This can be seen as a continuous version of the $\mathcal{O}(\varepsilon^{-7/4})$ bound on the number of iterations of a given algorithm, see Section 1.1.1. We adapt their proof strategy to the continuized framework. For example, inspired by their analysis, we introduce a Poisson averaging of $(x_t)_{t\geq 0}$ solution of (CNE)

$$\overline{x}_t := \int_0^t w_t(s) x_{s^-} dN_s, \quad \text{where } w_t(s) = \frac{\alpha e^{\alpha s}}{\int_0^t \alpha e^{\alpha s'} dN(s')}. \tag{1.17}$$

The stochasticity induced by the Poisson process complicates the analysis. Nonetheless, we could achieve the following result, which we state informally.

**Theorem 1.3.4** (Informal version)**.** *Let $f$ be with Lipschitz gradient and Hessian, and $\tilde{\varepsilon} \in (0, 1/2)$. For a number of iterations $n$ large enough,* (CNA) *with a suitable stochastic parameterization outputs a point $\tilde{x}$*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \|\nabla f(\tilde{x})\|\right] = \mathcal{O}(n^{-4/7}),$$

*where $\mathcal{A}_n$ is a subset of the realizations that satisfies $\mathbb{P}(\mathcal{A}_n) \geq 1 - \tilde{\varepsilon}$.*

Using the continuized method allows us to obtain that a stochastic parameterization of (NM) is able to achieve a point that satisfies $\mathbb{E}\left[\|\mathbb{1}_{\mathcal{A}_n} \nabla f(\tilde{x})\|\right] \leq \varepsilon$ in at most $\mathcal{O}(\varepsilon^{-7/4})$ gradient evaluations. Specifically, no momentum-reset mechanisms are required. This contribution is technically the most demanding of the manuscript. Most of the chapter is devoted to the proof of this theorem. Its content is based on the following paper.

> J. Hermant, J-F. Aujol, C. Dossal, L. Huang, A. Rondepierre and I. Waldspurger.
> **Continuized Nesterov Momentum Achieves the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity in Smooth Nonconvex Optimization**.
> *arXiv preprint*, 2026

## Other Contributions

We chose not to discuss the following contributions in this manuscript; we instead simply list them here. The sign "$*$" denotes equal contributions.

> M. Renaud*, J. Hermant*, D. Wei*, Y. Sun.
> **Provably Accelerated Imaging with Restarted Inertia and Score-based Image Priors**.
> In *The Thirteenth International Conference on Learning Representations*, 2026

> M. Renaud, J. Hermant, N. Papadakis.
> **Convergence Analysis of a Proximal Stochastic Denoising Regularization Algorithm**.
> In *Scale Space and Variational Methods in Computer Vision*, 2025

> E. Gbaguidi*, J. Hermant*.
> **Accelerated Zeroth-Order Quasar Convex Optimization**.
> *arXiv preprint, 2026*

## Acknowledgments

This work was supported by the ANR project PEPR PDE-AI. Part of this work was supported by the ANR MICROBLIND (grant ANR-21-CE48-0008), the ANR Masdol (grant ANR-19-CE23-0017), the ANR SOS2ID (grant ANR-24-CE40-3786) and the French Direction Générale de l'Armement. Experiments presented in Chapter 2 were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Universite de Bordeaux, Bordeaux INP and Conseil Regional d'Aquitaine (see https://www.plafrim.fr).

## 1.4 Appendices of Introduction

### 1.4.1 Convexity and Quasar Convexity

In the remainder of the section, we chose to elaborate on (strongly) quasar convex functions, as they might not be very well known, particularly compared with Polyak-Łojasiewicz functions (PL). Because it is a relaxation of (strong) convexity, we believe it might be interesting to start this section with reminders about convexity. The interplay between convexity and quasar convexity will appear clearly. Also, when considering convergence results of a given algorithm applied to quasar convex functions, it might be interesting to have in mind what results can be achieved in the convex case. In some situations, it appears that we achieve similar results, despite the relaxation of structure. Finally, we recall that unless stated otherwise, $f$ is always assumed to be differentiable.

#### Convex Functions

There exists a large body of theory about the minimization of convex functions (Rockafellar, 1997; Ben-Tal and Nemirovski, 2001; Boyd and Vandenberghe, 2004; Nesterov, 2013). These functions provide a framework that generalizes classical problems, such as quadratic minimization in unconstrained optimization or linear programming in constrained optimization, for which many efficient optimization algorithms have been designed. We now state different characterizations of convexity, depending on the level of information they require.

**Definition 1.4.1.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if it satisfies*

$$\forall x, y \in \mathbb{R}^d, \lambda \in [0,1], f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{1.18}$$

*It is equivalent to verifying*

$$\forall x, y \in \mathbb{R}^d, f(x) + \langle \nabla f(x), y - x \rangle \leq f(y). \tag{1.19}$$

*If $f$ is twice-differentiable, it is equivalent to verify*

$$\forall x, y \in \mathbb{R}^d, \left\langle \nabla^2 f(x) y, y \right\rangle \geq 0. \tag{1.20}$$

*In the case where all the above inequalities are strict, we say these functions are strictly convex.*

Each characterization emphasizes a different interpretation. The zeroth-order characterization (1.18) is equivalent to have that the epigraph of $f$, defined as $\text{epi}(f) = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq t\}$, is a convex set, namely for all $\lambda \in [0,1]$, $\lambda u + (1-\lambda)v \in \text{epi}(f)$ as long as $u, v \in \text{epi}(f)$. The first-order characterization (1.19) means that all tangent planes of $f$ lie below its graph. Finally, the second-order characterization (1.20) means that all the eigenvalues of the Hessian matrix of $f$ are nonnegative, or, in other words, it has nonnegative curvature at any point.

With convexity, we deduce interesting global information from local information. For example, this can be seen with (1.19): the sole knowledge of $(f(x), \nabla f(x))$ for some $x \in \mathbb{R}^d$ ensures the whole graph of $f$ lies over a hyperplane. Such a lower-bound may be of interest as we want to minimize $f$. Another fundamental example of such regularity is given by the following classical property. While it can be deduced directly from (1.19), it has very important consequences from an optimization point of view.

**Theorem 1.4.1.** *Let $f$ be convex. Let $x^* \in \mathbb{R}^d$ such that $\nabla f(x^*) = 0$. Then, $x^*$ is a global minimizer of $f$, i.e. $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^d$.*

For general functions, when an algorithm finds a local minimizer $\hat{x}$, this algorithm cannot ensure that there does not exist another minimizer $\overline{x}$ such that $f(\overline{x}) < f(\hat{x})$, except with some precise knowledge about the structure of the function such as its minimum value $\min_{x \in \mathbb{R}^d} f(x)$. Convexity precludes such situations.

**Example of convex problems** We give a few examples of optimization problems that write as the minimization of a convex function.

- **Regularized Linear Least Square.** A classical example is the minimization of $f(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda r(x)$, where $A \in \mathbb{R}^{n\times d}$ is a design or observation matrix and $b \in \mathbb{R}^n$ is a vector of observations, $\lambda \geq 0$ a regularization parameter and $r(\cdot)$ is a convex regularization function, classical examples being $r(x) = \|x\|_p^2$ for some $p \in [1, +\infty]$.
- **Logistic Regression.** Another important example is logistic regression, where for given data $(a_i, b_i)_{i=1}^n$, with features $a_i \in \mathbb{R}^d$ and labels $b_i \in \{-1, 1\}$, the logistic loss is defined by $f(x) = \frac{1}{n}\sum_{i=1}^n \log\big(1 + \exp(-b_i \langle a_i, x\rangle)\big)$.
- **Convexification of more complicated models.** Due to the favorable properties and efficient methods of convex optimization, there exists a line of work about "convexification" of originally non-convex problems. For example, in the case of neural networks, letting the number of neurons grow to infinity can make a convex optimization problem appear (Bengio et al., 2005; Bach, 2017; Chizat and Bach, 2018).

Beyond the example of convex problems, convexity provides a framework for the analysis of optimization algorithms. As it may offer strong convergence guarantees, it can serve as a natural test bed for the study of new methods. Even when an algorithm is ultimately designed for nonconvex problems, analyzing its behavior under convexity may be a useful first step towards understanding its theoretical properties. It also allows one to assess how the method performs in favorable regimes.

**Strongly Convex Functions** An important subclass of convex functions is the class of $\mu$-strongly convex functions, with some $\mu > 0$. These functions can be written in the form $g(x) + \frac{\mu}{2}\|x\|^2$, for some $g$ convex, which gives a direct interpretation of this class as convex functions that grow quadratically. We can provide a characterization of this class in a similar fashion to Definition 1.4.1.

**Definition 1.4.2.** *Let $\mu > 0$. A function $f$ is $\mu$-strongly convex if it satisfies*

$$\forall x, y \in \mathbb{R}^d,\ \forall \lambda \in [0, 1], \qquad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x - y\|^2 . \tag{1.21}$$

*This is equivalent to*

$$\forall x, y \in \mathbb{R}^d, \qquad f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2 \leq f(y). \tag{1.22}$$

*If $f$ is twice differentiable, this is equivalent to*

$$\forall x, y \in \mathbb{R}^d, \qquad \left\langle \nabla^2 f(x) y \right\rangle y \geq \mu \|y\|^2 . \tag{1.23}$$

*When such a constant $\mu > 0$ exists, we simply say that $f$ is strongly convex.*

More than having Theorem 1.4.1, these functions have a unique minimizer. A quadratic function whose Hessian matrix has only positive eigenvalues will be $\mu$-strongly quasar convex, where $\mu$ is the smallest of these eigenvalues. As we discussed in Section 1.1.2, the additional structure of these functions compared with convexity leads to fundamentally different convergence results of algorithms.

### Quasar Convexity

There exist several relaxations of convexity, such as quasi-convexity (Arrow and Enthoven, 1961; Daniilidis and Hadjisavvas, 1999; Grad et al., 2025) or pseudo-convexity (Mangasarian, 1975), to name a few classic examples. We refer to (Khademi, 2026) for a recent survey on the topic. An interest of quasar convex functions is that in some situations, it seems to benefit from similar convergence results as in the convex case. Even though we focused on strong quasar convexity and not on quasar convexity, we will discuss it there because both are connected.

**Quasar Convexity** Quasar convex functions are defined in a way that shows clear similarity with convexity (1.19).

**Definition 1.4.3.** *Let $\tau \in (0, 1]$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is $\tau$-quasar convex if it satisfies*

$$\forall x \in \mathbb{R}^d, x^* \in \operatorname{argmin} f, \lambda \in [0, 1], \quad f(\lambda x^* + (1 - \lambda)x) \leq \tau\lambda f(x^*) + (1 - \tau\lambda)f(x) \tag{1.24}$$

*This is equivalent to*

$$\forall x \in \mathbb{R}^d, x^* \in \operatorname{argmin} f, \quad f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle \leq f(x^*). \tag{1.25}$$

The equivalence between (1.24) and (1.25) can be found in (Hinder et al., 2020, Lemma 11). For the special case $\tau = 1$, classically named star-convexity, (1.24) and (1.25) can be obtained from the convex counterparts (1.18) and (1.19), by restricting $y \in \operatorname{argmin} f$. In this case, the epigraph is not convex, but star-convex with respect to the points $(x^*, f(x^*))$, namely for all $\lambda \in [0, 1]$, $\lambda u + (1 - \lambda)(x^*, f(x^*)) \in \operatorname{epi}(f)$ as long as $u \in \operatorname{epi}(f)$. For $\tau < 1$, star-convexity of the epigraph is not required anymore, and even more non-convex behavior is allowed, see Figure 1.3.

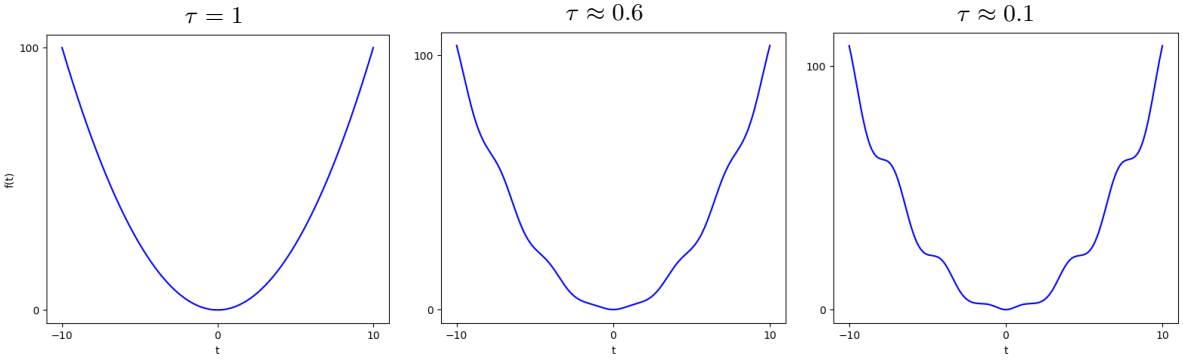



Figure 1.3: Examples of one-dimensional $\tau$-quasar convex functions. They are of the form $f(t) = (t + a\sin(bt))^2$, for some $a, b > 0$.

It becomes difficult to provide a second-order characterization. In particular, there is no reason to think that we can obtain a bound on $\langle \nabla^2 f(x)y, y\rangle$ that is uniform in $y$, as long as $x \notin \operatorname{argmin} f$. We will show in Chapter 3 that we can instead obtain characterizations that involve an average of $\nabla^2 f$ on segments that join the minimizer $x^*$ to points $x \in \mathbb{R}^d$. From (1.25), we can deduce the same result as Theorem 1.4.1.

**Theorem 1.4.2.** *Let $f$ be $\tau$-quasar convex. Let $x^* \in \mathbb{R}^d$ such that $\nabla f(x^*) = 0$. Then, $x^*$ is a global minimizer of $f$, i.e. $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^d$.*

**Examples of quasar convex functions** We give some examples of problems that can be written as the minimization of a quasar convex function. Naturally, it includes the examples given for convex functions in Section 1.4.1. There also exist examples where the function is not convex.

- **Learning linear dynamical systems.** The first known example arises in the training of a time-invariant linear dynamical system, that are of the form

$$\begin{aligned} h_{t+1} &= Ah_t + Bx_t, \\ y_t &= Ch_t + Dx_t + \xi_t, \end{aligned}$$

  where $\{x_t\}_{t\in\{1,\cdots T\}}$ and $\{y_t\}_{t\in\{1,\cdots T\}}$ are sequences, $\{h_t\}_{t\in\{1,\cdots T\}}$ are hidden state, $A$, $B$, $C$ and $D$ are linear transformations and $\{\xi_t\}_{t\in\{1,\cdots T\}}$ models noise, see the detailed definition of these objects in (Hardt et al., 2018). The authors assume access to examples of sequences $(x^{(i)}, y^{(i)})_{i=1}^n$ generated by

the above system, and set the model

$$\hat{h}_{t+1} = \hat{A}\hat{h}_t + \hat{B}x_t,$$
$$\hat{y}_t = \hat{C}\hat{h}_t + \hat{D}x_t,$$

where $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ are the parameters to learn. They then show that a relevant loss function to solve this problem satisfies the inequality (1.25) on a subset of the domain.

- **Generalized Linear Model.** Another example was given in the case of training a generalized linear model, namely

$$f(x) = \mathbb{E}_{a\sim\mathcal{D}}\left[\frac{1}{2}(\sigma(a^T x) - b)^2,\right]$$

where $\mathcal{D}$ is a distribution over $\mathbb{R}^d$, $\sigma : \mathbb{R} \to \mathbb{R}$ is a link-functions, and where there exists $x^* \in \mathbb{R}^d$ such that $b$ is generated with $b = \sigma(\langle a, x^*\rangle)$. The special case $\sigma = I_d$ corresponds to a linear model, and the associated $f$ is convex. (Wang and Wibisono, 2023) show that some choices of $\sigma$ that make $f$ quasar convex, at least on a domain including solutions or under assumptions on the distribution $\mathcal{D}$. Examples include the logistic function $\sigma(t) = 1/(1+e^{-t})$, Leaky-Relu $\sigma(t) = \max(\alpha t, t)$ for $\alpha \geq 0$ (including Relu with $\alpha = 0$) and quadratic $\sigma(t) = t^2$.

- **An Artificial Construction.** The idea of the following construction first appears in (Lee and Valiant, 2016), further generalized by (Hinder et al., 2020). Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $f$ is $\tau$-quasar convex and $\min_{\mathbb{R}} f(t) = f(0)$. Let $g : \mathbb{R}^d \to \mathbb{R}$ such that $g > 0$. Then, the following function

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right),\ x \in \mathbb{R}^d \tag{1.26}$$

is $\tau$-quasar convex. As this result holds even when $g$ is highly irregular, it shows that the quasar convexity property is not affected by strong oscillations that occur in a direction orthogonal to the one pointing to the minimizer. A visualization for $d = 2$ was provided in Figure 1.1.

**Brief History of quasar convexity** Just as convex functions naturally arise from convex sets through the epigraph characterization, star-convex functions can be traced back to the notion of star-convex sets. In the 1980s-1990s, related one-dimensional notions appeared, for instance under the name *stellarity* in (Toader, 1984), and under the name *starshaped functions* in (Toader, 1986; Mocanu et al., 1997; Toader, 1998). These functions were defined through inequalities of the form $tf(x) \geq f(tx)$, for $(t, x) \in [0, 1] \times \mathbb{R}$. If imposing $\min_x f(x) = f(0) = 0$, this coincides with star-convexity[5] in the sense of Definition 1.4.3 with $\tau = 1$. In optimization, this class was used in (Nesterov and Polyak, 2006, Section 4) for the analysis of a second-order method. Later, it was studied as a primary object in (Lee and Valiant, 2016), which presented star-convex functions as a relevant class for non-convex optimization. Independently, (Hardt et al., 2018) showed that learning linear dynamical systems leads to an objective satisfying a local form of $\tau$-quasar convexity, which they call $\tau$-weak-quasi-convexity. These works likely initiated the recent interest in quasar convex optimization, especially within the machine learning community. The terminology *quasar convex* was introduced in (Hinder et al., 2020), together with *strong quasar convexity*, whose relation to quasar convexity parallels that between strong convexity and convexity. This terminology has since become standard.

**Relevance of quasar convexity and convexity relaxation for more complicated functions** While relaxing convexity, quasar convexity may seem way too restrictive to encompass too complicated problems, such as the minimization problems arising in neural network training, which are, for example, expected to have non-global minimizers. An important characteristic of the classes we have introduced so far is that they are defined *globally*, namely, the underlying inequalities are required to hold on the whole space $\mathbb{R}^d$. Many problems indeed cannot be described globally by relaxations of convexity, such as quasar convexity. We emphasize that if the global geometry of a given function might be difficult to describe, optimization of such functions could be very efficient anyway. One of the reasons can be that the function is actually fairly

[5] We note that in some of these references the term *star-convexity* was used for a different notion.

regular (e.g., convex) on most of the space. For well-designed neural networks, the associated function to minimize has been observed to be "almost convex" around solutions (Li et al., 2018). Another reason can be that optimization algorithms will typically follow a path where the geometry of the function is favorable. For instance, along the optimization path of the stochastic gradient descent, the function appears empirically to satisfy an assumption very close to star-convexity (Zhou et al., 2019). We note also that optimization algorithms are observed to be able to find global minima in this context (Zhang et al., 2017), despite the possible existence of saddle-points or non-global minima.

**Strong quasar convexity**

In a similar fashion to how quasar convexity relaxes convexity, one can relax strong convexity.

**Definition 1.4.4.** *Let $\tau \in (0,1]$, $\mu > 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is $(\tau, \mu)$-strongly quasar convex if it satisfies*

$$\forall x \in \mathbb{R}^d, x^* \in \operatorname{argmin} f, \lambda \in [0,1], \quad f(\lambda x^* + (1-\lambda)x) + \lambda \left(1 - \frac{\lambda}{2-\tau}\right) \frac{\tau\mu}{2} \|x - x^*\|^2 \leq \tau\lambda f(x^*) + (1-\tau\lambda) f(x). \tag{1.27}$$

*This is equivalent to*

$$\forall x \in \mathbb{R}^d, \ f(x) + \frac{1}{\tau} \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2} \|x - x^*\|^2 \leq f(x^*). \tag{1.28}$$

The equivalence between (1.27) and (1.28) can be found in (Hinder et al., 2020, Lemma 11). The property of uniqueness of the minimizer holds as for strong convexity. Among the examples of quasar convex functions provided in the previous section, some generalized models are also strongly quasar convex, and the example of the artificial construction (1.26) also holds replacing quasar convexity with strong quasar convexity, as long as $f$ is strongly quasar convex and $g \geq 1$. Several results of the manuscript consider the class of strong quasar convexity, $L$-smooth functions, and we provide some new properties of this class in Chapter 3. We note that while adding a $\|x\|^2$ term to a convex function makes it strongly convex, it is not that simple in the quasar convex case. Generally speaking, for $f$ (strongly) quasar convex and $g$ (strongly) quasar convex, $f + g$ is not ensured, in general, to be (strongly) quasar convex. It is nonetheless ensured if they share the same minimizer (Hinder et al., 2020, Appendix D.3). Nonetheless, assuming that for all $x \in \mathbb{R}^d$, the following properties hold

$$\frac{\mu}{2} \|x - x^*\|^2 \leq f(x) - f^*, \text{ and } f(x) + \frac{1}{\tau} \langle \nabla f(x), x^* - x \rangle \leq f^*,$$

direct computations show that the function $f$ is $(\tau/2, \mu)$-strongly quasar convex. For the converse, a $(\tau, \mu)$-strongly quasar convex satisfies the two properties stated above. This means that strongly quasar convex functions can be interpreted as quasar convex functions that satisfy some quadratic growth property.

### 1.4.2 Comments on Algorithms

We provide some additional details and comments on (NM). Before doing so, we make a short comment on gradient methods in modern applications.

**On the resurgence of gradient descent** Not so long ago, gradient descent was considered a simple, yet slow algorithm. In an optimization book (Nocedal and Wright, 2006) first published in 1999, one can read (page 22):

> *"One advantage of the steepest[6] descent direction is that it requires calculation of the gradient $\nabla f$ but not of second derivatives. However, it can be excruciatingly slow on difficult problems."*

Indeed, there exist some arguably more sophisticated alternatives, such as second-order methods (*e.g.* Newton) or cutting-plane methods (Vaidya, 1996; Lee et al., 2015). Crucially, many of these methods become unusable when the dimension of the function to minimize becomes excessively large, which is the so called *large-scale problems*. This particularly happens in neural network optimization, whose number of parameters

[6]The term steepest refers to the fact that the stepsizes in (GD) are computed with an algorithmic procedure to maximize the decrease of the function along the line $\{\tilde{x}_k + \gamma \nabla f(\tilde{x}_k), \gamma \in \mathbb{R}\}$.

determines $d$ in (1.1). $d \sim 10^5$ is considered as corresponding to a small network, while state of the art models typically involve way more parameter and keep growing in size, see for instance GPT-3 in the case of large language models which involve 175 billion parameters (Brown et al., 2020). What makes the training (or optimization) of such big models possible is, in big part, fast differentiation techniques—notably reverse-mode automatic differentiation (backpropagation)— that have made, in many situations, the computation of the gradient costs at most four times the evaluation of the function itself (Griewank and Walther, 2008).

**Nesterov Momentum: Formulations, Interpretations and Names** Before Nesterov's momentum existed *Polyak's Momentum* algorithm, commonly named *Polyak's Heavy Ball* (Polyak, 1964). For an initialization $\tilde{x}_{-1}, \tilde{x}_0 \in \mathbb{R}^d$, it is defined as follows

$$\tilde{x}_{k+1} = \tilde{x}_k - \gamma_k \nabla f(\tilde{x}_k) + a_k(\tilde{x}_k - \tilde{x}_{k-1}), \quad k \in \mathbb{N}, \tag{PM}$$

with sequences $\{\alpha_k\}_{k\in\mathbb{R}^d}$ and stepsizes $\{\gamma_k\}_{k\in\mathbb{N}}$, that are fixed constant in his work. This method is traditionally attributed to Polyak, although he considers that this method is, already in 1964, *"widely used in linear algebra for speeding up the convergence of iterative methods of solving linear equations"* (Polyak, 1964). This algorithm benefits from a rather direct interpretation: it combines a *gradient step* $\tilde{x}_k - \gamma_k \nabla f(\tilde{x}_k)$ and a *momentum step* $\alpha_k(\tilde{x}_k - \tilde{x}_{k-1})$. The last term can indeed be interpreted as incorporating momentum, as it amplifies the previous displacement $\tilde{x}_k - \tilde{x}_{k-1}$, whose amplitude is controlled by the parameter $\alpha_k$ ($\alpha_k = 0$ means no momentum). The momentum interpretation may be more difficult to grasp in (NM). Crucially, this algorithm share other important formulations, and some reader may actually be more familiar with some of these other formulations rather than with (NM). An important instance of these formulations is the following

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + a_k(\tilde{x}_k - \tilde{x}_{k-1}) + b_k(\tilde{x}_k - \tilde{y}_{k-1}) \\ \tilde{x}_{k+1} = \tilde{y}_k - \gamma \nabla f(\tilde{y}_k) \end{cases} \tag{NM-I}$$

We show in Chapter 3, Proposition 3.B.1 that (NM) can be rewritten as (NM-I) for some parameters. In the case where $b_k = 0$, it is very similar to Polyak's Heavy Ball, with the difference that the momentum step and the gradient step are taken sequentially. This difference induces important consequences on the convergence results, Nesterov's momentum enabling better convergence results compared with Polyak's momentum in some situations, see *e.g.* (Goujaud et al., 2025). We may also mention another formulation of Nesterov's momentum, popular in the machine learning community, here with constant parameters

$$\begin{cases} \tilde{b}_k = p\,\tilde{b}_{k-1} + (1-\tau)\,\nabla f(\tilde{x}_k) \\ \tilde{x}_{k+1} = \tilde{x}_k - s\big(\nabla f(\tilde{x}_k) + p\tilde{b}_k\big) \end{cases} \tag{NM-II}$$

This algorithm is the version implemented in Pytorch with the function *torch.optim.SGD()* using the argument *nesterov = True*. We will see in Chapter 2 that (NM) with constant parameters and (NM-II) are equivalent in the sense that they can generate the same iterates, up to a suitable parameterization. Moreover, this formulation enables another interpretation. Here, $\tilde{x}_{k+1}$ can be seen as a gradient step where the gradient $\nabla f(\tilde{x}_k)$ is replaced by $\nabla f(\tilde{x}_k)$ plus an average of past gradients, through the sequence $\{\tilde{b}_k\}_{k\in\mathbb{N}}$. Here appears explicitly the notion of *memory*: the trajectory keeps using the information of past gradients when updating its iterates. There exists even more possible formulations, see *e.g.* (Defazio, 2019). This profusion of formulations can make the literature rather complicated, giving the impression that there exist lots of different algorithms, while in the end, they are just a specific formulation of Nesterov's momentum algorithm. Several traditional names coexist for Nesterov's momentum, independently of the formulations: examples include *Fast Gradient Method* (FGM) (Nesterov, 2013), Nesterov Accelerated Gradient (NAG) (Bubeck et al., 2015), or Accelerated Gradient Descent (AGD) (d'Aspremont et al., 2021). There may be confusion about the terms "accelerated" and "fast", *i.e.* it may arise from the improvement it enables when minimizing convex, $L$-smooth functions in (the setting in which it was introduced), or refer to the fact that the momentum term induces an accelerated, or faster displacement of the trajectory, independently of the convergence results. This ambiguity may become a nuisance in settings where such accelerated bounds are not possible with this algorithm. For this reason, we believe that naming this algorithm as being "fast" or "accelerated" in general might be slightly weird. This is the justification for our choice of name *Nesterov Momentum*, which might be non-standard.

# Chapter 2

# Nesterov Momentum with Batch Estimator: Effect of Gradient Correlation on Acceleration

We recall that the content of this chapter is based on the following paper.

> J. Hermant, M. Renaud, J-F. Aujol, C. Dossal, and A. Rondepierre.
> **Gradient correlation is a key ingredient to accelerate SGD with momentum**.
> In *The Thirteenth International Conference on Learning Representations*, 2025

## 2.1 Introduction

In this chapter, we address the limitations raised in Section 1.1.3. We recall that our motivation is supervised machine learning tasks, which can often be formulated as the following optimization problem (Hastie et al., 2009):

$$f^* = \min_{x \in \mathbb{R}^d} f(x), \text{ with } f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{FS}$$

where for all $i \in \{1, \dots, N\}$, $f_i : \mathbb{R}^d \to \mathbb{R}$ is associated to one data. In this chapter, we assume a (strongly) convex structure for $f$, which is not realistic for many machine learning models, such as neural networks (Li et al., 2018). However, even in the convex setting, there is still work to do concerning the possibility of accelerating SGD with SNM. We note that our core results about gradient correlation (Propositions 2.4.1-2.4.2) do not assume convexity, and thus could be used in future works beyond the convex setting. Following our discussion of Section 1.1.3, this chapter is devoted to the question:

**Can SNM accelerate over SGD in a convex, interpolation setting ?**

We summarize our contributions as follow.

**Contributions** **(i)** We give a new characterization of the Strong Growth Condition (SGC) constant by using the correlation between gradients, quantified by RACOGA (Propositions 2.4.1-2.4.2), and we exploit this link to study the efficiency of SNM. **(ii)** Using our framework, we study the theoretical impact of batch size on the algorithm performance, depending on the correlation between gradients (Theorem 2.4.1). **(iii)** We complete convergence results of (Vaswani et al., 2019; Gupta et al., 2024) with new almost sure convergence rates (Theorem 2.3.4). **(iv)** We provide numerical experiments that show that RACOGA is a key ingredient to have good performances of SNM compared to SGD.

## 2.2 Background

For a function $f : \mathbb{R}^d \to \mathbb{R}$, continuously differentiable, we recall the following definitions.

**Definition 2.2.1.** *Let $L > 0$. $f : \mathbb{R}^d \to \mathbb{R}$ is **L-smooth** if $\nabla f$ is L-Lipschitz.*

Definition 2.2.1 implies that $\forall x, y \in \mathbb{R}^d,\ f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2$ (Nesterov, 2013) and it ensures that the curvature of the function $f$ is upper-bounded by $L$.

**Definition 2.2.2.** *$f : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-**strongly convex** if there exists $\mu > 0$ such that: $\forall x, y \in \mathbb{R}^d,\ f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$. $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if it verifies this property with $\mu = 0$.*

Definition 2.2.2 implies that the curvature of the function $f$ is lower-bounded by $\mu \geq 0$.

**Assumption 2.2.1** (smoothness)**.** *Each $f_i$ in (FS) is $L_i$-smooth. We note $L_{(K)} := \max_B \frac{1}{K} \sum_{i \in B} L_i$ where $B \subset \{1, \dots, N\}$, Card$(B) = K$ and we note $L_{\max} := L_{(1)} = \max\limits_{1 \leq i \leq N} L_i$.*

Assumption 2.2.1 implies that $f$ is $L$-smooth with $L \leq \frac{1}{N} \sum_{i=1}^N L_i \leq L_{(K)} \leq L_{max}$, and will be used for the convergence results for SGD (Theorems 2.3.1-2.3.2). For some results, we will consider the interpolation setting, namely

**Assumption 2.2.2** (Interpolation)**.** $\exists x^* \in \arg\min f, \forall 1 \leq i \leq N,\ x^* \in \arg\min f_i$.

Such a phenomenom typically arise in over-parameterized models, that are generically able to perfectly fit the learning data (Cooper, 2021; Allen-Zhu et al., 2019; Nakkiran et al., 2021; Zhang et al., 2021). Under such as assumption, algorithms can obtain results close to the deterministic setting, see our discussion in Section 1.1.3.

**SGD** We recall the stochastic gradient descent (Algorithm 2), which can be viewed as a gradient descent where the exact gradient is replaced by a batch estimator

$$\tilde{\nabla}_K(x) := \frac{1}{K} \sum_{i \in B} \nabla f_i(x), \tag{2.1}$$

where $B$ is a batch of indices of size $K$ sampled uniformly in $\{B \subset \{1, \dots, N\} \mid \text{Card}(B) = K\}$. Note that $\tilde{\nabla}_K(x)$ is a random variable depending on $K$, $f$ and $x$.

**Algorithm 2** Stochastic Gradient Descent (SGD)

**input:** $x_0 \in \mathbb{R}^d$, $s > 0$
**for** $n = 0, 1, \dots, n_{it} - 1$ **do**
  $x_{n+1} = x_n - s\tilde{\nabla}_K(x_n)$
**end for**
**output:** $x_{n_{it}}$

**Algorithm 3** Stochastic Nesterov Momentum (SNM)

**input:** $x_0 = z_0 \in \mathbb{R}^d$, $s > 0$, $\beta \in [0, 1]$, $(\alpha_n)_{n \in \mathbf{N}} \in [0, 1]^{\mathbf{N}}$, $(\eta_n)_{n \in \mathbf{N}} \in \mathbf{R}_+^{\mathbf{N}}$
**for** $n = 0, 1, \dots, n_{it} - 1$ **do**
  $y_n = \alpha_n x_n + (1 - \alpha_n) z_n$
  $x_{n+1} = y_n - s\tilde{\nabla}_K(y_n)$
  $z_{n+1} = \beta z_n + (1 - \beta) y_n - \eta_n \tilde{\nabla}_K(y_n)$
**end for**
**output:** $x_{n_{it}}$

**SNM** The Nesterov momentum algorithm (Algorithm 8 in Appendix 2.B.1) allows to achieve faster convergence than gradient descent when considering $L$-smooth functions that are convex or strongly convex, see (Nesterov, 1983, 2013). Intuitively, a momentum mechanism accelerates the gradient descent. As proposed in (Nesterov, 2012), a stochastic version of the Nesterov momentum algorithm can be developed, see Algorithm 3. Note that there exists several ways to write it (see Appendix 2.B.2).

**Strong Growth Condition** We restate the noise assumption that we consider in this chapter. To our knowledge, this assumption was introduced in (Polyak, 1987), and further used by (Cevher and Vu, 2019) as a relaxation of the maximal strong growth condition (Tseng, 1998; Solodov, 1998).

**Definition 2.2.3.** *The function $f$, with a gradient estimator $\tilde{\nabla}_K$ (Equation 2.1), is said to verify the **Strong Growth Condition** if there exists $\rho_K \geq 1$ such that*

$$\forall x \in \mathbb{R}^d : \ \mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] \leq \rho_K \|\nabla f(x)\|^2. \tag{SGC}$$

$\rho_K$ quantifies the amount of noise: the larger $\rho_K$, the higher the noise. In some sense, the strong growth condition allows to replace the norm of the stochastic gradient by the norm of the exact gradient, up to a degrading constant $\rho_K$. We will see in Section 2.3 that it allows to recover similar convergence results as for the deterministic case.

**Example 2.2.1.** *(Vaswani et al., 2019) show that if the function $f$ is $\mu$-Polyak-Łojasiewicz (2.114) and verifies Assumptions 2.2.2 and 2.2.1, then $f$ verifies the strong growth condition with $\rho_K = \frac{L_{(K)}}{\mu}$. In particular, this is also true if replacing (2.114) by strong convexity, see (Necoara et al., 2019).*

**Remark 2.2.1.** *The SGC implies interpolation (Assumption 2.2.2) if each $f_i$ is convex, see Remark 2.H.2 in Appendix 2.H.3. However in the following results (Theorems 2.3.1-2.3.4), as we will only assume that the sum of $f_i$ is convex, SGC will not enforce interpolation. It will imply instead that minimizers of $f$ are critical points of all $f_i$.*

**Remark 2.2.2.** *Considering linear regression, one can choose functions such that the SGC is verified only for arbitrary large values of $\rho_K$ (see Appendix 2.D.1). Worse, if we discard the convexity assumption, then one can construct examples such that $\rho_K$ does not exist (see Appendix 2.D.2). Finding classes of functions such that the SGC is verified for an interesting $\rho_K$ is thus not an obvious task.*

## 2.3 Convergence speed of SNM and comparison with SGD

In this section, we present convergence results for SNM (Algorithm 3) under the Strong Growth Condition (SGC). Before doing so, we introduce convergence results for SGD (Algorithm 2) in order to compare the performance of these two algorithms. All the results introduced in this section are summarized in Table 2.1. For $\varepsilon > 0$, we say that an algorithm $\{x_n\}_n$ reaches an $\varepsilon$-precision at rank $\nu_\varepsilon \in \mathbb{R}_+$ if

$$\forall n \geq \nu_\varepsilon, \ \mathbb{E}\left[f(x_n) - f^*\right] \leq \varepsilon. \tag{2.2}$$

We denote by $\Omega$ the set of realization of the noise. We say that an algorithm $\{x_n\}_n$ converges almost surely with a rate negligible compared to $a_n \in \mathbb{R}_{++}^{\mathbb{N}}$, denoted by $f(x_n) - f^* \overset{a.s.}{=} o(a_n)$, if and only if $\exists A \subset \Omega$, such that $\mathbb{P}(A) = 1$ and $\forall \omega \in A$, $\forall \epsilon > 0$, $\exists n_0 \in \mathbb{N}$, such that $\forall n \geq n_0$,

$$|f(x_n(\omega)) - f^*| \leq \epsilon a_n. \tag{2.3}$$

### 2.3.1 Convergence results for SGD

First, we state convergence results of SGD (Algorithm 2), in expectation and almost surely. The two following theorems are variations of (Gower et al., 2019) and (Gower et al., 2021) (results in expectation) and (Sebbouh et al., 2021) (result almost surely). The difference is that our setting does not assume the convexity of each $f_i$ in (FS), but rather only the convexity of the sum.

**Theorem 2.3.1.** *Under Assumptions 2.2.2 and 2.2.1, SGD (Algorithm 2) guarantees to reach an $\varepsilon$-precision (2.2) at the following iterations:*
- *If $f$ is convex, $s = \frac{1}{2L_{(K)}}$,*

$$n \geq 2\frac{L_{(K)}}{\varepsilon}\left\|x_0 - x^*\right\|^2. \tag{2.4}$$

| Algorithm | Assumption over $f$ | Convergence | |
|---|---|---|---|
| SGD | convex | $\mathbb{E}$, $\varepsilon$ solution | $\mathcal{O}\left(\frac{L_{(K)}}{\varepsilon}\right)$, Thm 2.3.1 |
| | strongly convex | $\mathbb{E}$, $\varepsilon$ solution | $\mathcal{O}\left(\frac{L_{(K)}}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$, Thm 2.3.1 |
| | convex | a.s., c.r | $o\left(\frac{1}{n}\right)$, Thm 2.3.2 |
| | strongly convex | a.s., c.r | $o\left((1-\frac{\mu}{L}+\varepsilon')^n\right)$, Thm 2.3.2 |
| SNM | Convex | $\mathbb{E}$, $\varepsilon$ solution | $\mathcal{O}\left(\rho_K\sqrt{\frac{L}{\varepsilon}}\right)$, Thm 2.3.3 |
| | Strongly Convex | $\mathbb{E}$, $\varepsilon$ solution | $\mathcal{O}\left(\rho_K\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)\right)$, Thm 2.3.3 |
| | Convex | a.s., c.r | $o\left(\frac{1}{n^2}\right)$, Thm 2.3.4 |
| | Strongly | a.s., c.r | $o\left((1-\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}+\varepsilon')^n\right)$, Thm 2.3.4 |

Table 2.1: Summary of all the convergence results presented in Section 2.3. Results stated as $\epsilon$-solution refer to convergence results of the form of Equation (2.2). *c.r.* stands for *convergence rate*. These results are stated as an upper bound of the form $f(x_n) - f^* = \mathcal{O}(\psi_n)$, where $\psi_n$ is a sequence decreasing to 0.

• *If $f$ is $\mu$-strongly convex, $s = \frac{1}{L_{(K)}}$,*

$$n \geq 2\frac{L_{(K)}}{\mu}\log\left(2\frac{f(x_0)-f^*}{\mu\varepsilon}\right). \tag{2.5}$$

For the convex case, the bound is of the order $\mathcal{O}\left(\frac{L_{(K)}}{\varepsilon}\right)$, while for the strongly convex case the key factor is $\frac{L_{(K)}}{\mu}$ that may be very large for ill conditioned problems. These results are very similar to those obtained in a deterministic setting, see Appendix 2.B.1.
Additionally, almost sure convergence gives guarantees that apply to a single run of SGD.

**Theorem 2.3.2.** *Under Assumptions 2.2.2 and 2.2.1, SGD (Algorithm 2) guarantees, in the sense of (2.3), the following asymptotic results, :*
• *If $f$ is convex, $s = \frac{1}{2L_{(K)}}$,*

$$f(\overline{x}_n) - f^* \overset{a.s.}{=} o\left(\frac{1}{n}\right). \tag{2.6}$$

• *If $f$ is $\mu$-strongly convex, $s = \frac{1}{L_{(K)}}$,*

$$f(x_n) - f^* \overset{a.s.}{=} o\left((q+\varepsilon')^n\right), \tag{2.7}$$

*for all $\varepsilon' > 0$, where $q := 1 - \frac{\mu}{L_{(K)}}$, $\overline{x}_0 = x_0$ and $\overline{x}_{n+1} = \frac{2}{n+1}x_n + \frac{n-1}{n+1}\overline{x}_n$.*

Note that in the convex case, there is a need of averaging the trajectory along iterations. Proofs of Theorem 2.3.1 and Theorem 2.3.2 are in Appendix 2.E.1.

### 2.3.2 Convergence in expectation for SNM

We now state the convergence speed of SNM in expectation under the Strong Growth Condition (SGC) and we compare it with the convergence speed of SGD.

**Theorem 2.3.3.** *Assume $f$ is $L$-smooth, and that $\tilde{\nabla}_K$ verifies the SGC for $\rho_K \geq 1$. Then the SNM (Algorithm 3) allows to reach an $\varepsilon$-precision (2.2) at the following iterations:*
• *If $f$ is convex, $s = \frac{1}{L\rho_K}$, $\eta_n = \frac{1}{L\rho_K^2}\frac{n+1}{2}$, $\beta = 1$, $\alpha_n = \frac{\frac{n^2}{n+1}}{2+\frac{n^2}{n+1}}$,*

$$n \geq \rho_K\sqrt{\tfrac{2L}{\varepsilon}}\left\|x_0 - x^*\right\|. \tag{2.8}$$

• *If $f$ is $\mu$-strongly convex,* $s = \frac{1}{L\rho_K},\ \eta_n = \eta = \frac{1}{\rho_K\sqrt{\mu L}},\ \beta = 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}},\ \alpha_n = \alpha = \frac{1}{1+\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}}$,

$$n \geq \rho_K \sqrt{\frac{L}{\mu}} \log\left(2\frac{f(x_0)-f^*}{\varepsilon}\right). \tag{2.9}$$

Theorem 2.3.3 is a variation of (Vaswani et al., 2019; Gupta et al., 2024), see Appendix 2.C.3. Indeed, our proof (Appendix 2.G) leads to the same convergence result as (Vaswani et al., 2019), although resulting in a slightly simpler formulation of the algorithm, as we do not have intermediate sequences of parameters, see Appendix 2.C. Note that we only use the $L$-smoothness of $f$ and the SGC instead of Assumptions 2.2.2-2.2.1 because SGC allows us to make weaker assumptions.

Theorem 2.3.3 indicates that the performance degrades linearly with $\rho_K$. For the special case $\rho_K = 1$, bounds of Theorem 2.3.3 are the same as in the deterministic case (see Appendix 2.B.1).

**Remark 2.3.1.** *According to Theorem 2.3.3 and Theorem 2.3.1, SNM (Algorithm 3) is faster than SGD (Algorithm 2) when $\rho_K$ is small enough, more precisely when*

- $\rho_K < \sqrt{\frac{2L_{(K)}^2}{\varepsilon L}}\,\|x_0 - x^*\|$ *if $f$ convex.*
- $\rho_K < 2\sqrt{\frac{L_{(K)}^2}{\mu L}}$ *if $f$ $\mu$-strongly convex, ignoring the differences between logarithm terms.*

*If $f$ is convex and the required precision $\varepsilon$ small enough, SNM is faster than SGD. It is not necessarily the case if $f$ is $\mu$-strongly convex, as the dependence on $\varepsilon$ disappears. In particular, the bound $\rho_K \leq \frac{L_{(K)}}{\mu}$ offered by strong convexity (see Example 2.2.1) does not guarantee acceleration.*

**Remark 2.3.2.** *In our comparison, we neither considered a convergence result for SGD (Algorithm 2) that assumes the SGC, nor considered a result for SNM (Algorithm 3) that does not assume SGC. In both cases, doing so would lead to misleading comparisons, see summary in Remark 2.F.1.*

### 2.3.3 Almost sure convergence for SNM

We provide new asymptotic almost sure convergence results for SNM (Algorithm 3). Almost sure convergence has already been addressed in (Gupta et al., 2024) without convergence rates.

**Theorem 2.3.4.** *Assume $f$ is $L$-smooth,*

• *If $f$ is convex,* $s = \frac{1}{\rho_K L},\ \eta_n = \frac{1}{4}\frac{n^2}{n+1},\ \beta = 1,\ \alpha_n = \frac{\frac{n^2}{n+1}}{4+\frac{n^2}{n+1}}$,

$$f(x_n) - f^* \overset{a.s.}{=} o\left(\frac{1}{n^2}\right). \tag{2.10}$$

• *If $f$ is $\mu$-strongly convex,* $s = \frac{1}{\rho_K L}$, $\eta_n = \eta = \frac{1}{\rho_K\sqrt{\mu L}}$, $\beta = 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}, \alpha_n = \alpha = \frac{1}{1+\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}}$,

$$f(x_n) - f^* \overset{a.s.}{=} o\left((q+\varepsilon')^n\right) \tag{2.11}$$

*for all $\varepsilon' > 0$, where $q := 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}$.*

See proofs in Appendix 2.G. These bounds are asymptotically better than the finite time bounds, with $o(\frac{1}{n^2})$ compared to $\mathcal{O}(\frac{1}{n^2})$ (Theorem 2.3.3) for instance in the convex setting. A similar asymptotic speedup phenomenon happens in the deterministic setting (Attouch and Peypouquet, 2016).

**Remark 2.3.3.** *Theorem 2.3.4 states that in the convex case, the parameter $\rho_K$ has a negligible impact on the asymptotic convergence, and thus SNM always asymptotically outperforms SGD. For the strongly convex case, we need to ensure that $\rho_K < \sqrt{L_{(K)}^2/\mu L}$ to have SNM faster than SGD.*

The possibility of acceleration of SNM over SGD is highly depending on the SGC constant $\rho_K$. We need to investigate for a fine characterization of $\rho_K$ to ensure acceleration in realistic contexts.

## 2.4 Characterizing convergence with strong growth condition and gradient correlation

Although general bounds on the constant $\rho_K$ are difficult to obtain (see Remark 2.2.2), Example 2.4.1 shows that the characterization of the SGC constant given in Example 2.2.1 can be improved.

**Example 2.4.1** (Motivating example)**.** *Consider the function* $f(x) = \frac{1}{2}\Big(\frac{\mu}{2}\langle e_1, x\rangle^2 + \frac{L}{2}\langle e_2, x\rangle^2\Big)$*, with* $0 < \mu < L$ *and* $e_1, e_2$ *standard basis vectors. This function satisfies Assumption 2.2.2, Assumption 2.2.1 with* $L_{\max} = L$*, and it is* $\frac{\mu}{2}$*-strongly convex. Following Example 2.2.1,* $f$ *satisfies the SGC with* $\rho_1 = 2\frac{L}{\mu}$*, which can be arbitrary large. However, by developing* $\|\nabla f(x)\|^2$*, we get that the SGC is actually verified for* $\rho_1 = 2 < 2\frac{L}{\mu}$*.*

Example 2.4.1 motivates to seek for new, eventually tighter, characterizations of the SGC constant.

### 2.4.1 Average positive correlation condition

In this section, we show how we can exploit the finite sum structure of $f$ to exhibit a condition that, if verified, allows for a new computation of $\rho_K$.

**Proposition 2.4.1.** *Considering batches of size* $K$*, we have*

$$\|\nabla f(x)\|^2 = \frac{K}{N}\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] + \frac{2}{N^2}\frac{N-K}{N-1}\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle. \tag{2.12}$$

Proposition 2.4.1 is proved in Appendix 2.H.1. Without any assumption on $f$, Proposition 2.4.1 splits the norm of $\nabla f$ into two terms. One relies on the gradient estimator, while the other one involves the average correlation of gradients. From Proposition 2.4.1, we deduce the following consequence.

**Corollary 2.4.1.** *Considering batches of size* $K$*,* $f$ *satisfies the SGC with* $\rho_K = \frac{N}{K}$ *if its gradients are, on average, positively correlated, i.e. if we have*

$$\forall x \in \mathbb{R}^d, \quad \sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle \geq 0. \tag{PosCorr}$$

Using condition PosCorr, Corollary 2.4.1 ensures that $f$ verifies a SGC for a constant $\rho_K$ only depending on $N$ and the batch size $K$, and not on geometrical parameters of $f$, *e.g.* $\mu$ or $L$.

**Example 2.4.2.** *Assume* $f_i(x) = \Phi(\langle x, u_i\rangle)$*, for some* $\Phi : \mathbb{R} \to \mathbb{R}$ *that are non necessarily convex, and some orthogonal basis* $\{u_i\}_i$*. We have* $\nabla f_i(x) = \Phi'(\langle x, u_i\rangle)u_i$*. Then, condition PosCorr is verified, and* $f$ *satisfies the SGC with* $\rho_K = \frac{N}{K}$*. Note that the upper-bound* $\rho_K \leq \frac{L_{(K)}}{\mu}$ *given in Example 2.2.1 can be arbitrary large independently of* $N$ *and* $K$ *(see Example 2.4.1), meaning that eventually* $\frac{N}{K} \ll \frac{L_{(K)}}{\mu}$*. Thus, the new upper-bound for* $\rho_K$ *(Corollary 2.4.1) can be much tighter, resulting in improved convergence bounds for SNM (Theorem 2.3.3-2.3.4).*

In some cases, condition PosCorr could be too restrictive (see Appendix 2.I). In the following section, we show how to ensure the SGC with a relaxed version of PosCorr, named RACOGA.

### 2.4.2 RACOGA: relaxing the PosCorr condition

We introduce a new condition named RACOGA which is related, but not the same as two other conditions named gradient diversity (Yin et al., 2018) and gradient confusion (Sankararaman et al., 2020). We discuss these different conditions in Appendix 2.C.

**Definition 2.4.1.** *We say that* $f$ *verifies the Relaxed Averaged COrrelated Gradient Assumption (RACOGA) if there exists* $c \in \mathbb{R}$ *such that the following inequality holds:*

$$\forall x \in \mathbb{R}^d \backslash \overline{\mathcal{X}}, \ \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2} \geq c, \tag{RACOGA}$$

*where* $\overline{\mathcal{X}} = \{x \in \mathbb{R}^d, \forall i \in \{1, \dots, N\}, \|\nabla f_i(x)\| = 0\}$*.*

RACOGA is a generalisation of the condition PosCorr which allows to quantify anti correlation ($c < 0$) or correlation ($c > 0$) between gradients.

**Proposition 2.4.2.** *Assume RACOGA holds with $c > -\frac{1}{2}$. Then, considering batch of size 1, $f$ verifies the SGC with $\rho_1 = \frac{N}{1+2c}$.*

Proposition 2.4.2 is proved in Appendix 2.H.2. Note that RACOGA is always verified with $c = -\frac{1}{2}$, and we have $c \leq \frac{N-1}{2}$ (Appendix 2.H.5).

**Remark 2.4.1.** *Proposition 2.4.2 creates a direct link between RACOGA and SGC. It indicates that the more correlation between gradients there is, the lower is the SGC constant, which results in improved convergence bounds for SNM, see Theorems 2.3.3-2.3.4. Importantly if the gradients are too anti correlated, Proposition 2.4.2 could only be verified with c arbitrary close to $-\frac{1}{2}$, resulting in a bound for $\rho_1$ going to $+\infty$.*

**Remark 2.4.2.** *It is well known that considering high dimensional vectors drawn uniformly on the unit sphere, they will be pairwise quasi orthogonal with high probability (Milman and Schechtman, 1986). (Sankararaman et al., 2020) show theoretically and empirically that considering neural networks with data drawn uniformly on the unit sphere, under some assumptions, linked to over parameterization, each scalar product $\langle \nabla f_i(x), \nabla f_j(x) \rangle$ is not too negative. In this case, RACOGA is thus verified for a c that is at worst close to zero.*

### 2.4.3 The strong growth condition with batch size 1 determines how the performance scales with batch size

In this section, we build over Proposition 2.4.2, which only covers batches of size 1, to take into account bigger batches.

**Lemma 2.4.1.** *Assume that for batches of size 1, $f$ verifies the SGC with constant $\rho_1$. Then, for batches of size $K$, $f$ verifies the strong growth condition with constant $\rho_K$ where*

$$\rho_K \leq \frac{1}{K(N-1)} \left(\rho_1(N-K) + (K-1)N\right). \tag{2.13}$$

Lemma 2.4.1 is proved in Appendix 2.H.3. It shows that if the SGC is verified for batches of size 1, it is verified for any size of batch $K$ and we can compute an estimation of $\rho_K$.

Strikingly, Lemma 2.4.1 allows to study the effect of increasing batch size $K$ on the number of $\nabla f_i$ evaluations we need to reach a $\varepsilon$-solution. We only consider the convex case, as the same reasoning and results hold for strongly convex functions.

**Theorem 2.4.1.** *Assume $f$ is convex and $L$-smooth, and that $\tilde{\nabla}_1$ verifies the SGC with $\rho_1 \geq 1$. Then, SNM (Algorithm 3) with batch size $K$ allows to reach an $\varepsilon$-precision (2.2) at this amount of $\nabla f_i$ evaluations:*

$$\Delta_K . \rho_1 \sqrt{\frac{2L}{\varepsilon}} \left\| x_0 - x^* \right\|, \tag{2.14}$$

*where $\Delta_K := \left(\frac{N-K}{N-1} + \frac{N}{\rho_1}\frac{K-1}{N-1}\right)$, and $\rho_1\sqrt{\frac{2L}{\varepsilon}}\left\|x_0 - x^*\right\|$ is the number of $\nabla f_i$ evaluations needed to reach an $\varepsilon$-precision when using batches of size 1 according to Theorem 2.3.3.*

Compared to the theorems of section 2.3, Theorem 2.4.1 gives a bound on the number of $\nabla f_i$ we have to evaluate, not the number of iterations of the algorithm, see the proof in Appendix 2.H.4.

Theorem 2.4.1 indicates that using batches of size $K$, we need $\Delta_K$ times the number of $\nabla f_i$ that is needed when using batches of size 1 to reach an $\epsilon$-precision. Note also that it assumes the knowledge of $\rho_1$, that can be determined using RACOGA, see Proposition 2.4.2.

**Remark 2.4.3.** *From Theorem 2.4.1, we distinguish 3 regimes, among which the orthogonality of gradients is a critical state.*

1. *$\rho_1 = N$. This is notably true when the gradients are orthogonal. $\Delta_K = 1$ for any value of $K$, and the number of $\nabla f_i$ evaluations is exactly the same independently of batch size.*

2. $\rho_1 < N$. *The gradients are in average positively correlated, i.e. RACOGA is verified with* $c > 0$. $\Delta_K > 1$, *and increasing batch size leads to an increasing amount of* $\nabla f_i$ *evaluations. So, increasing batch size will make parallelization sublinearily efficient, a phenomenon known as performance saturation, see (Ma et al., 2018; Liu and Belkin, 2020).*

3. $\rho_1 > N$. *The gradients are in average negatively correlated, i.e. RACOGA is verified with* $c < 0$. $\Delta_K < 1$ *and larger batches leads to a decreasing amount of* $\nabla f_i$ *evaluations.*

Theorem 2.4.1 and Remark 2.4.3 state that, considering convergence speed, replacing the exact gradient by a stochastic approximation is not necessarily cheaper, in term of number of $\nabla f_i$ we evaluate.

## 2.5 Numerical experiments

Our theory indicates that the correlation between gradients, evaluated through RACOGA is needed to have SNM (Algorithm 3) outperforming SGD (Algorithm 2). We provide numerical experiments to validate this statement by running SNM to optimize classic machine learning models such as linear regression (Section 2.5.1) or classification neural network (Section 2.5.2). We also compare its performance with its deterministic version NM (Algorithm 8), together with GD (Algorithm 7). **In our figures, SNM and NM are named SNAG and NAG**. Our code is available at `https://github.com/J-Hermant/Momentum_Stochastic_GD`.

**RACOGA in practice** We introduced the RACOGA as an inequality that holds over all the space. However, in practice, one only needs to consider the RACOGA values along the optimization path. So this quantity will be computed only along this path.

**Performance metrics** Our interest will be how the algorithms make the training loss function decrease. In order to make a fair comparison between algorithms, our $x$-axis is the number of $\nabla f_i$ evaluations, not $n_{it}$, the number of iterations of the algorithms.

### 2.5.1 Linear regression

For a dataset $\{a_i, b_i\}_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$, we want to solve a linear regression formulated as Problem LR.

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(\langle a_i, x\rangle - b_i)^2. \tag{LR}$$

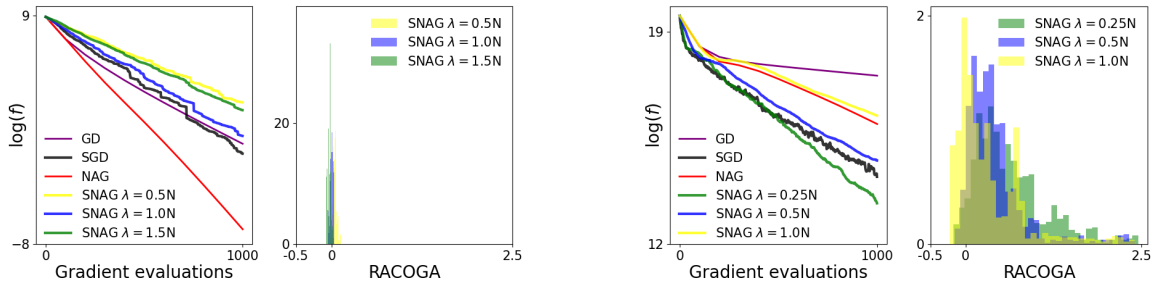
We consider the overparameterized case, *i.e.* $d > N$. Note that the linear regression problem is convex and smooth, so we are in the theoretical setting of this chapter. Moreover, we can directly compute the parameters involved in the algorithms except for a parameter $\lambda$ that replaces the unknown $\rho_K$ constant in the case of SNM, see details in Appendix 2.A.1.

An interesting characteristic of linear regression is that as $\nabla f_i(x) = (\langle a_i, x\rangle - b_i)a_i$, we have

$$\underbrace{\langle \nabla f_i(x), \nabla f_j(x)\rangle}_{\text{gradient correlation}} = (\langle a_i, x\rangle - b_i)(\langle a_j, x\rangle - b_j) \underbrace{\langle a_i, a_j\rangle}_{\text{data correlation}}. \tag{2.15}$$

The pairwise correlation between the gradients depends explicitly on the pairwise correlation inside the data. Therefore, we expect the correlation inside data to impact on the RACOGA values, and thus on the performance of stochastic algorithms such as SNM (Algorithm 3).

To validate this intuition experimentally, we build two different datasets with $N = 100$ and $d = 1000$. The first set of $\{a_i\}_{i=1}^N$ is generated uniformly onto the $d$-dimensional sphere, such that the data are fewly correlated. The second one is generated by a Gaussian mixture law with ten modes, which induces correlation inside data. In both cases, the $\{b_i\}_{i=1}^N$ are generated by a Gaussian law.




(a) Low correlation within data

(b) High correlation within data

Figure 2.1: Illustration of the convergence speed of GD (Algorithm 7), SGD (Algorithm 2, batch size 1), NM (Algorithm 8, NAG in the figure) and SNM (Algorithm 3, batch size 1, SNAG in the figure) on a linear regression problem, together with an histogram distribution of RACOGA values along the iterations of SNM. Stochastic algorithms results are averaged over ten runs. On the left, data are generated by a law that make them fewly correlated, while on the right the data are generated by a gaussian mixture, leading to higher correlation. The $\lambda$ parameter replaces the unknown SGC constant in the algorithm, see Appendix 2.A.1. Note that the data correlation results in better performance of SNM, whereas uncorrelated data lead to smaller RACOGA values, reducing the benefit of using SNM.

**SGD vs SNM** Comparing right parts of Figure 2.1a and Figure 2.1b, we observe that the lack of correlation inside data leads to smaller values of RACOGA. In the case of Figure 2.1a, these lower RACOGA values coincide with SGD being faster than SNM. On the other hand we see on Figure 2.1b that the presence of correlation inside the data makes the optimization path crosses areas with higher RACOGA values, allowing SNM to be faster than SGD. These experimental results support our theoretical findings.

**Deterministic vs stochastic** Strikingly, one can also observe that the lack of correlation results in poor performance of all the stochastic algorithms, especially compared to NM (Figure 2.1a). Conversely, presence of correlation results in the opposite phenomenon (Figure 2.1b). These observations are consistent with our theoretical findings of Section 2.4.3, which indicate that is some cases, stochastic algorithms are not necessarily cheaper to use than their deterministic counterparts in term of $\nabla f_i$ evaluations.

**Role of $\lambda$** In the case of SNM (Algorithm 3), the choice of parameters from Theorem 2.3.3 involves the $\rho_K$ constant, that we do not know. We thus replace $\rho_K$ by a parameter $\lambda$. The higher the RACOGA are, the smaller $\rho_K$ is, and the smaller $\lambda$ can be chosen which results to more aggressive steps of the algorithm, as $s = \frac{1}{L\lambda}$. See details in Appendix 2.A.1.

### 2.5.2 Neural networks

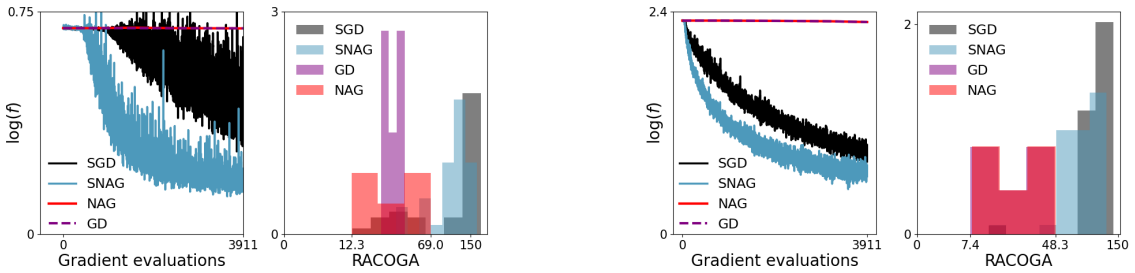
In this second experiment, we aim to test if the crucial role of correlation inside data observed for linear regression (Section 2.5.1) extend to more general models.

For a dataset $\{a_i, b_i\}_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$, we consider a classification problem tackled with a neural network model with the cross-entropy loss. Importantly, this problem is **not convex**, so we are not anymore in the setting of our theoretical results.

We use the SNM version implemented in *Pytorch* (Algorithm 4), that is equivalent to Algorithm 3, see Appendix 2.B.2

As for the linear regression problem, we use two different datasets. The first one, CIFAR10 (Krizhevsky and Hinton, 2009), is composed of images and serves as the correlated dataset. The second one, generated onto the $d$-dimensional sphere with 2 different labels according to which hemisphere belongs each data (see details in Appendix 2.A.2), serves as the uncorrelated dataset. For the CIFAR10 experiment, we use a Convolutional Neural Network (CNN, (LeCun et al., 1998)), and for the sphere experiment we use a Multi Layer Perceptron (MLP, (Rumelhart et al., 1986)).

Strikingly, it appears on Figures 2.2a and 2.2b that the correlation inside data has no longer direct impact on the RACOGA values along the iterations path. In each case the RACOGA values are high, which results




(a) SPHERE - MLP

(b) CIFAR10 - CNN

Figure 2.2: Illustration of the convergence speed of GD (Algorithm 7), SGD (Algorithm 2, batch size 64), NM (Algorithm 4, full batch, NAG in the figure) and SNM (Algorithm 4, batch size 64, SNAG in the figure) averaged over 10 different initializations, together with an histogram distribution of RACOGA values taken along the optimization path, averaged over 10 different initialisations, where the $x$-axis scale is logarithmic. On the left, we use a MLP to classify data sampled from a law such that they are fewly correlated. On the right we use a CNN to classify CIFAR10 images. Note that contrarily to Figure 2.1, the presence of correlation within data no longer influence the RACOGA values, that remains high in both cases, resulting in better performances of SNM.

is SNM outperforming other algorithms. In particular, both deterministic algorithms are significantly less efficient.

These experiments suggest that neural networks offer high RACOGA values, that SNM can take advantage of to converge faster.

## 2.6 Conclusion

In this chapter, we introduced RACOGA to help us to understand in which case the Stochastic Nesterov Momentum algorithm (SNM) allows to outperform the Stochastic Gradient Descent (SGD) for convex or strongly convex functions, as it happens for the deterministic counterparts of these algorithms. We demonstrate theoretically and empirically that **large RACOGA values allows to accelerate SGD with momentum**. RACOGA may be the, up to now, missing ingredient to understand the acceleration possibilities offered by SNM in this setting, outside of the linear regression problems.

## Appendices Organization

These appendices contain additional details on the numerical experiments, proofs of our theoretical results, and some additional insights. In Supplement 2.A, we present additional details about numerical experiments and supplementary experiments. In Supplement 2.B, an optimization background is provided for completeness. In Supplement 2.C, we present and comment related works. In Supplement 2.D, we provide simple examples of functions that do not verify SGC, or verify it for a large constant. In Supplement 2.E, we provide convergence proofs for SGD (Algorithm 2). In Supplement 2.F, we introduce new convergence results for SNM (Algorithm 3) without assuming that the SGC holds. In Supplement 2.G, we present convergence proof for SNM (Algorithm 3). In Supplement 2.H, we provide the proof of the results presented in Section 2.4. Finally in Supplement 2.I, we give deeper explanations concerning the link between RACOGA and algorithms considered in this chapter when applied to the problem of linear regression.

## 2.A Additional experiments and details

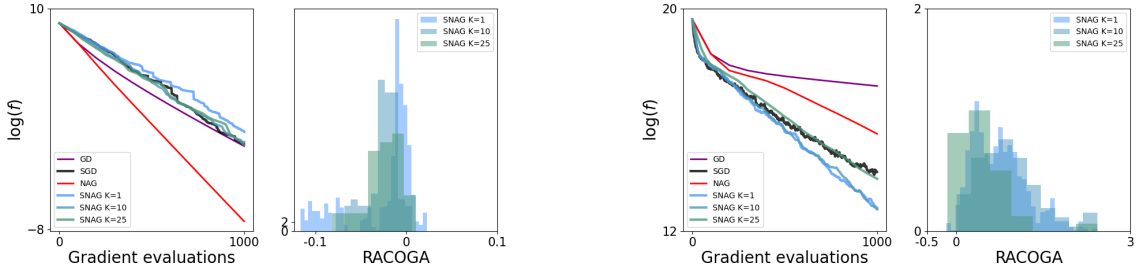
This section presents additional details on the experiments for the sake of reproducibility. We also provide additional experiments for a deeper analysis.

### 2.A.1 Linear regression

**Algorithms and parameters** For the problem LR, we can explicitly compute geometrical constants. Indeed, $f$ is $L$-smooth with $L = \frac{1}{N}\lambda_{\max}\left(\sum_{i=1}^N a_i a_i^T\right)$, and each $f_i$ is $L_i$-smooth with $L_i = \lambda_{\max}\left(a_i a_i^T\right)$. In the overparametrized case *i.e.* $N < d$, $f$ is not $\mu$-strongly convex. However, up to a restriction to the vectorial subspace $\mathbf{V}$ spanned by $\{a_1, \dots, a_N\}$, $f|_{\mathbf{V}}$ is $\mu = \frac{1}{N}\lambda_{\min}\left(\sum_{i=1}^N a_i a_i^T\right)$-strongly convex. Therefore in our experiments, in order to run GD, SGD and NM (respectively Algorithms 7, 2 and 8), we chose the parameters respectively according to Theorem 2.B.1, 2.3.1 and 2.B.2. In the case of SNM (Algorithm 3), in order to apply Theorem 2.3.3, we also need to know the SGC constant $\rho_K$, where $K$ is the selected batch size. However this constant is hard to compute. The knowledge of RACOGA along iterates would be sufficient, but we do not know this path before launching the algorithm. Thus, we run SNM with this choice of parameters

$$s = \frac{1}{L\lambda},\ \eta = \frac{1}{\sqrt{\mu L}\lambda},\ \beta = 1 - \frac{1}{\lambda}\sqrt{\frac{\mu}{L}},\ \alpha = \frac{1}{1 + \frac{1}{\lambda}\sqrt{\frac{\mu}{L}}} \tag{2.16}$$

where $\lambda \geq 1$. In order to achieve better performance, provided that the iterates cross areas with higher RACOGA values, $\lambda$ can be chosen more aggressively, *i.e.* $\lambda$ smaller, as on Figure 2.1b. Decreasing the $\lambda$ parameter leads to a more aggressive, or less safe algorithm, because it will increase $s$ and $\eta$ in Equation (2.16). Taking a glance at Algorithm 3, we see that it results in making larger gradient steps. Recall that larger gradients steps allows the trajectory to move faster, and eventually to converge faster, but steps that are too big will make the algorithm diverge.




(a) Isotropic data

(b) Correlation within data

Figure 2.3: Illustration of the convergence speed of GD (Algorithm 7), NM (Algorithm 8, NAG in the figure), SGD (Algorithm 2) and SNM (Algorithm 3, SNAG in the figure) with varying batch sizes $K$, applied to a linear regression problem, together with an histogram distribution of RACOGA values along the iterations of SNM. The stochastic algorithms results are average on ten runs. On the left, data are generated by a law such that they are fewly correlated, while on the right the data are generated by a gaussian mixture, such that some of the data are highly correlated. Note that the presence of correlation in data results in a decrease of performance for SNM (Algorithm 3) when increasing too much the batch size, whereas uncorrelated data results in an improvement of performance when increasing batch size.

**Algorithm 4** Stochastic Nesterov Momentum - Machine learning version (SNM ML)

**input:** $x_0, b_0 \in \mathbb{R}^d$, $s > 0$, $p \in [0,1]$, $(\eta_n)_{n\in\mathbf{N}} \in \mathbf{R}_+^{\mathbf{N}}$
**for** $n = 0, 1, \dots, n_{it} - 1$ **do**
  $b_n \leftarrow p b_{n-1} + \tilde{\nabla}_K(x_{n-1})$
  $x_n \leftarrow x_{n-1} - s\left(\tilde{\nabla}_K(x_{n-1}) + p b_n\right)$
**end for**
**output:** $x_{n_{it}}$

**Batch size influence** In Section 2.5.1, we observed that contrarily to data generated uniformly onto the sphere, the presence of correlation inside data coincides with high RACOGA values and good performance of

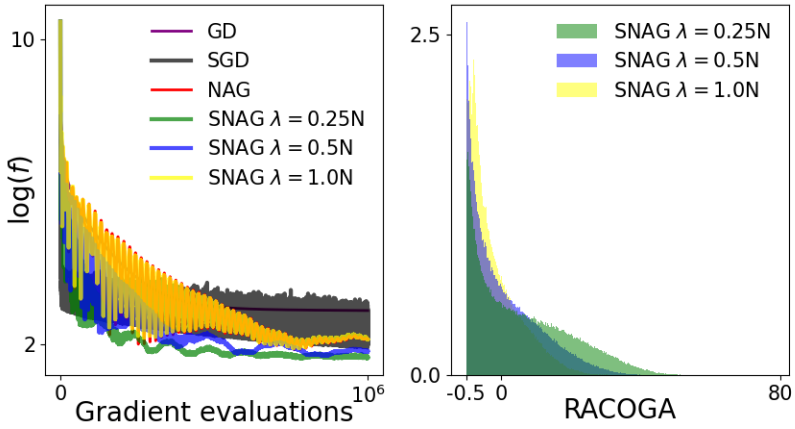



Figure 2.4: Illustration of the convergence speed of GD (Algorithm 7), NM (Algorithm 8, NAG in the figure), SGD (Algorithm 2) and SNM (Algorithm 3, SNAG in the figure) for the linear regression problem with the Boston dataset. The parameters of SNM are choosen accordingly to Equation 2.16. Some RACOGA values are high, while other are also close to the worst anti-correlation case $-\frac{1}{2}$. Note that SNM appears to converge faster than the other algorithms when choosing $\lambda$ small enough.

SNM (Algorithm 3) with batch size 1 compared to other algorithms. Now in the same experimental setting, we study the impact of varying batch size. According to Remark 2.4.3, in the case of correlated data, we should observe a decrease of the performance up to a certain batch size. On Figure 2.3b, we observe this phenomenon. We see that we can multiply the batch size by a factor 10, and keeping the same performance. If performing parallelization, this results in 10 times faster computations. However, when increasing batch size from 10 to 25, we lose performance, inducing paralellization will not results in 2.5 times faster computation. This phenomenon is often referred to as *performance saturation*, see (Ma et al., 2018; Liu and Belkin, 2020). On Figure 2.3a, we observe that conversely, increasing batch size improve the performance. Figure 2.3 is thus consistent with our theoretical findings, see Remark 2.4.3.

**Under-parameterization with a realistic dataset** The Boston dataset (Harrison and Rubinfeld, 1978) concerns house pricing in the Boston area. This dataset can be used for linear regression, where we aim to predict house prices using 13 variables such as tax rates. There are 200 data, so $N = 200$ and $d = 13$, which means we are in an under-parameterized regime. In this case, we can hardly expect SGC to hold, as the noise of the gradient has no reason to be zero at the minimizers of $f$. On Figure 2.4, we tackle the linear regression problem on this dataset, and we observe that despite the under-parameterized setting, SNM (Algorithm 3) appears to outperform GD (Algorithm 7), NM (Algorithm 8) and SGD (Algorithm 2). RACOGA values can be either very large (*e.g.* for aggressive, *i.e.* small $\lambda$), but also close to $-\frac{1}{2}$, which corresponds to high anti-correlation. In this under-parameterized regime, the role of RACOGA is less clear. A possible explanation of the good performance of SNM could be the variance reduction effect of momentum, see (Gao et al., 2024).

### 2.A.2 Neural networks

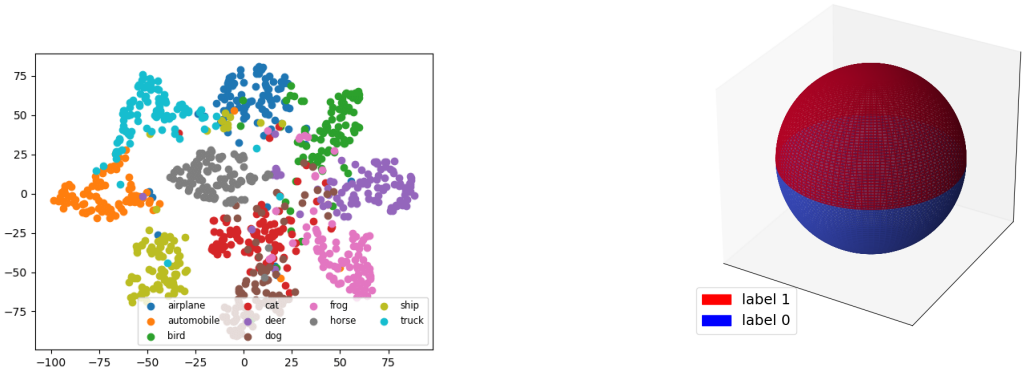
For a dataset $\{a_i, b_i\}_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$, we want to solve a classification problem formulated as Problem C.

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) := \frac{1}{N}\sum_{i=1}^{N} CROSS(x; a_i, b_i). \tag{C}$$

Where $CROSS()$ is a cross entropy loss, as it is implemented in Pytorch with the function *nn.CrossEntropyLoss()*.

**Classification problem** In Section 2.5.2, we considered two classification problems. The first one involves the classic CIFAR10 dataset (Krizhevsky and Hinton, 2009), that contains 60000 color images (dimension

$32 \times 32$) with 10 different labels. See Figure 2.5a to see a data visualisation of the dataset, taken from (Balasubramanian et al., 2022). This dataset serves as our correlated dataset. For the second classification problem, we generated data drawn uniformly onto the $d$-dimensional sphere, where $d$ is the same dimension as the image of CIFAR10, *i.e.* $3 * 32 * 32 = 3072$. We created 2 different labels depending on the positivity of the first coordinate of each data, which remains to associate a different label depending on which hemisphere belong each data, see Figure 2.5b for a 3d visualisation. This dataset is named SPHERE.

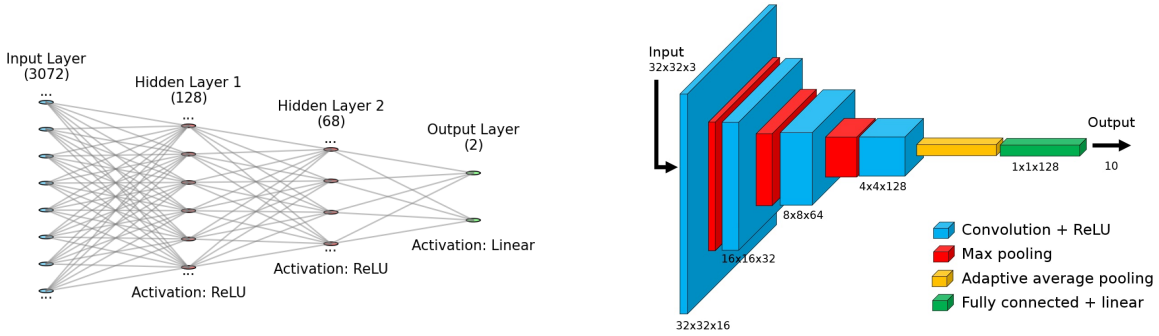



(a) CIFAR10 data vizualisation

(b) Sphere classification problem

Figure 2.5: Illustration of the two classification problems we consider in Section 2.5.2. On the left part, wee see a 2D vizualisation of CIFAR10 data set, proposed by (Balasubramanian et al., 2022). On the right, we illustrate SPHERE dataset on the $3d$-sphere, where each hemisphere correspond to a different label.

**Network architecture** For the classification problem involving the CIFAR10 dataset, we use a Convolutionnal Neural Network (CNN, (LeCun et al., 1998)). For the classification problem involving the spherical data, we use a Multi Layer Perceptron (MLP, (Rumelhart et al., 1986)). CNN are efficient architecture when it comes to image classification, at is exploit local information of the images. However, this architecture makes less sense for our classification problem on the sphere. This model indeed performed poorly in our experiment, and so our choice of a MLP architecture. We detail the architectures on Figure 2.6.




(a) MLP - 401,730 parameters

(b) CNN - 98,730 parameters

Figure 2.6: Scheme of the architecture of the MLP and the CNN used in our experiments.

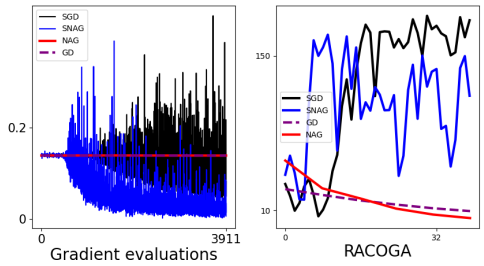
**Algorithms and parameters** We ran the experiments using the *Pytorch* library. We used the *Pytorch* implementation to run the optimization algorithms, through the function *torch.optim.SGD*. This function contains a *nesterov* = *True* argument, which allows to run Algorithm 4. Note that this is indeed a formulation of the Nesterov algorithm, see Appendix 2.B.2. The detailed parameters used to run the experiment,

determined by grid search, are displayed on Figure 2.2 are presented on Table 2.2, together with the final test accuracy, averaged over 10 initialisations.

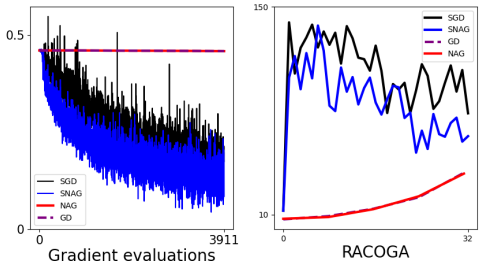
| | MLP (SPHERE) | | | CNN (CIFAR10) | | |
|---|---|---|---|---|---|---|
| | $s$ | $p$ | Accuracy test(%) | $s$ | $p$ | Accuracy test(%) |
| GD | 3 | - | 49.72 | 4 | - | 17.54 |
| SGD | 0.3 | - | 80.6 | 0.3 | - | 65.40 |
| NM | 2 | 0.9 | 50.02 | 2 | 0.7 | 17.07 |
| SNM | 0.1 | 0.9 | **88.73** | 0.05 | 0.9 | **70.88** |

Table 2.2: Parameters (the learning rate $s$ and the momentum $p$) used to run the experiments presented in Section 2.5.2, together with the final accuracy test in percent averaged over 10 different initializations. For precise formulation of the algorithms, see Algorithm 7 (GD), Algorithm 2 (SGD), Algorithm 9 (NM) and Algorithm 4 (SNM).

**Single run and RACOGA along iterations** As a complement, we display on Figure 2.7 a slightly different view of Figure 2.2. On the left part of Figure 2.7a-Figure 2.7b, we display the typical behaviour of one single run of optimization algorithms, namely without averaging for several initialisations. As we can expect, we observe a higher variability, although the general behaviour remains similar. On the right part of Figure 2.7a-Figure 2.7b, we displayed the evolution of RACOGA values along the iterations, instead of the histogram distribution of Figure 2.2, which does not keep any temporal information. For correlated data, we observe that RACOGA values decrease when iterations converge. This phenomenon can be interpreted as the convergence of the iterations to a minimum with low curvature which is related to a low RACOGA value (see Appendix 2.I for more details).




(a) SPHERE-MLP




(b) CIFAR10-CNN

Figure 2.7: Illustration of the convergence speed of on run of GD (Algorithm 7 batch size N ), SGD (Algorithm 2, batch size 64), NM (Algorithm 8, NAG in the figure) and SNM (Algorithm 3, batch size 64, SNAG in the figure), together with a display of the RACOGA values taken along the path, averaged over 5 different initializations. On the left, we use a MLP, and data are generated by an isotropic law and are fewly correlated. On the right we use a CNN, and the dataset used is CIFAR10.

**Algorithm 5** Root Mean Square Propagation (RMSprop)

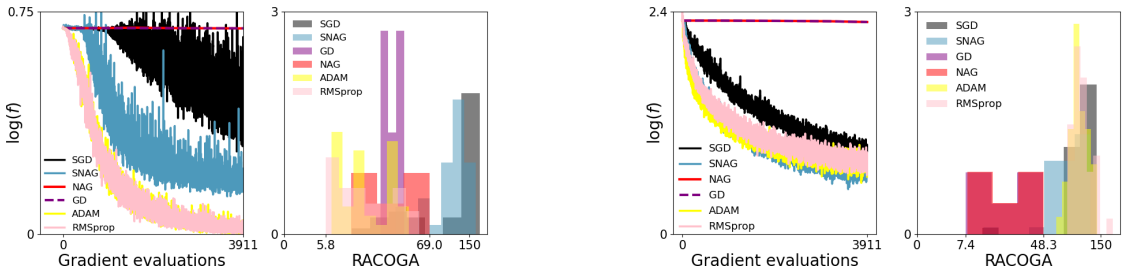
**Input:** $\alpha, \beta, \epsilon > 0$, $x_0 \in \mathbb{R}^d$
**for** $n = 0, \dots, n_{it} - 1$ **do**
  $g_n = \tilde{\nabla}_K(x_n)$
  $v_{n+1} = \beta v_n + (1-\beta) g_n^2$
  $x_{n+1} = x_n - \alpha \frac{g_n}{\sqrt{v_{n+1}+\epsilon}}$
**end for**

**Algorithm 6** Adaptive Moment Estimation (ADAM)

1: **Input:** $\alpha, \beta_1, \beta_2, \epsilon > 0$, initial parameters $x_0 \in \mathbb{R}^d$, $m_0 = v_0 = 0$
2: **for** $n = 0, \dots, n_{it}$ **do**
3: $\quad g_n = \tilde{\nabla}_K(x_n)$
4: $\quad m_{n+1} = \beta_1 m_n + (1-\beta_1) g_n$
5: $\quad v_{n+1} = \beta_2 v_n + (1-\beta_2) g_n^2$
6: $\quad \hat{m}_{n+1} = \frac{m_{n+1}}{1-\beta_1^{n+1}}$
7: $\quad \hat{v}_{n+1} = \frac{v_{n+1}}{1-\beta_2^{n+1}}$
8: $\quad x_{n+1} = x_n - \alpha \frac{\hat{m}_{n+1}}{\sqrt{\hat{v}_{n+1}+\epsilon}}$
9: **end for**

**Comparison with ADAM and RMSprop** RMSprop (Hinton, 2012) and ADAM (Kingma and Ba, 2015) are popular algorithms when it comes to optimize neural networks. We state them in Algorithm 5 and Algorithm 6 respectively, where we in line 5 and line 8, respectively, we denote as the standard division the coordinate-wise division. RMSprop is similar to gradient descent, at the difference that, grossly, it divides component-wise the gradient by an average of the squared norm of the past gradients. There exists a variant of RMSprop that incorporates momentum. Adam combines both techniques of RMSprop and momentum, plus other mechanisms such as bias corrections.

On Figure 2.8, compared to Figure 2.2, we add the training convergence curve and RACOGA values of RMSprop and ADAM. On the CIFAR10 dataset, both algorithms do not converge faster than SNM, and share similar RACOGA values. Interestingly for the SPHERE dataset, both algorithms are significantly faster than others. We observe that all the algorithms are, at the begining of the optimization process, stuck in a tray. SNM steps out of this tray faster than SGD, and RMSprop and ADAM step out of it even faster. One may think that the normalization by the average of squared gradients induces larger stepsize and boost the convergence speed, as in this tray the gradient values are low. The average test accuracy at the end of the training for ADAM (Algorithm 6) is 93.73% for MLP, 69.12% for CNN. The average test accuracy at the end of the training for RMSprop (Algorithm 5) is 94.02% for MLP, 67.31% for CNN.




(a) SPHERE (b) CIFAR10

Figure 2.8: Illustration of the convergence speed of GD (Algorithm 7), SGD (Algorithm 2, batch size 64), NM (Algorithm 4, full batch, NAG in the figure) and SNM (Algorithm 4, batch size 64, SNAG in the figure), ADAM (Algorithm 6) and RMSprop (Algorithm 5) averaged over 10 different initializations, together with an histogram distribution of RACOGA values taken along the optimization path, averaged over 10 different initialisations. On the left, we use a MLP to classify fewly correlated data sampled from an isotropic law. On the right we use a CNN to classify CIFAR10 images. Note that if ADAM and RMSprop are both better than other algorithms for the SPHERE experiment, they are not faster than SNM in the CIFAR10 experiment.

**Computation time** Using the Python library *time*, the computation time needed to choose the best parameters for our two models (CNN for CIFAR10 and MLP for SPHERE) have been saved. Moreover, the computational time needed to generate Figures 2.1, 2.2, 2.3 and 2.8 is added to our computational budget. We saved the computation time needed to train the models and to compute the RACOGA values,

10 times per algorithms (due to the 10 initializations) and for both networks. Note that for the stochastic algorithms, RACOGA was computed only every 100 iterations, because of the heavy computation time it demands. The experiments for the linear regression problem take less than one minute of computation. Experiments with SNM, NM, SGD and GD took approximately 4 hours. Additional experiments with ADAM and RMSprop took approximately 8 hours. Note that ADAM requires to tune 3 hyperparameters with the grid search, making this step significantly longer. The total computation time needed for all these experiments is approximately 12 hours of computation on a GPU NVIDIA A100 Tensor Core.

## 2.B Background : convex optimization and the Nesterov algorithm

In this section, we give the reader some optimization background for completeness of our chapter. First, we expose known results related to the convergence of the deterministic counterparts of SGD and SNM, and then we exhibit different forms of the Nesterov algorithm that can be found in the literature.

### 2.B.1 Performance of GD and NM

We present the speed of convergence of GD (Algorithm 7) and NM (Algorithm 8). The two following theorems are well known results demonstrated by (Nesterov, 1983, 2013).

**Algorithm 7** Gradient Descent (GD)

**input:** $x_0 \in \mathbb{R}^d$, $s > 0$
**for** $n = 0, 1, \dots, n_{it} - 1$ **do**
  $x_{n+1} = x_n - s\nabla f(x_n)$
**end for**
**output:** $x_{n_{it}}$

**Algorithm 8** Nesterov Momentum (NM)

**input:** $x_0, z_0 \in \mathbb{R}^d$, $s > 0$, $\beta \in [0, 1]$, $(\alpha_n)_{n\in\mathbf{N}} \in [0, 1]^{\mathbf{N}}$, $(\eta_n)_{n\in\mathbf{N}} \in \mathbf{R}_+^{\mathbf{N}}$
**for** $n = 0, 1, \dots, n_{it} - 1$ **do**
  $y_n \leftarrow \alpha_n x_n + (1 - \alpha_n) z_n$
  $x_{n+1} \leftarrow y_n - s\nabla f(y_n)$
  $z_{n+1} \leftarrow \beta z_n + (1 - \beta) y_n - \eta_n \nabla f(y_n)$
**end for**
**output:** $x_{n_{it}}$

**Theorem 2.B.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth, $\{x_n\}_{n\in\mathbb{N}}$ be generated by Algorithm 7 with stepsize $s = \frac{1}{L}$. Let $\nu_\epsilon \in \mathbb{N}$ the smallest integer such that $\forall n \geq \nu_\epsilon,\ f(x_n) - f^* \leq \varepsilon$.*

1. *If $f$ is convex, we have*

$$\nu_\epsilon \leq \frac{L}{2} \frac{\|x_0 - x^*\|^2}{\varepsilon}. \tag{2.17}$$

2. *If $f$ is $\mu$-strongly convex, we have*

$$\nu_\epsilon \leq \frac{L}{\mu} \log\left(\frac{f(x_0) - f^*}{\varepsilon}\right). \tag{2.18}$$

**Theorem 2.B.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth, $\{x_n\}_{n\in\mathbb{N}}$ be generated by Algorithm 8 with stepsize $s = \frac{1}{L}$. Let $\nu_\epsilon \in \mathbb{N}$ the smallest integer such that $\forall n \geq \nu_\epsilon,\ f(x_n) - f^* \leq \varepsilon$.*

1. *If $f$ is convex, choosing $\alpha_n = \frac{n-1}{n+1}, \beta = 1, \eta_n = \frac{1}{L}\frac{n+1}{2}$, we have*

$$\nu_\epsilon \leq \sqrt{2L}\frac{\|x_0 - x^*\|}{\sqrt{\varepsilon}}. \tag{2.19}$$

2. *If $f$ is $\mu$-strongly convex, choosing $\alpha_n = \frac{1}{1+\sqrt{\frac{\mu}{L}}}, \beta = 1 - \sqrt{\frac{\mu}{L}}, \eta_n = \frac{1}{\sqrt{\mu L}}$, we have*

$$\nu_\epsilon \leq \sqrt{\frac{L}{\mu}} \log\left(\frac{2(f(x_0) - f^*)}{\varepsilon}\right). \tag{2.20}$$

By comparing the convergence speed for convex functions given by Theorems 2.B.1 and 2.B.2, NM is faster than GD as long as the desire precision $\epsilon$ is small enough. As $L \geq \mu$, considering $\mu$-strongly convex functions NM is always at least as good as GD. In practice, in particular for ill-conditioned problems, we can have $\mu \ll L$. In theses cases, NM is a significant improvement over GD.

**Optimality of NM** Note that NM (Algorithm 8) not only outperforms GD (Algorithm 7). It also offers bounds that are optimal among first-order algorithms when minimizing strongly or non strongly convex functions, in the sense that it is possible to find examples of functions within these classes such that these bounds are achieved up to a constant (Nemirovskij and Yudin, 1983).

### 2.B.2 The different forms of Nesterov algorithm

There exists several ways of writing Nesterov momentum algorithm, that can be linked together (see *e.g.* (Defazio, 2019) for 6 different forms in the strongly convex case). Different research fields are used to their own typical formulation. In this section, we present formulations that can be found in the optimization and machine learning communities. Within the optimization community NM (Algorithm 8) (Allen-Zhu and Orecchia, 2017; Hinder et al., 2020) or NM 2S (Algorithm 10) (Su et al., 2016; Aujol et al., 2023b, 2024c) are often used. The machine learning community is rather used to the formulation of NM ML (Algorithm 9). In fact, this last algorithm is the version implemented in Pytorch with the function *torch.optim.SGD()* with the argument *nesterov = True*, up to the condition that $\tau = 0$.

**Algorithm 9** Nesterov Momentum - Machine Learning version (NM ML)

**input:** $x_0, b_0 \in \mathbb{R}^d$, $s > 0$, $p \in [0,1]$, $\tau \in \mathbb{R}$, $(\eta_n)_{n\in\mathbf{N}} \in \mathbf{R}_+^{\mathbf{N}}$
**for** $n = 0, 1, \ldots, n_{it} - 1$ **do**
  $b_n \leftarrow p b_{n-1} + (1-\tau)\nabla f(x_{n-1})$
  $x_n \leftarrow x_{n-1} - s\left(\nabla f(x_{n-1}) + p b_n\right)$
**end for**
**output:** $x_{n_{it}}$

**Algorithm 10** Nesterov Momentum - Two Sequences version (NM 2S)

**input:** $x_0 \in \mathbb{R}^d$, $s > 0$ , $(a_n)_{n\in\mathbf{N}} \in [0,1]^{\mathbf{N}}$, $(b_n)_{n\in\mathbf{N}} \in [0,1]^{\mathbf{N}}$
**for** $n = 0, 1, \ldots, n_{it} - 1$ **do**
  $y_n \leftarrow x_n + a_n(x_n - x_{n-1}) + b_n(x_n - y_{n-1})$
  $x_{n+1} \leftarrow y_n - s\nabla f(y_n)$
**end for**
**output:** $x_{n_{it}}$

The links between the several forms have been studied previously (Defazio, 2019; Lee et al., 2021). As our results are related to the three points version of NM (Algorithm 8), we state now results that allow to generate the same optimisation scheme than Algorithm 8 with Algorithms 9-10.

**Proposition 2.B.1** (Chapter 3)**.** *Consider $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ generated by NM (Algorithm 8). The same sequences are generated by Algorithm 10 with choice of parameters*

$$a_n = \frac{1-\alpha_n}{1-\alpha_{n-1}}\alpha_{n-1}\beta_{n-1},\ b_n = (1-\alpha_n)\left(\frac{\eta_{n-1}}{s} - \frac{\alpha_{n-1}\beta_{n-1}}{1-\alpha_{n-1}} - 1\right). \tag{2.21}$$

Note that Algorithm 9 have constant parameters in the Pytorch implementation. The link between Algorithm 8 and Algorithm 9 can be deduced from (Defazio, 2019) in the case when $\tau = 1$ for Algorithm 9. We demonstrate a generalisation of this result for $\tau \neq 1$ and we present it in Proposition 2.B.2

**Proposition 2.B.2.** *Consider $(y_n)_{n\in\mathbb{N}}$ generated by Algorithm 8. The sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 9 is the same as $(y_n)_{n\in\mathbb{N}}$ with choices of parameters*

$$p = \alpha\beta,\ \tau = \frac{s(1+\alpha(\beta-1)) - (1-\alpha)\eta}{\alpha\beta s}. \tag{2.22}$$

*Proof.* Our strategy is to write both algorithms in a one point form algorithm to compare the parameters.

*(i) Algorithm 9.* By the line 4 of Algorithm 9, we have the relation

$$b_n = \frac{x_{n-1} - x_n}{ps} - \frac{1}{p}\nabla f(x_{n-1}). \tag{2.23}$$

Now we can replace the $b_n$ and $b_{n-1}$ in line 3 of Algorithm 9, thereby

$$\frac{x_{n-1} - x_n}{ps} - \frac{1}{p}\nabla f(x_{n-1}) = p\left(\frac{x_{n-2} - x_{n-1}}{ps} - \frac{1}{p}\nabla f(x_{n-2})\right) + (1-\tau)\nabla f(x_{n-1}). \tag{2.24}$$

We multiply both sides by $-ps$, and rearrange

$$\begin{aligned} x_n &= x_{n-1} + p(x_{n-1} - x_{n-2}) - s\nabla f(x_{n-1}) + ps\nabla f(x_{n-2}) - ps(1-\tau)\nabla f(x_{n-1}) && (2.25)\\ &= x_{n-1} + p(x_{n-1} - x_{n-2}) - s\nabla f(x_{n-1}) - ps\left(\nabla f(x_{n-1}) - \nabla f(x_{n-2})\right) + \tau ps\nabla f(x_{n-1}) && (2.26)\end{aligned}$$

Interestingly, we recognize with the three first terms on the right hand side Polyak's Heavy Ball (HB) equation (Polyak, 1964)

$$x_n = x_{n-1} + p(x_{n-1} - x_{n-2}) - s\nabla f(x_{n-1}). \tag{HB}$$

The $-ps\left(\nabla f(x_{n-1}) - \nabla f(x_{n-2})\right)$ term is often referred as a gradient correction term (Shi et al., 2022), and is characteristic of the difference between Polyak's Heavy Ball and Nesterov's algorithm. The last $+\tau ps\nabla f(x_{n-1})$ is considered as a damping term, and is often needed to obtain the tightest convergence results (Kim and Fessler, 2016), also to achieve accelerated convergence outside of the deterministic convex realm, see Chapter 3.

*(ii) Algorithm 8* We set that $\forall n \in \mathbb{N}$, $\alpha_n = \alpha \in [0,1]$, and $\eta_n = \eta > 0$. First, combining lines 3 and 4 of Algorithm 8, we have:

$$y_n = \alpha x_n + (1-\alpha)z_n = \alpha(y_{n-1} - s\nabla f(y_{n-1})) + (1-\alpha)z_n \tag{2.27}$$

Hence:

$$z_n = \frac{y_n - \alpha y_{n-1} + \alpha s\nabla f(y_{n-1})}{1-\alpha}. \tag{2.28}$$

Now we inject this expression in line 5 of the Algorithm 8

$$\frac{y_{n+1} - \alpha y_n + \alpha s\nabla f(y_n)}{1-\alpha} = \beta\left(\frac{y_n - \alpha y_{n-1} + \alpha s\nabla f(y_{n-1})}{1-\alpha}\right) + (1-\beta)y_n - \eta\nabla f(y_n). \tag{2.29}$$

Then, multiply both sides by $1-\alpha$ and rearrange to get

$$\begin{aligned} y_{n+1} &= \alpha y_n - \alpha s\nabla f(y_n) + \beta y_n - \beta\alpha y_{n-1} + \alpha\beta s\nabla f(y_{n-1}) && (2.30)\\ &\quad +(1-\alpha)(1-\beta)y_n - (1-\alpha)\eta\nabla f(y_n). && (2.31)\end{aligned}$$

By simplifying and regrouping terms, we get

$$\begin{aligned} y_{n+1} &= y_n + \alpha\beta(y_n - y_{n-1}) - s\nabla f(y_n) - \alpha\beta s\left(\nabla f(y_n) - \nabla f(y_{n-1})\right) \\ &\quad + s\nabla f(y_n) + \alpha\beta s\nabla f(y_n) - \alpha s\nabla f(y_n) - (1-\alpha)\eta\nabla f(y_n) \qquad (2.32) \\ &= y_n + \alpha\beta(y_n - y_{n-1}) - s\nabla f(y_n) - \alpha\beta s\left(\nabla f(y_n) - \nabla f(y_{n-1})\right) \\ &\quad + \left[s(1+\alpha(\beta-1)) - (1-\alpha)\eta\right]\nabla f(y_n). \qquad (2.33) \end{aligned}$$

By comparing, Equation 2.26 and Equation 2.33, we can identify the different parameters

$$p = \alpha\beta \qquad (2.34)$$
$$\tau ps = s(1+\alpha(\beta-1)) - (1-\alpha)\eta \qquad (2.35)$$

So, we have the correspondence

$$p = \alpha\beta \qquad (2.36)$$
$$\tau = \frac{s(1+\alpha(\beta-1)) - (1-\alpha)\eta}{\alpha\beta s} \qquad (2.37)$$

□

**Extension to Stochastic Nesterov algorithms** Proposition 2.B.1 and Proposition 2.B.2 can be extended to the stochastic versions of NM (Algorithm 8), NM ML (Algorithm 9) and NM 2S (Algorithm 10), where the $\nabla f$ terms are replaced by $\tilde{\nabla}_K$. The same correspondence between parameters of these algorithms holds.

## 2.C Related works

### 2.C.1 Gradient correlation conditions

In this chapter, we introduce two assumptions related to the average correlations of the gradients of the functions that form the sum in problem (FS), namely

$$\forall x \in \mathbb{R}^d, \quad \sum_{1\leq i<j\leq N} \langle \nabla f_i(x), \nabla f_j(x)\rangle \geq 0. \qquad \text{(PosCorr)}$$

$$\forall x \in \mathbb{R}^d \backslash \overline{\mathcal{X}}, \quad \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2} \geq c, \qquad \text{(RACOGA)}$$

where $\overline{\mathcal{X}} = \{x \in \mathbb{R}^d, \forall i \in \{1,\dots,N\}, \|\nabla f_i(x)\| = 0\}$. PosCorr is a special case of RACOGA, with the choice $c = 0$. The key role of correlation between gradients has been already observed in previous works, through related but different assumptions.

**Gradient diversity (Yin et al., 2018)** Gradient diversity is defined at a point $x \in \mathbb{R}^d$ as the following ratio

$$\Delta(x) = \frac{\sum_{i=1}^N \|\nabla f_i(x)\|^2}{\left\|\sum_{i=1}^N \nabla f_i(x)\right\|^2}. \qquad \text{(GradDiv)}$$

This quantity is closely related to the SGC condition. Indeed, for batch size 1, we clearly have

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x)\|^2 \leq N \sup_{x\in\mathbb{R}^d} \Delta(x) \|\nabla f(x)\|^2. \qquad (2.38)$$

Thus, assuming $\sup_{x\in\mathbb{R}^d} \Delta(x) < +\infty$, $f$ verifies SGC with $\rho_1 \leq N \sup_{x\in\mathbb{R}^d} \Delta(x)$. The authors show that increasing batch size is less efficient with a ratio depending on the values that $N\Delta(\cdot)$ takes along the optimization path. The smaller these quantities (high correlation), the smaller this ratio. Conversely, the higher it is

(low or anti correlation), the higher the ratio is, inducing a large gain with parallelization of large batches. GradDiv is another measure to quantify gradient correlation. Indeed by developing the squared norm, we have the relation

$$\Delta(x)^{-1} = 1 + 2\frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^{N}\|\nabla f_i(x)\|^2} \tag{2.39}$$

However, RACOGA appears naturally in our study of the SGC, and is a direct measure of the correlation between gradients.

**Gradient confusion (Sankararaman et al., 2020)** $f$ has gradient confusion $\eta \geq 0$ at a point $x \in \mathbb{R}^d$ if

$$\langle \nabla f_i(x), \nabla f_j(x)\rangle \geq -\eta,\ i \neq j \in \{1, \dots, j\}. \tag{GradConf}$$

The authors show that for some classes of functions and considering SGD, satisfying this assumption allows to reduce the size of the neighbourhood of a stationary point towards which the algorithm converges. Also, they study theoretically and empirically how the gradient confusion behaves when considering neural networks. They show that practices that improve the learning, such as increasing width, batch normalization and skip connections, induce a lower gradient confusion at the end of the training, *i.e.* GradConf is verified with a small $\eta$. **In other words, well tuned neural networks avoid anti-correlation between gradients.** However, compared to RACOGA, GradConf asks for a uniform bound over the correlation of each pair of gradients, which can be much more restrictive.

### 2.C.2 Linear regression acceleration results

We mentioned in the introduction of this chapter that there exists positive results concerning the possibility of acceleration considering the linear regression problem (Jain et al., 2018; Liu and Belkin, 2020; Varre and Flammarion, 2022). However, our setting is a bit different. Indeed, we consider the finite sum setting, while they consider the following problem

$$\min_{x\in\mathbb{R}^d} P(x),\ P(x) := \mathbb{E}_{(a,b)\sim\mathcal{D}}\left[\frac{1}{2}(\langle a, x\rangle - b)^2\right]. \tag{2.40}$$

They assume that they have access to a stochastic first-order oracle, that returns the quantity

$$\tilde{\nabla} P(x) = (\langle a, x\rangle - b)a \tag{2.41}$$

where $(a, b) \sim \mathcal{D}$. Noting $H = \mathbb{E}_{(a,b)\sim\mathcal{D}}[aa^T]$, they denote $R^2$ and $\tilde{\kappa}$ the smallest positive constants such that

$$\mathbb{E}_{(a,b)\sim\mathcal{D}}[\|a\|^2 aa^T] \preceq R^2 H,\ \mathbb{E}_{(a,b)\sim\mathcal{D}}[\|a\|^2_{H^{-1}} aa^T] \preceq \tilde{\kappa} H. \tag{2.42}$$

Finally, they note $\mu$ the smallest eigenvalue of $H$ which is assumed to be invertible, and $\kappa := \frac{R^2}{\mu}$. In this setting, there is interpolation if there exists $x^* \in \mathbb{R}^d$ such that almost surely, $b = \langle x^*, a\rangle$. Under Assumption 2.2.2, noting $n_{it}$ the algorithm iterations, the convergence speed of GD (Algorithm 7) is of the order $\mathcal{O}(e^{-\frac{n_{it}}{\kappa}})$, while momentum style algorithms allow to have a $\mathcal{O}(e^{-\frac{n_{it}}{\sqrt{\kappa\tilde{\kappa}}}})$ convergence (Jain et al., 2018; Liu and Belkin, 2020). The acceleration is data dependant. In particular, fixing the distribution to be uniform over the orthonormal basis $\{e_1, \dots, e_d\}$, one have $\kappa = \tilde{\kappa}$ and there is no acceleration. Our results (Example 2.4.2 + Theorem 2.4.1) extend outside linear regression this non-acceleration result, to the case of convex functions with orthogonal gradients. However, the ideas behind those results and ours are different, and $\tilde{\kappa}$ is not a correlation measure. According to (Jain et al., 2018), $\tilde{\kappa}$ measure the number of $a_i$ we need to sample such that the empirical covariance is close to the Hessian matrix, $H$.

### 2.C.3 SGC related convergence results

The linear convergence of SGD under a variation of SGC, for smooth convex and strongly convex functions, has been addressed in (Schmidt and Roux, 2013). Later, convergence of SNM type algorithms under strong growth condition, for functions that are convex or strongly convex, also attracted interest. In this section, we discuss how these works relate to Theorems 2.3.3-2.3.4.

**(Vaswani et al., 2019)** As mentioned in Section 2.3, the bounds of Theorem 2.3.3 were already achieved by (Vaswani et al., 2019). It was achieved with a quite unusual formulation of the SNM algorithm, which is equivalent, to the following one

**Algorithm 11** (Vaswani et al., 2019))

**input:** $x_0 = z_0 \in \mathbb{R}^d$, $s > 0$, $\beta \in [0,1]$, $(\alpha_n)_{n\in\mathbf{N}} \in [0,1]^{\mathbf{N}}$, $(\eta_n)_{n\in\mathbf{N}} \in \mathbf{R}_+^{\mathbf{N}}$, $g_n(.)$ stochastic approximation of $\nabla f(.)$ at iteration $n$.
**for** $n = 0, 1, \ldots, n_{it} - 1$ **do**
  $y_n = \alpha_n x_n + (1-\alpha_n) z_n$
  $x_{n+1} = y_n - s g_n(y_n)$
  $z_{n+1} = \beta z_n + (1-\beta) y_n - \gamma_n s g_n(y_n)$
**end for**
**output:** $x_{n_{it}}$

For $f$ $L$-smooth, and such that SGC is verified with constant $\rho$, the bounds of Theorem 2.3.3 are verified considering SNM (Vaswani et al., 2019) with the following parameters:

• If $f$ $\mu$-strongly convex:

$$\gamma_n = \frac{1}{\sqrt{\mu s \rho}}, \quad \beta_n = 1 - \sqrt{\frac{\mu s}{\rho}}, \quad b_{n+1} = \frac{\sqrt{\mu}}{\left(1 - \sqrt{\frac{\mu s}{\rho}}\right)^{(n+1)/2}},$$
$$a_{n+1} = \frac{1}{\left(1 - \sqrt{\frac{\mu s}{\rho}}\right)^{(n+1)/2}}, \quad \alpha_n = \frac{\gamma_n \beta_n b_{n+1}^2 s}{\gamma_n \beta_n b_{n+1}^n s + a_n^2}, \quad s = \frac{1}{\rho L}.$$

• If $f$ convex:

$$\gamma_n = \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{n-1}^2}}{2}, \quad a_{n+1} = \gamma_n \sqrt{s\rho},$$
$$\alpha_n = \frac{\gamma_n s}{\gamma_n s + a_n^2}, \quad s = \frac{1}{\rho L}.$$

In the strongly convex case, intermediate sequences of parameters $(a_n)$ and $(b_n)$ appear. In the convex case, the sequence $(\gamma_n)$ is defined with a recursive formula. Our different proof of Theorem 2.3.3 does not make these features appear.

**(Gupta et al., 2024)** Another line of research leads to a similar expectation result to ours, using AGNES (Algorithm 12), and so with a different proof. The authors of (Gupta et al., 2024) also get almost sure convergence, nevertheless without convergence rates contrarily to our Theorem 2.3.4.

**Algorithm 12** Accelerated Gradient descent with Noisy EStimators (AGNES), (Gupta et al., 2024)

**input:** $f$ (objective/loss function), $x_0$ (initial point), $\alpha = 10^{-3}$ (learning rate), $\eta = 10^{-2}$ (correction step size), $\rho = 0.99$ (momentum), $N$ (number of iterations)
$v_0 \leftarrow 0$
**for** $n = 0, 1, \ldots, N$ **do**
  $g_n \leftarrow \nabla_x f(x_n)$  (gradient estimator)
  $v_{n+1} \leftarrow \rho(v_n - g_n)$
  $x_{n+1} \leftarrow x_n + \alpha v_{n+1} - \eta g_n$
**end for**
$g_N \leftarrow \nabla_x f(x_N)$
$x_N \leftarrow x_N - \eta g_N$
**output:** $x_N$

## 2.D Some examples with critical strong growth condition constant

In this section, we show that even for some simple examples, the SGC constant can be very large or not existing, justifying that finding interesting characterizations of it is a challenging problem.

### 2.D.1 Large $\rho_1$ with linear regression

We consider the function

$$f(x) = \frac{1}{2}(\frac{1}{2}\underbrace{\langle e_1, x\rangle^2}_{:=f_1(x)} + \frac{1}{2}\underbrace{\langle a, x\rangle^2}_{:=f_2(x)}),$$

where $e_1 = (1, 0)$, $a = (1, \varepsilon)$. We have $\nabla f_1(x) = \langle e_1, x\rangle e_1$ and $\nabla f_2(x) = \langle a, x\rangle a$. Assume $x_0 = (-\frac{\varepsilon}{2}, \lambda)$. We have

$$\nabla f(x_0) = -\frac{\varepsilon}{4}(1, 0) + \frac{\varepsilon}{2}(\lambda - \frac{1}{2})(1, \varepsilon) \tag{2.43}$$

$$= \frac{\varepsilon}{2}(\lambda - 1, (\lambda - \frac{1}{2})\varepsilon). \tag{2.44}$$

Thus we obtain

$$\|\nabla f(x_0)\|^2 = \frac{\varepsilon^2}{4}((\lambda - 1)^2 + (\lambda - \frac{1}{2})^2\varepsilon^2), \tag{2.45}$$

whereas

$$\mathbb{E}[\left\|\tilde{\nabla}_1(x_0)\right\|^2] = \frac{\varepsilon^2}{8} + \frac{\varepsilon^2}{2}(\lambda - \frac{1}{2})^2(1 + \varepsilon^2). \tag{2.46}$$

Simply note that with the choice $\lambda = 1$, we have $\|\nabla f(x_0)\|^2 = \frac{\varepsilon^4}{16}$, while $\mathbb{E}[\left\|\tilde{\nabla}_1(x_0)\right\|^2] = \varepsilon^2\frac{1}{4} + o(\varepsilon^2)$. Thus

$$\frac{\mathbb{E}[\left\|\tilde{\nabla}_1(x_0)\right\|^2]}{\|\nabla f(x_0)\|^2} \approx \frac{1}{\varepsilon^2}. \tag{2.47}$$

So the strong growth condition can be arbitrarily large as $\varepsilon$ vanishes.

### 2.D.2 Non convex functions such that $\rho_1$ does not exist

We consider:

$$f(x) = \frac{1}{2}(f_1(x) + f_2(x))$$

where $x = (x_1, x_2) \in \mathbb{R}^2$, $f_1(x) = \frac{1}{2}(x_1 - \tanh(x_2))^2$, $f_2(x) = \frac{1}{2}(x_1 + \tanh(x_2))^2$. We have:

$$\nabla f_1(x) = (x_1 - \tanh(x_2), -(x_1 - \tanh(x_2))(1 - \tanh^2(x_2))) \tag{2.48}$$

$$\nabla f_2(x) = (x_1 + \tanh(x_2), (x_1 + \tanh(x_2))(1 - \tanh^2(x_2))) \tag{2.49}$$

Consider the line $y = (0, y_0),\ y_0 \in \mathbb{R}$. We have:

$$\nabla f_1(y) = (-\tanh(y_0), \tanh(y_0)(1 - \tanh^2(y_0))) \tag{2.50}$$

$$\nabla f_2(y) = (\tanh(y_0), \tanh(y_0)(1 - \tanh^2(y_0))). \tag{2.51}$$

Then:

$$\nabla f(y) = \frac{1}{2}(\nabla f_1(y) + \nabla f_2(y)) = (0, \tanh(y_0)(1 - \tanh(y_0)^2) \underset{|y_0|\to+\infty}{\to} (0, 0). \tag{2.52}$$

However, we have $\nabla f_1(y) \to (-1, 0)$ and $\nabla f_2(y) \to (1, 0)$ as $y_0 \to +\infty$. Thus, for any $\rho_1 > 0$, the SGC condition is not verified when considering the whole space $\mathbb{R}^2$.

## 2.E Convergence of SGD

In this section, we prove convergence results of SGD (Algorithm 2) stated in Section 2.3, namely Theorem 2.3.1 and Theorem 2.3.2. Also, in Subsection 2.E.2 we present a convergence result for SGD under SGC and we justify that it is not a relevant result in order to compare the convergence speed of SGD and SNM (Algorithm 3).

### 2.E.1 Proofs of Theorem 2.3.1 and Theorem 2.3.2

We can deduce Theorem 2.3.1-Theorem 2.3.2 by respectively adapting the proof from (Garrigos and Gower, 2023) and (Sebbouh et al., 2021). We can not directly apply their original results because their setting is slightly different: in problem (FS) their convexity assumption holds for each $f_i$, while in our case it solely holds over the whole sum. In both expectation and almost sure cases, the core Lemma is to bound $\mathbb{E}\left[\left\|\tilde{\nabla}_K(x)\right\|^2\right]$, which bounds the variance of the estimator. It is Lemma 6.7 in (Garrigos and Gower, 2023), and Lemma 1.3 in (Sebbouh et al., 2021). These two lemmas do not hold in our setting. We will use instead the following result.

**Lemma 2.E.1.** *Under assumptions (2.2.1)-(2.2.2), we have*

$$\mathbb{E}\left[\left\|\tilde{\nabla}_K(x)\right\|^2\right] \leq 2L_{(K)}(f(x) - f^*). \tag{2.53}$$

We note that the above inequality is named the weak growth condition Vaswani et al. (2019), here with constant equal to $2L_{(K)}$.

*Proof.* Let $x \in \mathbb{R}^d$. Consider a fixed batch $B$ with $|B| = K$. We note $f_B(x) = \frac{1}{K}\sum_{i\in B} f_i(x)$, $\nabla f_B(x) := \frac{1}{K}\sum_{i\in B}\nabla f_i(x)$. We first show that $f_B$ is $\frac{\sum_{i\in B} L_i}{K}$-smooth.

$$\forall x, y \in \mathbb{R}^d,\ \|\nabla f_B(x) - \nabla f_B(y)\| \leq \frac{1}{K}\sum_{i\in B}\|\nabla f_i(x) - \nabla f_i(y)\| \tag{2.54}$$

$$\leq \frac{1}{K}\sum_{i\in B} L_i \|x - y\| \tag{2.55}$$

$$= \frac{\sum_{i\in B} L_i}{K}\|x - y\|. \tag{2.56}$$

The first inequality uses triangular inequality, the second inequality uses the assumption that each $f_i$ is $L_i$-smooth, *i.e.* $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x - y\|$. Now, it is well known that if $f_B$ is $\frac{\sum_{i\in B} L_i}{K}$-smooth, fixing $x^* \in \arg\min f$ we have $\|\nabla f_B(x)\|^2 \leq 2\frac{\sum_{i\in B} L_i}{K}(f_B(x) - f_B(x^*))$, see (Nesterov, 2013) page 30. We upper bound this quantity uniformly in the constants $L_i$ over all the batches.

$$\|\nabla f_B(x)\|^2 \leq 2\frac{\sum_{i\in B} L_i}{K}(f_B(x) - f_B(x^*) \leq 2L_{(K)}(f_B(x) - f_B(x^*). \tag{2.57}$$

where $L_{(K)} := \max_{B',|B'|=K}\left(\frac{1}{K}\sum_{i\in B'} L_i\right)$. Now, we get back to the random variable $\tilde{\nabla}_K(\cdot)$, we take the expectation over all the batches of size $K$, and we get

$$\mathbb{E}\left[\left\|\tilde{\nabla}_K(x)\right\|^2\right] \leq 2L_{(K)}\mathbb{E}\left[\sum_{i\in B}\frac{1}{K}(f_i(x) - \min f_i)\right]. \tag{2.58}$$

Note that $\mathbb{E}\left[\frac{1}{K}\sum_{i\in B} f_i(x)\right] = f(x)$. By interpolation (Assumption 2.2.2), there exists $x^* \in \arg\min f$ such that $x^* \in \arg\min f_i$ for all $1 \leq i \leq N$, which implies

$$\mathbb{E}\left[\frac{1}{K}\sum_{i\in B}\min f_i\right] = \mathbb{E}\left[\frac{1}{K}\sum_{i\in B} f_i(x^*)\right] = f(x^*) := f^*. \tag{2.59}$$

□

**Proof of Theorem 2.3.1** Now to prove Theorem 2.3.1, note that the same proof as for Theorem 6.8 and Theorem 6.12 from (Garrigos and Gower, 2023) holds, replacing their Lemma 6.7 by our Lemma 2.E.1, setting in their proof $\sigma_b^* = 0$ (Assumption 2.2.2) and replacing their $2\mathcal{L}_b$ by $L_{(K)}$, allowing in our case to take a stepsize $s \leq \frac{1}{2L_{(K)}}$, instead of their $\frac{1}{4\mathcal{L}_b}$ in the convex case, and $s \leq \frac{1}{L_{(K)}}$ instead of their $\frac{1}{2\mathcal{L}_b}$ in the strongly convex case. Note that for the convex case, the $\varepsilon$-precision is actually reach with the sequence $\overline{x}_n := \frac{1}{n}\sum_{i=0}^{N-1} x_i$, *i.e.* we get a number of iterations such that $\mathbb{E}\left[f(\overline{x}_n) - f^*\right] \leq \varepsilon$.

**Proof of Theorem 2.3.2** For the case of a convex function, the almost sure result from (Sebbouh et al., 2021) follows from a decrease in expectation. It is obtained in our case replacing their Lemma 1.3 by our Lemma 2.E.1. As for the result in expectation, Lemma 2.E.1 allows us to choose $s \leq \frac{1}{L_{(K)}}$. The rest of the proof follows as in theirs, as no supplementary assumption is needed in the interpolated case (Assumption 2.2.2). For the case of strongly convex functions, we apply the same proof as for Proposition 2.G.2, fixing $E_n = \|x_n - x^*\|^2$, where $x^* \in \arg\min f$ with $s = \frac{1}{L_{(K)}}$.

## 2.E.2 Convergence of SGD with strong growth condition

We state and prove a convergence result for SGD (Algorithm 2) when making use of the strong growth condition. A similar result can be found in (Gupta et al., 2024).

**Theorem 2.E.1.** *Assume $f$ is $L$-smooth, and that $\tilde{\nabla}_K$ verifies the SGC for $\rho_K \geq 1$. Then SGD (Algorithm 2) with stepsize $s = \frac{1}{\rho_K L}$ allows to reach an $\varepsilon$-precision (2.2) at the following iterations*

$$n \geq \frac{\rho_K}{2}\frac{L}{\varepsilon}\|x_0 - x^*\|^2 \quad \textit{(f convex)}, \tag{2.60}$$

$$n \geq \rho_K \frac{L}{\mu}\log\left(2\frac{f(x_0) - f^*}{\varepsilon}\right) \quad \textit{(f $\mu$-strongly convex)}. \tag{2.61}$$

*Proof.* These results can be obtained by adapting Theorem 3.4 and Theorem 3.6 from (Garrigos and Gower, 2023). Indeed, the two equations from their Lemma 2.28 can be adapted to our case in the following way, $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[f(x - s\tilde{\nabla}_K(x)) - f(x)\right] \leq -s\left(1 - \frac{\rho_K s L}{2}\right)\|\nabla f(x)\|^2. \tag{2.62}$$

Then, we can reuse their proof just replacing the condition $s \leq \frac{1}{L}$ by $s \leq \frac{1}{\rho_K L}$. The conclusion proceeds fixing $s = \frac{1}{\rho_K L}$ from their Corollaries 3.5 and 3.8. □

It is straightforward to compare these bounds with those of Theorem 2.3.3. This comparison indicates that the bounds of SNM always outperform the bounds of SGD. The only exception that could occur is in the convex case if $\|x_0 - x^*\| \ll 1$, *i.e.* if we start close from the minimum.
Making such conclusions would be misleading. Indeed in some cases, the characterization of Theorem 2.E.1 is suboptimal compared to the one of Theorem 2.3.1. Consider the strongly convex bounds: we then compare $\rho_K \frac{L}{\mu}\log\left(2\frac{f(x_0)-f^*}{\varepsilon}\right)$ (Theorem 2.E.1) with $4\frac{L_{\max}}{\mu}\log\left(2\frac{f(x_0)-f^*}{\mu\varepsilon}\right)$ (Theorem 2.3.1). Now, the question is:

*How do we compare $\rho_K L$ with $L_{\max}$?*

We recall the example of Section 2.D.1, that is

$$f(x) = \frac{1}{2}(\underbrace{\frac{1}{2}\langle e_1, x\rangle^2}_{:=f_1(x)} + \underbrace{\frac{1}{2}\langle a, x\rangle^2}_{:=f_2(x)}). \tag{2.63}$$

where $e_1 = (1, 0)$, $a = (1, \varepsilon)$, $\varepsilon > 0$. We already computed the gradient, that is

$$\nabla f(x) = \frac{1}{2}\langle e_1, x\rangle e_1 + \frac{1}{2}\langle a, x\rangle a. \tag{2.64}$$

One can check that the Hessian matrix is

$$\nabla^2 f(x) = \frac{1}{2}\begin{pmatrix} 2 & \varepsilon \\ \varepsilon & \varepsilon^2 \end{pmatrix}. \tag{2.65}$$

The $L$-smoothness constant is the larger eigenvalue, that is $\frac{1}{4}\left(\left(\varepsilon^4+4\right)^{\frac{1}{2}}+\varepsilon^2+2\right) = 1+o(\varepsilon)$. Moreover, $L_{(1)} = L_{\max} = 1+\varepsilon^2$. Now, we saw in Section 2.D.1 that $\rho_K$ is at least of the order $O\left(\frac{1}{\varepsilon^2}\right)$. Thereby, for small values of $\varepsilon$, we have $\rho_K L \gg L_{(1)}$, inducing the bound of Theorem 2.E.1 is significantly suboptimal. Therefore, Theorem 2.3.1 is more relevant than Theorem 2.E.1 to compare SGD and SNM convergence speeds.

# 2.F Convergence of SNM without strong growth condition

In this Section, we derive a finite time convergence result in expectation for SNM (Algorithm 3) without assuming SGC. In this case, as for Theorems 2.3.1-2.3.2, the bound over the noise is derived from the geometrical properties of the functions.

**Lemma 2.F.1.** *Assume $f$ is such that assumptions (2.2.2)-(2.2.1) hold.*
*If $f$ is $\mu$-strongly convex, we have*

$$\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla}_K(x)\right\|^2\right] \leq 2(L_{(K)}-\mu)(f(x)-f^*). \tag{2.66}$$

*If $f$ is convex, we have*

$$\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla}_K(x)\right\|^2\right] \leq 2L_{(K)}(f(x)-f^*). \tag{2.67}$$

We prove Lemma 2.F.1 in Appendix 2.F.3. We note $L$ the smoothness constant of $f$. Under assumption (2.2.1), we have $L \leq \frac{1}{N}\sum_{i=1}^N L_i \leq L_{(K)}$.

**Theorem 2.F.1.** *Under Assumptions 2.2.2 and 2.2.1, SNM (Algorithm 3) with batch size $K$ guarantees to reach an $\varepsilon$-precision (2.2) at the following iterations:*

- *If $f$ is convex, $s_n = \frac{1}{2L_{(K)}}\frac{1}{n+1}$, $\alpha_n = \frac{n}{n+2}$, $\eta_n = \frac{1}{4}\frac{1}{L_{(K)}}$ , $\beta = 1$,*

$$n \geq \frac{4L_{(K)}}{\varepsilon}\left\|x_0 - x^*\right\|^2. \tag{2.68}$$

- *If $f$ is $\mu$-strongly convex, $s = \frac{1}{16}\frac{\mu}{(L_{(K)}-\mu)^2}$, $\beta = 1-\frac{1}{8}\frac{\mu}{(L_{(K)}-\mu)}$, $\eta = \frac{1}{4}\frac{1}{(L_{(K)}-\mu)}$, $\alpha = \frac{1}{1+\frac{1}{4}\frac{\mu}{(L_{(K)}-\mu)}}$,*

$$n \geq 8\frac{L_{(K)}-\mu}{\mu}\log\left(\frac{2(f(x_0)-f^*)}{\varepsilon}\right). \tag{2.69}$$

In this Section, we denote $\mathcal{F}_n$ the $\sigma$-algebra generated by the $n+1$ first iterates $\{x_i\}_{i=0}^n$ generated by SNM (Algorithm 3), *i.e.* $\mathcal{F}_n = \sigma(x_0,\dots,x_n)$. Also, we note $\mathbb{E}_n[\cdot]$ the conditional expectation with respect to $\mathcal{F}_n$.

**Comparison with the bounds of Theorem 2.3.1 on SGD (Algorithm 2)**

- **$f$ convex.** In this case, the bound for SGD of Theorem 2.3.1 is

$$n \geq 2\frac{L_{(K)}}{\varepsilon}\left\|x_0 - x^*\right\|^2.$$

This bound for SGD is always better than the bound of SNM from Theorem 2.F.1. However this comparison would be misleading, as Theorems 2.3.3 show that in the convex case, if aiming for a small enough precision $\varepsilon$, SNM's bound can always be smaller than the one of SGD.

- $f$ **strongly convex.** In this case, the bound for SGD of Theorem 2.3.1 is

$$n \geq 2\frac{L_{(K)}}{\mu}\log\left(2\frac{f(x_0)-f^*}{\mu\varepsilon}\right).$$

The bound on SNM from Theorem 2.F.1 is better if $\frac{L_{(K)}}{\mu} \leq \frac{4}{3}$. In realistic setting, $\frac{L_{(K)}}{\mu} \gg 1$. In particular if $f$ is quadratic strongly convex, to ensure that the bound of SNM is better, we need to ensure that $\frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{4}{3}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the highest and the lowest eigenvalues of the Hessian matrix of $f$.

Using Assumptions 2.2.2-2.2.1 to study the convergence of SNM almost always leads to a worst bound than SGD independently of the value of $L_{(K)}$, which is misleading. This is because, in this case, we have to take safer parameters for SNM. For example in the convex case, $s_n \to 0$ in order to ensure convergence, compared to the Theorem using SGC (Theorem 2.3.3) where $s_n = s > 0$. This means that this is the SGC characterization of the noise that allows to stabilize SNM and which tells us how to chose parameters such that the convergence is of the order $\mathcal{O}\left(n^{-2}\right)$. Our results of Section 2.4 indicate that what is behind, in the finite sum setting, is the question of measuring the correlation between gradients, that is encapsulated within the SGC.

**Comparison with Theorem 2.3.3** Compared to Theorem 2.3.3 that makes use of SGC, in Theorem 2.F.1 we had to choose more conservative parameters in order to ensure convergence, leading to similar convergence bound as for SGD (Algorithm 2). The comparison of the bounds of Theorems 2.3.3-2.F.1 are similar to the ones of Remark 2.3.1.

**Remark 2.F.1.** *In Remark 2.3.1, we compare SGD and SNM under different assumptions, in the sense that we assume that the SGC holds when studying convergence of SNM, but not for SGD. In Section 2.E.2, we see that using SGC to study SGD leads to convergence bounds that are always worse than for SNM, which is misleading. In this Section, with Theorem 2.F.1, we show that not using SGC when studying SNM leads to an opposite phenomenon, with bounds for SGD that are almost always better than for SNM, which is also misleading. Indeed, in particular, as seen in Remark 2.3.1, in the convex case, using SGC to study SNM allows to have acceleration over SGD in finite-time, as long as we aim for a small enough $\varepsilon$-precision. Also, in our experiments, e.g. on Figures 2.1-2.2, we see that both cases are possible, namely SGD outperforming SNM or SNM outperforming SGD.*

## 2.F.1 Proof of Theorem 2.F.1, convex case

We first recall the SNM algorithm (Algorithm 3), with step-size $s_n$ and $\beta = 1$

$$\left\{\begin{array}{l} y_n = \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} = y_n - s_n \tilde{\nabla}_K(y_n) \\ z_{n+1} = z_n - \eta_n \tilde{\nabla}_K(y_n) \end{array}\right. \tag{SNM}$$

with $\tilde{\nabla}_K(\cdot)$ defined in (2.1).

$$\frac{1}{2}\|z_{n+1}-x^*\|^2 = \frac{1}{2}\|z_n - x^*\|^2 + \frac{\eta_n^2}{2}\|\tilde{\nabla}_K(y_n)\|^2 + \eta_n\langle x^* - z_n, \tilde{\nabla}_K(y_n)\rangle \tag{2.70}$$

$$= \frac{1}{2}\|z_n - x^*\|^2 + \frac{\eta_n^2}{2}\|\tilde{\nabla}_K(y_n)\|^2 + \eta_n\langle x^* - y_n, \tilde{\nabla}_K(y_n)\rangle \tag{2.71}$$

$$+ \eta_n \frac{\alpha_n}{1-\alpha_n}\langle x_n - y_n, \tilde{\nabla}_K(y_n)\rangle \tag{2.72}$$

Taking the expectation, using, under the assumption $s_n \leq \frac{1}{L}$, Lemma 2.F.2, we have

$$\mathbb{E}_n\left[\frac{1}{2}\|z_{n+1}-x^*\|^2\right] \leq \frac{1}{2}\|z_n - x^*\|^2 + \eta_n\langle x^* - y_n, \nabla f(y_n)\rangle \tag{2.73}$$

$$+ \eta_n \frac{\alpha_n}{1-\alpha_n}\langle x_n - y_n, \nabla f(y_n)\rangle + \frac{\eta_n^2}{s_n}\mathbb{E}_n\left[f(y_n) - f(x_{n+1})\right] + \eta_n^2 \mathbb{E}_n\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right]. \tag{2.74}$$

Using convexity of $f$ on both scalar products, we get

$$\mathbb{E}_n\left[\frac{1}{2}\|z_{n+1}-x^*\|^2+\frac{\eta_n^2}{s_n}(f(x_{n+1})-f^*)\right]\leq\frac{1}{2}\|z_n-x^*\|^2+\eta_n\frac{\alpha_n}{1-\alpha_n}(f(x_n)-f(y_n)) \tag{2.75}$$

$$+\left(\frac{\eta_n^2}{s_n}-\eta_n\right)(f(y_n)-f^*)+\eta_n^2\mathbb{E}_n\left[\|\nabla f(y_n)-\tilde{\nabla}_K(y_n)\|^2\right]. \tag{2.76}$$

In order to bound $\mathbb{E}_n\left[\|\nabla f(y_n)-\tilde{\nabla}_K(y_n)\|^2\right]$, we use Lemma 2.F.1

$$\mathbb{E}_n\left[\frac{1}{2}\|z_{n+1}-x^*\|^2+\frac{\eta_n^2}{s_n}(f(x_{n+1})-f^*)\right]\leq\frac{1}{2}\|z_n-x^*\|^2+\eta_n\frac{\alpha_n}{1-\alpha_n}(f(x_n)-f^*)$$
$$+\left(\frac{\eta_n^2}{s_n}-\eta_n\frac{\alpha_n}{1-\alpha_n}-\eta_n+2\eta_n^2L_{(K)}\right)(f(y_n)-f^*). \tag{2.77}$$

We set

$$\frac{\eta_n^2}{s_n}=\frac{C}{L_{(K)}}(n+1)^\alpha,\quad \eta_n\frac{\alpha_n}{1-\alpha_n}=\frac{C}{L_{(K)}}n^\alpha,$$

for some positive constants $C$ and $\alpha$. We want to have $\alpha$ **to be the highest possible**, while having the factor behind $(f(y_n)-f^*)$ non-positive. We have

$$\frac{\eta_n^2}{s_n}=\frac{C}{L_{(K)}}(n+1)^\alpha\Rightarrow\eta_n=\sqrt{\frac{Cs_n}{L_{(K)}}}(n+1)^{\frac{\alpha}{2}}. \tag{2.78}$$

And then

$$\eta_n\frac{\alpha_n}{1-\alpha_n}=\frac{C}{L_{(K)}}n^\alpha\Rightarrow\frac{\alpha_n}{1-\alpha_n}=\frac{1}{\eta_n}\frac{C}{L_{(K)}}n^\alpha \tag{2.79}$$

$$\Rightarrow\frac{\alpha_n}{1-\alpha_n}=\sqrt{\frac{C}{L_{(K)}s_n}}\frac{n^\alpha}{(n+1)^{\frac{\alpha}{2}}}. \tag{2.80}$$

We plug the above equations in the factor behind $(f(y_n)-f^*)$

$$\frac{\eta_n^2}{s_n}-\eta_n\frac{\alpha_n}{1-\alpha_n}-\eta_n+2\eta_n^2L_{(K)}=\frac{C}{L_{(K)}}(n+1)^\alpha-\frac{C}{L_{(K)}}n^\alpha \tag{2.81}$$

$$-\sqrt{\frac{Cs_n}{L_{(K)}}}(n+1)^{\frac{\alpha}{2}}+2s_nC(n+1)^\alpha \tag{2.82}$$

$$=\frac{C}{L_{(K)}}\left((n+1)^\alpha-n^\alpha-\sqrt{\frac{L_{(K)}s_n}{C}}(n+1)^{\frac{\alpha}{2}}\right. \tag{2.83}$$

$$\left.+2s_nL_{(K)}(n+1)^\alpha\right). \tag{2.84}$$

Now, we note $s_n=\frac{C_1}{(n+1)^\beta}$. Then, we have

$$\frac{\eta_n^2}{s_n}-\eta_n\frac{\alpha_n}{1-\alpha_n}-\eta_n+2\eta_n^2L_{(K)}=\frac{C}{L_{(K)}}\left((n+1)^\alpha-n^\alpha-\sqrt{\frac{L_{(K)}C_1}{C}}(n+1)^{\frac{\alpha-\beta}{2}}\right.$$
$$\left.+\,2L_{(K)}C_1(n+1)^{\alpha-\beta}\right). \tag{2.85}$$

We have to ensure that the positive term are not of larger order than the negative ones. We have $(n+1)^\alpha-n^\alpha\approx\alpha n^{\alpha-1}$. So, to have that this term is controlled, we need to have

$$\begin{cases}2(\alpha-1)\leq\alpha-\beta\\ \alpha-\beta\geq2(\alpha-\beta)\end{cases} \tag{2.86}$$

The second constraint implies $\alpha \leq \beta$. As we want $\alpha$ to be maximal, we set $\alpha = \beta$. The first constraint implies $\alpha + \beta \leq 2$, we get $\alpha = \beta = 1$. Plugging these values into Equation (2.85)

$$\frac{\eta_n^2}{s_n} - \eta_n \frac{\alpha_n}{1-\alpha_n} - \eta_n + 2\eta_n^2 L_{(K)} = 0 \Rightarrow \frac{C}{L_{(K)}} \left(1 - \sqrt{\frac{L_{(K)} C_1}{C}} + 2L_{(K)} C_1\right) = 0. \tag{2.87}$$

$$\Rightarrow \sqrt{\frac{L_{(K)} C_1}{C}} = 1 + 2L_{(K)} C_1 \tag{2.88}$$

$$\Rightarrow C = \frac{C_1 L_{(K)}}{(1 + 2L_{(K)} C_1)^2} \tag{2.89}$$

We choose $C_1$ such that $C$ is maximal, and $\frac{C_1 L_{(K)}}{(1+2L_{(K)}C_1)^2}$ is maximized at $C_1 = \frac{1}{2L_{(K)}} \Rightarrow C = \frac{1}{8}$.

So we have

$$s_n = \frac{1}{2L_{(K)}} \frac{1}{n+1}, \quad \eta_n = \frac{1}{4L_{(K)}}, \tag{2.90}$$

and $\frac{\alpha_n}{1-\alpha_n} = \frac{1}{\eta_n} \frac{C}{L_{(K)}} n = \frac{n}{2}$, so $\alpha_n = \frac{n}{n+2}$. Note that as $L_{(K)} \geq L$, this choice of $s_n$ satisfies the constraint $s_n \leq \frac{1}{L}$, needed to apply Lemma 2.F.2. With these choices of parameters, taking expectation on Equation (2.77), we have

$$\mathbb{E}\left[\frac{1}{2}\|z_{n+1} - x^*\|^2 + \frac{(n+1)}{8L_{(K)}}(f(x_{n+1}) - f^*)\right] \leq \mathbb{E}\left[\frac{1}{2}\|z_n - x^*\|^2 + \frac{n}{8L_{(K)}}(f(x_n) - f(x^*))\right]. \tag{2.91}$$

By induction, we get

$$\mathbb{E}\left[\frac{1}{2}\|z_{n+1} - x^*\|^2 + \frac{(n+1)}{8L_{(K)}}(f(x_{n+1}) - f^*)\right] \leq \mathbb{E}\left[\frac{1}{2}\|z_0 - x^*\|^2\right] = \frac{1}{2}\|x_0 - x^*\|^2. \tag{2.92}$$

Now, as $\frac{1}{2}\|z_{n+1} - x^*\|^2 + \frac{(n+1)}{8L_{(K)}}(f(x_{n+1}) - f^*) \geq \frac{(n+1)}{8L_{(K)}}(f(x_{n+1}) - f^*)$, we get

$$\mathbb{E}\left[(f(x_{n+1}) - f^*)\right] \leq \frac{4L_{(K)}}{(n+1)}\|x_0 - x^*\|^2 \tag{2.93}$$

We see that, compared to the case where we assumed SGC, where we had a decrease of the order $\mathcal{O}\left(n^{-2}\right)$, the decrease here is of the order $\mathcal{O}\left(n^{-1}\right)$. In term of $\varepsilon$ solution, it leads to

$$n \geq \frac{4L_{(K)}}{\varepsilon}\|x_0 - x^*\|^2. \tag{2.94}$$

### 2.F.2 Proof of Theorem 2.F.1, strongly convex case

We use the following Lyapunov function.

$$E_n := f(x_n) - f^* + \frac{\mu}{2}\|z_n - x^*\|^2 \tag{2.95}$$

Proceeding to the same computations as in the proof of Theorem 2.3.3 in the strongly convex case, we arrive to Equation (2.148), namely

$$\begin{aligned}\mathbb{E}_n\left[E_{n+1}\right] &= \beta E_n + \mathbb{E}_n\left[f(x_{n+1}) - f^*\right] - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 \\ &+ \frac{\mu}{2}\eta^2 \mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \\ &- \frac{\alpha\beta\eta\mu}{1-\alpha}\langle \mathbb{E}_n\left[\tilde{\nabla}_K(y_n)\right], y_n - x_n\rangle - \mu\eta\langle \mathbb{E}_n\left[\tilde{\nabla}_K(y_n)\right], y_n - x^*\rangle.\end{aligned} \tag{2.96}$$

Recall $\mathbb{E}_n\left[\tilde{\nabla}_K(y_n)\right] = \nabla f(y_n)$. By Lemma 2.F.2, and Lemma 2.F.1 we respectively have

$$\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] \leq \frac{2}{s_n}\mathbb{E}_n\left[(f(y_n) - f(x_{n+1}))\right] + 2\mathbb{E}_n\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right], \tag{2.97}$$

and

$$\mathbb{E}\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right] \leq 2(L_{(K)} - \mu)(f(y_n) - f^*). \tag{2.98}$$

So, by combining the two previous equations, we have

$$\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] \leq \frac{2}{s}\mathbb{E}_n\left[(f(y_n) - f(x_{n+1}))\right] + 4(L_{(K)} - \mu)(f(y_n) - f^*). \tag{2.99}$$

Now, we inject this in (2.96), also using strong convexity.

$$\begin{aligned}\mathbb{E}_n\left[E_{n+1}\right] &\leq \beta E_n + \mathbb{E}_n\left[f(x_{n+1}) - f^*\right] - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 \\ &+ \frac{\mu}{s}\eta^2\mathbb{E}_n\left[f(y_n) - f(x_{n+1})\right] + 2\mu\eta^2(L_{(K)} - \mu)(f(y_n) - f^*) \\ &- \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}(f(y_n) - f(x_n)) - \mu\eta(f(y_n) - f^*) \\ &- \frac{\mu^2\eta}{2}\|y_n - x^*\|^2.\end{aligned} \tag{2.100}$$

Collecting terms and removing the $\|y_n - x_n\|^2$ term, we get

$$\begin{aligned}\mathbb{E}_n\left[E_{n+1}\right] \leq \beta E_n + \left(1 - \frac{\mu}{s}\eta^2\right)\mathbb{E}_n\left[f(x_{n+1}) - f^*\right] + \beta\left(\frac{\alpha\eta\mu}{1-\alpha} - 1\right)\left(f(x_n) - f^*\right) \\ + \frac{\mu}{2}(1 - \beta - \mu\eta)\|y_n - x^*\|^2 + \mu\eta\left(\frac{\eta}{s} - \frac{\alpha\beta}{1-\alpha} - 1 + 2\eta(L_{(K)} - \mu)\right)\left(f(y_n) - f^*\right)\end{aligned} \tag{2.101}$$

We fix $s = \mu\eta^2$ and $\frac{\alpha}{1-\alpha} = \frac{1}{\eta\mu}$, which cancels $\mathbb{E}_n\left[f(x_{n+1}) - f^*\right]$ and the $f(x_n) - f^*$ terms. With these choices, we want

$$\frac{\eta}{s} - \frac{\alpha\beta}{1-\alpha} - 1 + 2\eta(L_{(K)} - \mu) = \frac{1}{\mu\eta} - \frac{\beta}{\mu\eta} - 1 + 2\eta(L_{(K)} - \mu) = 0 \tag{2.102}$$

$$\Rightarrow 1 - \beta = \mu\eta(1 - 2\eta(L_{(K)} - \mu)) \tag{2.103}$$

We want to maximize the right quantity with respect to $\eta$. It is maximized for $\eta = \frac{1}{4(L_{(K)} - \mu)}$, and in this case:

$$\beta = 1 - \mu\eta(1 - 2\eta(L_{(K)} - \mu)) = 1 - \frac{1}{4}\frac{\mu}{(L_{(K)} - \mu)}(1 - \frac{1}{2}) = 1 - \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)} \tag{2.104}$$

Also with that choice of $\eta$ and $\beta$, we have $1 - \beta - \mu\eta = \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)} - \frac{1}{4}\frac{\mu}{(L_{(K)} - \mu)} = -\frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)} < 0$, which ensures that we can control the $\|y_n - x^*\|^2$ term. Thus, with the choice of parameters

$$s = \frac{1}{16}\frac{\mu}{(L_{(K)} - \mu)^2}, \quad \beta = 1 - \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)}, \quad \eta = \frac{1}{4}\frac{1}{(L_{(K)} - \mu)}, \quad \alpha = \frac{1}{1 + \frac{1}{4}\frac{\mu}{(L_{(K)} - \mu)}}, \tag{2.105}$$

we have

$$\mathbb{E}_n\left[E_{n+1}\right] \leq \left(1 - \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)}\right)E_n - \frac{1}{16}\frac{\mu^2}{(L_{(K)} - \mu)}\|y_n - x^*\|^2 \tag{2.106}$$

Ignoring norm term, taking expectation and by induction

$$\mathbb{E}\left[E_n\right] \leq \left(1 - \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)}\right)^n \mathbb{E}\left[E_0\right] \tag{2.107}$$

Finally, we just bound $E_0$, by strong convexity

$$E_0 = f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2 \leq 2(f(x_0) - f^*), \tag{2.108}$$

as we assumed $x_0 = z_0$. Thus, we get

$$\mathbb{E}[f(x_n) - f^*)] \leq 2\left(1 - \frac{1}{8}\frac{\mu}{(L_{(K)} - \mu)}\right)^n \mathbb{E}\left[f(x_0) - f^*)\right] \tag{2.109}$$

We obtain a bound in term of $\varepsilon$ solution following the same reasoning as for the proof of Theorem 2.3.3 in the strongly convex case, and obtain jfathat such a solution is achieved if

$$n \geq 8\frac{L_{(K)} - \mu}{\mu}\log\left(\frac{2(f(x_0) - f^*)}{\varepsilon}\right) \tag{2.110}$$

### 2.F.3 Additional Lemma

In this section, we prove Lemma 2.F.1. Then, we introduce and prove Lemma 2.F.2.

First, we recall the statement of Lemma 2.F.1.

**Lemma.** *Assume $f$ is such that assumptions (2.2.2)-(2.2.1) hold.*

*If $f$ is $\mu$-strongly convex, we have*

$$\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla}_K(x)\right\|^2\right] \leq 2(L_{(K)} - \mu)(f(x) - f^*). \tag{2.111}$$

*If $f$ is convex, we have*

$$\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla}_K(x)\right\|^2\right] \leq 2L_{(K)}(f(x) - f^*). \tag{2.112}$$

*Proof.* We have the elementary relation:

$$\mathbb{E}\left[\|\nabla f(x) - \tilde{\nabla}_K(x)\|^2\right] = \mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] - \|\nabla f(x)\|^2. \tag{2.113}$$

If $f$ is $\mu$-strongly convex, it satisfies the Polyak-Łojasiewicz inequality (Necoara et al., 2019), namely

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*). \tag{2.114}$$

The convex case can be obtained by letting $\mu \to 0$. In this case, we only have $\|\nabla f(x)\|^2 \geq 0$. To conclude, we apply Lemma 2.E.1 and we get the result. □

**Lemma 2.F.2.** *Assuming $f$ is $L$-smooth and $s_n \leq \frac{1}{L}$, iterates of SNM give:*

$$\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] \leq \frac{2}{s_n}\mathbb{E}_n\left[(f(y_n) - f(x_{n+1}))\right] + 2\mathbb{E}_n\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right]. \tag{2.115}$$

*Proof.* Using smoothness, we get

$$f(x_{n+1}) \leq f(y_n) + \langle \nabla f(y_n), x_{n+1} - y_n\rangle + \frac{L}{2}\|x_{n+1} - y_n\|^2 \tag{2.116}$$

$$f(x_{n+1}) \leq f(y_n) - s_n\langle \nabla f(y_n), \tilde{\nabla}_K(y_n)\rangle + \frac{Ls_n^2}{2}\|\tilde{\nabla}_K(y_n)\|^2. \tag{2.117}$$

We have

$$\langle \nabla f(y_n), \tilde{\nabla}_K(y_n)\rangle = \langle \nabla f(y_n) - \tilde{\nabla}_K(y_n), \tilde{\nabla}_K(y_n)\rangle + \|\tilde{\nabla}_K(y_n)\|^2 \tag{2.118}$$

$$= -\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2 + \|\tilde{\nabla}_K(y_n)\|^2 - \langle \nabla f(y_n) - \tilde{\nabla}_K(y_n), \nabla f(y_n)\rangle. \tag{2.119}$$

Note that taking conditional expectation with respect to $\mathcal{F}_n$, the scalar product in equation (2.119) gets cancelled. Inserting (2.119) in (2.117) and taking conditional expectation, we have

$$\mathbb{E}_n\left[f(x_{n+1})\right] \leq \mathbb{E}_n\left[f(y_n)\right] + s_n \mathbb{E}_n\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right] \tag{2.120}$$

$$- s_n\left(1 - \frac{Ls_n}{2}\right)\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] \tag{2.121}$$

$$\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] \leq \frac{2}{s_n}\mathbb{E}_n\left[f(y_n) - f(x_{n+1})\right] + 2\mathbb{E}_n\left[\|\nabla f(y_n) - \tilde{\nabla}_K(y_n)\|^2\right]. \tag{2.122}$$

Where we used in the second inequality that $s_n\left(1 - \frac{Ls_n}{2}\right) \geq \frac{1}{2}$, provided that $s_n \leq \frac{1}{L}$. □

## 2.G Convergence of SNM with strong growth condition

In Sections 2.G.1 and 2.G.2, we provide for completeness a proof of Theorem 2.3.3, that is a similar result to the one from (Vaswani et al., 2019). Our proof is a slightly simpler formulation of the algorithm. In Sections 2.G.3 and 2.G.4, we extend these results proving new almost sure convergences (Theorem 2.3.4), that are asymptotically better that the results in expectation.

In this Section 2.G, we denote $\mathcal{F}_n$ the $\sigma$-algebra generated by the $n+1$ first iterates $\{x_i\}_{i=0}^n$ generated by SNM (Algorithm 3), *i.e.* $\mathcal{F}_n = \sigma(x_0, \ldots, x_n)$. Also, we will note $\mathbb{E}_n[\cdot]$ the conditional expectation with respect to $\mathcal{F}_n$.

First, we present a technical result (Lemma 2.G.1) that will be useful in our proofs.

**Lemma 2.G.1.** *Assume $f$ is $L$-smooth, and that $\tilde{\nabla}_K$ verifies the SGC for $\rho_K \geq 1$. If $y_n$ and $x_{n+1}$ are generated by SNM (Algorithm 3), then with $s = \frac{1}{L\rho_K}$*

$$\|\nabla f(y_n)\|^2 \leq 2L\rho_K \mathbb{E}_n\left[f(y_n) - f(x_{n+1})\right] \tag{2.123}$$

*where $\mathbb{E}_n$ stands for the conditional expectation with respect to $\mathcal{F}_n$.*

*Proof.* By $L$-smoothness, we have

$$f(x_{n+1}) \leq f(y_n) + \langle \nabla f(y_n), x_{n+1} - y_n\rangle + \frac{L}{2}\|x_{n+1} - y_n\|^2 \tag{2.124}$$

$$= f(y_n) - s\langle \nabla f(y_n), \tilde{\nabla}_K(y_n)\rangle + \frac{Ls^2}{2}\|\tilde{\nabla}_K(y_n)\|^2. \tag{2.125}$$

By taking conditional expectation,

$$\mathbb{E}_n\left[f(x_{n+1}) - f(y_n)\right] \leq -s\|\nabla f(y_n)\|^2 + \frac{Ls^2}{2}\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right]. \tag{2.126}$$

Then, by strong growth condition (SGC), we have

$$\mathbb{E}_n\left[f(x_{n+1}) - f(y_n)\right] \leq s\left(\frac{L\rho_K}{2}s - 1\right)\|\nabla f(y_n)\|^2. \tag{2.127}$$

To maximize the decrease, we choose $s = \frac{1}{L\rho_K}$, leading to

$$\mathbb{E}_n\left[f(x_{n+1}) - f(y_n)\right] \leq -\frac{1}{2L\rho_K}\|\nabla f(y_n)\|^2. \tag{2.128}$$

□

### 2.G.1 Convex-Expectation

In this section, we prove statement (2.8) of Theorem 2.3.3 which is the convergence rate of SNM algorithm for a convex function.

We first recall the SNM algorithm (Algorithm 3), with a fixed step-size $s$ and $\beta = 1$

$$\left\{ \begin{array}{l} y_n = \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} = y_n - s\tilde{\nabla}_K(y_n) \\ z_{n+1} = z_n - \eta_n \tilde{\nabla}_K(y_n) \end{array} \right. \tag{SNM}$$

with $\tilde{\nabla}_K(\cdot)$ defined in (2.1).

$$\frac{1}{2}\|z_{n+1} - x^*\|^2 = \frac{1}{2}\|z_n - x^*\|^2 + \frac{\eta_n^2}{2}\|\tilde{\nabla}_K(y_n)\|^2 + \eta_n\langle x^* - z_n, \tilde{\nabla}_K(y_n)\rangle \tag{2.129}$$

$$= \frac{1}{2}\|z_n - x^*\|^2 + \frac{\eta_n^2}{2}\|\tilde{\nabla}_K(y_n)\|^2 + \eta_n\langle x^* - y_n, \tilde{\nabla}_K(y_n)\rangle \tag{2.130}$$

$$+ \eta_n \frac{\alpha_n}{1-\alpha_n}\langle x_n - y_n, \tilde{\nabla}_K(y_n)\rangle. \tag{2.131}$$

After taking conditional expectation with respect to $\mathcal{F}_n$, by the convexity of $f$, SGC and Lemma 2.G.1, we have

$$\frac{1}{2}\mathbb{E}_n[\|z_{n+1} - x^*\|^2] \leq \frac{1}{2}\|z_n - x^*\|^2 - \eta_n(f(y_n) - f(x^*)) + \eta_n \frac{\alpha_n}{1-\alpha_n}(f(x_n) - f(y_n)) \tag{2.132}$$

$$+ L\rho_K^2\eta_n^2\mathbb{E}_n[f(y_n) - f(x_{n+1})]. \tag{2.133}$$

We can reformulate as

$$\mathbb{E}_n[L\rho_K^2\eta_n^2\,(f(x_{n+1}) - f^*) + \frac{1}{2}\|z_{n+1} - x^*\|^2] \leq \eta_n \frac{\alpha_n}{1-\alpha_n}(f(x_n) - f^*) + \frac{1}{2}\|z_n - x^*\|^2 \tag{2.134}$$

$$+ \left(L\rho_K^2\eta_n^2 - \eta_n - \eta_n\frac{\alpha_n}{1-\alpha_n}\right)(f(y_n) - f^*). \tag{2.135}$$

We define parameters as

$$L\rho_K^2\eta_n^2 = \frac{C}{L}(n+1)^2, \quad \eta_n\frac{\alpha_n}{1-\alpha_n} = \frac{C}{L}n^2, \tag{2.136}$$

with $C \geq 0$.

This parameter setting implies $\eta_n = \frac{\sqrt{C}}{L\rho_K}(n+1)$. Thus, we have

$$L\rho_K^2\eta_n^2 - \eta_n - \eta_n\frac{\alpha_n}{1-\alpha_n} = \frac{C}{L}(2n+1) - \frac{\sqrt{C}}{L\rho_K}(n+1) \leq 0 \tag{2.137}$$

$$\Rightarrow \sqrt{C} \leq \frac{1}{\rho_K}\frac{n+1}{2n+1}. \tag{2.138}$$

As we have, for all $n \in \mathbb{N}$, $\frac{1}{2} \leq \frac{n+1}{2n+1} \leq 1$, at best we can set $\sqrt{C} = \frac{1}{2\rho_K}$. With this choice, we have $C = \frac{1}{4\rho_K^2}$, $\eta_n = \frac{1}{L\rho_K^2}\frac{n+1}{2}$ and

$$\frac{\alpha_n}{1-\alpha_n} = \frac{C}{L}n^2\eta_n^{-1} = \frac{1}{2}\frac{n^2}{n+1}. \tag{2.139}$$

This implies that $\alpha_n = \frac{\frac{n^2}{n+1}}{2+\frac{n^2}{n+1}}$. With this choice of parameter, we have

$$\mathbb{E}_n\left[\frac{(n+1)^2}{4L\rho_K^2}(f(x_{n+1}) - f^*) + \frac{1}{2}\|z_{n+1} - x^*\|^2\right] \leq \frac{n^2}{4L\rho_K^2}(f(x_n) - f^*) + \frac{1}{2}\|z_n - x^*\|^2. \tag{2.140}$$

Finally, we get the convergence rate

$$\mathbb{E}[f(x_{n+1}) - f^*] \leq \frac{2L\rho_K^2}{(n+1)^2} \|x_0 - x^*\|^2 . \tag{2.141}$$

To conclude:

$$\frac{2L\rho_K^2}{n^2} \|x_0 - x^*\|^2 \leq \varepsilon \Rightarrow \sqrt{\frac{2L}{\varepsilon}}\rho_K \|x_0 - x^*\| \leq n. \tag{2.142}$$

### 2.G.2 Strongly convex - Expectation

In this section, we prove statement (2.9) of Theorem 2.3.3. Let us remind the algorithm

$$\left\{ \begin{array}{l} y_n = \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} = y_n - s\tilde{\nabla}_K(y_n) \\ z_{n+1} = \beta z_n + (1-\beta) y_n - \eta_n \tilde{\nabla}_K(y_n) \end{array} \right. \tag{SNM}$$

with $\tilde{\nabla}_K(\cdot)$ defined in (2.1). We introduce the following Lyapunov energy:

$$E_n = f(x_n) - f^* + \frac{\mu}{2}\|z_n - x^*\|^2 \tag{2.143}$$

We compute:

$$E_{n+1} - E_n = f(x_{n+1}) - f(x_n) + \frac{\mu}{2}\|z_{n+1} - x^*\|^2 - \frac{\mu}{2}\|z_n - x^*\|^2. \tag{2.144}$$

We start considering the right term

$$\begin{array}{rcl} \Delta_n & = & \|z_{n+1} - x^*\|^2 - \|z_n - x^*\|^2 \\ & = & \|\beta z_n + (1-\beta) y_n - \eta\tilde{\nabla}_K(y_n) - x^*\|^2 - \|z_n - x^*\|^2 \\ & = & (\beta^2 - 1)\|z_n - x^*\|^2 + (1-\beta)^2\|y_n - x^*\|^2 + \eta^2\|\tilde{\nabla}_K(y_n)\|^2 \\ & & +2\beta\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\tilde{\nabla}_K(y_n)\rangle - 2(1-\beta)\eta\langle\tilde{\nabla}_K(y_n), y_n - x^*\rangle \end{array}$$

by construction of Algorithm 3. We now control the first scalar product: using the definition of Algorithm 3, we have $z_n = y_n + \frac{\alpha}{1-\alpha}(y_n - x_n)$, therefore

$$\begin{array}{l} \langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\tilde{\nabla}_K(y_n)\rangle \\ = \langle y_n - x^*, (1-\beta)(y_n - x^*) - \eta\tilde{\nabla}_K(y_n)\rangle \\ +\frac{\alpha}{1-\alpha}\langle y_n - x_n, (1-\beta)(y_n - x^*) - \eta\tilde{\nabla}_K(y_n)\rangle \\ = (1-\beta)\|y_n - x^*\|^2 - \eta\langle y_n - x^*, \tilde{\nabla}_K(y_n)\rangle - \frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \tilde{\nabla}_K(y_n)\rangle \\ +\frac{\alpha}{1-\alpha}(1-\beta)\langle y_n - x_n, y_n - x^*\rangle \end{array}$$

Now, applying the relation $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ to $a = y_n - x^*$ and $b = \frac{\alpha}{1-\alpha}(y_n - x_n)$, we get

$$\frac{\alpha}{1-\alpha}\langle y_n - x_n, y_n - x^*\rangle = \frac{1}{2}\|z_n - x^*\|^2 - \frac{1}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{1}{2}\|y_n - x^*\|^2, \tag{2.145}$$

so that

$$\begin{array}{l} \langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\tilde{\nabla}_K(y_n)\rangle \\ = \frac{1-\beta}{2}\left(\|z_n - x^*\|^2 + \|y_n - x^*\|^2 - \left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2\right) - \eta\langle y_n - x^*, \tilde{\nabla}_K(y_n)\rangle \\ -\frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \tilde{\nabla}_K(y_n)\rangle \end{array}$$

and

$$\begin{aligned}\Delta_n &= -(1-\beta)\|z_n - x^*\|^2 + (1-\beta)\|y_n - x^*\|^2 + \eta^2\|\tilde{\nabla}_K(y_n)\|^2 \\ &\quad -\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - 2\frac{\alpha\beta\eta}{1-\alpha}\langle\tilde{\nabla}_K(y_n), y_n - x_n\rangle \\ &\quad -2\eta\langle\tilde{\nabla}_K(y_n), y_n - x^*\rangle.\end{aligned}$$

Reinjecting $\Delta_n$ in the expression of $E_{n+1} - E_n$ and by definition of $E_n$, we get

$$\begin{aligned}E_{n+1} - E_n &= -(1-\beta)E_n + f(x_{n+1}) - f^* - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 \\ &\quad + \frac{\mu}{2}\eta^2\|\tilde{\nabla}_K(y_n)\|^2 - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \qquad (2.146) \\ &\quad - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\tilde{\nabla}_K(y_n), y_n - x_n\rangle - \mu\eta\langle\tilde{\nabla}_K(y_n), y_n - x^*\rangle. \qquad (2.147)\end{aligned}$$

We take the conditional expectation with respect to $\mathcal{F}_n$:

$$\begin{aligned}\mathbb{E}_n\left[E_{n+1}\right] &= \beta E_n + \mathbb{E}_n\left[f(x_{n+1}) - f^*\right] - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 \qquad (2.148) \\ &\quad + \frac{\mu}{2}\eta^2\mathbb{E}_n\left[\|\tilde{\nabla}_K(y_n)\|^2\right] - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \\ &\quad - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\mathbb{E}_n\left[\tilde{\nabla}_K(y_n)\right], y_n - x_n\rangle - \mu\eta\langle\mathbb{E}_n\left[\tilde{\nabla}_K(y_n)\right], y_n - x^*\rangle. \qquad (2.149)\end{aligned}$$

Using strong convexity of $f$, the strong growth condition (SGC), and then Lemma 2.G.1, we have:

$$\begin{aligned}\mathbb{E}_n\left[E_{n+1}\right] &\leq \beta E_n + \mathbb{E}_n\left[f(x_{n+1}) - f^*\right] - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 \\ &\quad + \mu L\rho_K^2\eta^2\mathbb{E}_n\left[f(y_n) - f(x_{n+1})\right] - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \qquad (2.150) \\ &\quad - \frac{\alpha\beta\eta\mu}{1-\alpha}(f(y_n) - f(x_n)) - \mu\eta(f(y_n) - f^*) - \frac{\mu^2\eta}{2}\|y_n - x^*\|^2. \qquad (2.151) \\ &\leq \beta E_n + \left(1 - \mu L\rho_K^2\eta^2\right)\mathbb{E}_n\left[f(x_{n+1}) - f^*\right] + \beta\left(\frac{\alpha\eta\mu}{1-\alpha} - 1\right)\left(f(x_n) - f^*\right) \\ &\quad + \frac{\mu}{2}(1 - \beta - \mu\eta)\|y_n - x^*\|^2 + \mu\eta\left(L\rho_K^2\eta - \frac{\alpha\beta}{1-\alpha} - 1\right)\left(f(y_n) - f^*\right) \qquad (2.152)\end{aligned}$$

We make the following choices: $\eta = \frac{1}{\sqrt{\mu L}\rho_K}$, $\beta = 1 - \mu\eta = 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}$ and $\frac{\alpha}{1-\alpha} = \frac{1}{\mu\eta} = \rho_K\sqrt{\frac{L}{\mu}} \Rightarrow \alpha = \frac{1}{1+\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}}$. As we have:

$$L\rho_K^2\eta - \frac{\alpha\beta}{1-\alpha} - 1 = \rho_K\sqrt{\frac{L}{\mu}} - \rho_K\sqrt{\frac{L}{\mu}}(1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}) - 1 = 0, \qquad (2.153)$$

these choices cancel all the terms. Thus we have:

$$\mathbb{E}_n\left[E_{n+1}\right] \leq \left(1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)E_n \Rightarrow \mathbb{E}\left[E_{n+1}\right] \leq \left(1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)^{n+1}E_0. \qquad (2.154)$$

Now, note that $E_0 = f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2 \leq 2(f(x_0) - f^*)$, because $f$ is $\mu$-strongly convex. We deduce the following convergence rate:

$$\mathbb{E}[f(x_n) - f^*] \leq 2\left(1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)^n(f(x_0) - f^*). \qquad (2.155)$$

A sufficient condition on the number $n$ of iterations needed to achieve a precision $\varepsilon$ is then naturally given by

$$2\left(1-\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)^n (f(x_0)-f^*) \leq \varepsilon. \tag{2.156}$$

By taking the log, we get

$$n\log\left(1-\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)+\log\left(\frac{2(f(x_0)-f^*)}{\varepsilon}\right)\leq 0. \tag{2.157}$$

Or equivalently

$$n \geq \left|\log\left(1-\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)\right|^{-1}\log\left(\frac{2(f(x_0)-f^*)}{\varepsilon}\right). \tag{2.158}$$

Using the inequality: $|\log(1-x)| \geq x$ for any $x \in (0,1)$, we observe that

$$\left|\log\left(1-\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}\right)\right|^{-1}\log\left(\frac{2(f(x_0)-f^*)}{\varepsilon}\right) \leq \rho_K\sqrt{\frac{L}{\mu}}\log\left(\frac{2(f(x_0)-f^*)}{\varepsilon}\right) \tag{2.159}$$

Hence a sufficient condition on the number of iterations to reach a given precision $\varepsilon$ is

$$\begin{aligned} n &\geq \rho_K\sqrt{\tfrac{L}{\mu}}\log\left(\tfrac{2(f(x_0)-f^*)}{\varepsilon}\right) \\ \Rightarrow n &\geq \left|\log\left(1-\tfrac{1}{\rho_K}\sqrt{\tfrac{\mu}{L}}\right)\right|^{-1}\log\left(\tfrac{2(f(x_0)-f^*)}{\varepsilon}\right) \\ &\Rightarrow \mathbb{E}[f(x_n)-f^*] \leq \varepsilon \end{aligned} \tag{2.160}$$

### 2.G.3 Convex - almost sure

In this section we extend statement (2.8) of Theorem 2.3.3 to get a new almost sure convergence rate.

**Proposition 2.G.1.** *Assume $f$ is $L$-smooth, convex, and that $\tilde{\nabla}_K$ verifies the SGC for $\rho_K \geq 1$. Then SNM (Algorithm 3) with parameter setting $s=\frac{1}{\rho_K L}$, $\beta=1$, $\alpha_n=\frac{\frac{n^2}{n+1}}{4+\frac{n^2}{n+1}}$, $\eta_n=\frac{1}{4}\frac{n^2}{n+1}$ generates a sequence $\{x_n\}_{n\in\mathbb{N}}$ such that*

$$f(x_n)-f^* \overset{a.s.}{=} o\left(\frac{1}{n^2}\right). \tag{2.161}$$

This result is asymptotically better than the result in expectation of Theorem 2.3.3. This asymptotic speedup happens similarly considering the deterministic version of the algorithm (Attouch and Peypouquet, 2016).
Following the scheme of the proof of Theorem 3.1 in (Sebbouh et al., 2021), the Theorem 2.G.1 is the key tool of our proof.

**Theorem 2.G.1** ((Robbins and Siegmund, 1971))**.** *Let $V_n$, $A_n$, $B_n$ and $\alpha_n$ be positive sequences, adapted to some filtration $\mathcal{F}_n$. Assume the following inequality is verified for all $n \in \mathbb{N}$ :*

$$\mathbb{E}[V_{n+1}|\ \mathcal{F}_n] \leq V_n(1+\alpha_n)+A_n-B_n \tag{2.162}$$

*Then, on the set $\{\sum_{i\geq 0}\alpha_i<+\infty, \sum_{i\geq 0}A_i<+\infty\}$, $V_n$ converges almost surely to a random variable $V_\infty$, and we also have $\sum_{i\geq 0}B_i<+\infty$.*

Note that the choice of parameters stated in Proposition 2.G.1 are less agressive (multiplied by a factor $\frac{1}{2}$) compared to the results in expectation (Theorem 2.8).

*Proof.* We start back from Equation (2.134) that we recall

$$\mathbb{E}_n[L\rho_K^2\eta_n^2(f(x_{n+1})-f^*)+\frac{1}{2}\|z_{n+1}-x^*\|^2] \leq \eta_n\frac{\alpha_n}{1-\alpha_n}(f(x_n)-f^*)+\frac{1}{2}\|z_n-x^*\|^2 \tag{2.163}$$

$$+\left(L\rho_K^2\eta_n^2-\eta_n-\eta_n\frac{\alpha_n}{1-\alpha_n}\right)(f(y_n)-f^*). \tag{2.164}$$

We set

$$L\rho_K^2\eta_n^2 = \frac{C}{L}(n+1)^2, \quad \eta_n\frac{\alpha_n}{1-\alpha_n} = \frac{C}{L}n^2, \tag{2.165}$$

with $C \geq 0$.

In this proof, compared to the one of Theorem 2.3.3, we do not want to cancel the last term but to exploit it. More precisely, we are looking forward to the following inequality

$$L\rho_K^2\eta_n^2 - \eta_n - \eta_n\frac{\alpha_n}{1-\alpha_n} \leq -\frac{\eta_n}{2} \tag{2.166}$$

Hence:

$$\frac{C}{L}(2n+1) \leq \frac{\eta_n}{2} = \frac{1}{2}\frac{\sqrt{C}}{L\rho_K}(n+1) \quad \Leftrightarrow \quad \sqrt{C} \leq \frac{1}{2\rho_K}\frac{n+1}{2n+1} \tag{2.167}$$

$$\Leftrightarrow \quad C \leq \frac{1}{16\rho_K^2}\left(\frac{n+1}{n+\frac{1}{2}}\right)^2. \tag{2.168}$$

With the choice $C = \frac{1}{16\rho_K^2}$, the last inequality is verified. Then, we get the following parameters

$$\eta_n = \frac{1}{\rho_K^2 L}\frac{n+1}{4}, \ \frac{\alpha_n}{1-\alpha_n} = \frac{1}{4}\frac{n^2}{n+1}. \tag{2.169}$$

The latter induces $\alpha_n = \frac{\frac{n^2}{n+1}}{4+\frac{n^2}{n+1}}$. Thus, we have

$$\mathbb{E}_n\left[\frac{(n+1)^2}{16L\rho_K^2}(f(x_{n+1}) - f^*) + \frac{1}{2}\|z_{n+1} - x^*\|^2\right] \leq \frac{n^2}{16L\rho_K^2}(f(x_n) - f^*) + \frac{1}{2}\|z_n - x^*\|^2 - \frac{\eta_n}{2}(f(y_n) - f^*).$$

We can now apply Theorem 2.G.1 with

$$\begin{aligned} V_n &:= \frac{n^2}{16L\rho_K^2}(f(x_n) - f^*) + \frac{1}{2}\|z_n - x^*\|^2, \\ A_n &:= 0, \\ B_n &:= \frac{\eta_n}{2}(f(y_n) - f^*), \\ \alpha_n &:= 0. \end{aligned}$$

So we have almost surely

$$\sum_{n\geq 0} B_n = \frac{1}{8\rho_K^2 L}\sum_{n\geq 0}(n+1)(f(y_n) - f^*) < +\infty, \tag{2.170}$$

which implies that

$$\sum_{n\geq 0}(n+1)(f(y_n) - f^*) < +\infty. \tag{2.171}$$

By the definition of Algorithm 3 ($\beta = 1$ in the convex setting), we have

$$x_n - z_n = \alpha_{n-1}(x_{n-1} - z_{n-1}) + (\eta_{n-1} - s)\tilde{\nabla}_K(y_{n-1}). \tag{2.172}$$

Moreover, we have

$$x_n - y_n = (1-\alpha_n)(x_n - z_n). \tag{2.173}$$

By combining Equation (2.172) and Equation (2.173), we get

$$\|x_{n+1} - y_{n+1}\|^2 = (1-\alpha_{n+1})^2\left(\frac{\alpha_n}{1-\alpha_n}\right)^2\|x_n - y_n\|^2 + (1-\alpha_{n+1})^2(\eta_n - s)^2\|\tilde{\nabla}_K(y_n)\|^2 \tag{2.174}$$

$$+ 2(1-\alpha_{n+1})^2\frac{\alpha_n}{1-\alpha_n}(\eta_n - s)\langle x_n - y_n, \tilde{\nabla}_K(y_n)\rangle. \tag{2.175}$$

By taking the expectation with respect to $\mathcal{F}_n$, and using $(\eta_n - s)^2 \leq \eta_n^2$ for $n$ large enough, we have

$$\mathbb{E}_n[\|x_{n+1} - y_{n+1}\|^2] \leq (1-\alpha_{n+1})^2 \left(\frac{\alpha_n}{1-\alpha_n}\right)^2 \|x_n - y_n\|^2 \tag{2.176}$$

$$+ (1-\alpha_{n+1})^2 \eta_n^2 \mathbb{E}_n[\|\tilde{\nabla}_K(y_n)\|^2] + 2(1-\alpha_{n+1})^2 \frac{\alpha_n}{1-\alpha_n}(\eta_n - s)\langle x_n - y_n, \nabla f(y_n)\rangle. \tag{2.177}$$

Using the convexity of $f$, we have $\langle x_n - y_n, \nabla f(y_n)\rangle \leq f(x_n) - f(y_n)$. Thanks to SGC and Lemma 2.G.1, we have

$$\mathbb{E}_n[\|x_{n+1} - y_{n+1}\|^2] \leq (1-\alpha_{n+1})^2 \left(\frac{\alpha_n}{1-\alpha_n}\right)^2 \|x_n - y_n\|^2 \tag{2.178}$$

$$+ 2(1-\alpha_{n+1})^2 \eta_n^2 L \rho_K^2 \mathbb{E}_n[f(y_n) - f(x_{n+1})] \tag{2.179}$$

$$+ 2(1-\alpha_{n+1})^2 \frac{\alpha_n}{1-\alpha_n}(\eta_n - s)(f(x_n) - f(y_n)). \tag{2.180}$$

We divide the previous inequality by $(1-\alpha_{n+1})^2$

$$\mathbb{E}_n\left[\frac{\|x_{n+1} - y_{n+1}\|^2}{(1-\alpha_{n+1})^2}\right] \leq \left(\frac{\alpha_n}{1-\alpha_n}\right)^2 \|x_n - y_n\|^2 + 2\eta_n^2 L \rho_K^2 \mathbb{E}_n[f(y_n) - f(x_{n+1})] \tag{2.181}$$

$$+ 2\frac{\alpha_n}{1-\alpha_n}(\eta_n - s)(f(x_n) - f(y_n)). \tag{2.182}$$

Thus

$$\mathbb{E}_n\left[\frac{\|x_{n+1} - y_{n+1}\|^2}{(1-\alpha_{n+1})^2}\right] \leq \left(\frac{\alpha_n}{1-\alpha_n}\right)^2 \|x_n - y_n\|^2 + 2\eta_n^2 L \rho_K^2 \mathbb{E}_n[(f^* - f(x_{n+1}))] \tag{2.183}$$

$$+ 2\frac{\alpha_n}{1-\alpha_n}(\eta_n - s)(f(x_n) - f^*) \tag{2.184}$$

$$+ 2\left(\eta_n^2 L \rho_K^2 - \frac{\alpha_n}{1-\alpha_n}(\eta_n - s)\right)(f(y_n) - f^*). \tag{2.185}$$

By the parameter setting (Equation (2.169)) and the step-size $s = \frac{1}{\rho_K L}$, we have

$$\eta_n^2 L \rho_K^2 - \frac{\alpha_n}{1-\alpha_n}(\eta_n - s) = \frac{1}{16}\frac{2n+1}{L\rho_K^2} + \frac{1}{L\rho_K}\frac{1}{4}\frac{n^2}{n+1} = \mathcal{O}(n). \tag{2.186}$$

By setting $C_n := 2\left(\eta_n^2 L \rho_K^2 - \frac{\alpha_n}{1-\alpha_n}(\eta_n - s)\right)(f(y_n) - f^*)$, Equation (2.171) and Equation (2.186) gives that almost surely

$$\sum_n C_n < +\infty. \tag{2.187}$$

By defining $\lambda_n := \frac{1}{1-\alpha_n}$ and the parameter setting (Equation (2.169)), Equation 2.183 can be transformed into

$$\mathbb{E}_n[\lambda_{n+1}^2 \|x_{n+1} - y_{n+1}\|^2 + \frac{1}{8}\frac{(n+1)^2}{L\rho_K^2}(f(x_{n+1}) - f^*)] \tag{2.188}$$

$$\leq (1-\lambda_n)^2 \|x_n - y_n\|^2 + \frac{1}{8}\frac{n^2}{L\rho_K^2}(f(x_n) - f^*) - \frac{n^2}{2L\rho_K(n+1)}(f(x_n) - f^*) + C_n \tag{2.189}$$

$$\leq \lambda_n^2 \|x_n - y_n\|^2 + \frac{1}{8}\frac{n^2}{L\rho_K^2}(f(x_n) - f^*) - \frac{n^2}{2L\rho_K(n+1)}(f(x_n) - f^*) + C_n \tag{2.190}$$

$$- (2\lambda_n - 1)\|x_n - y_n\|^2. \tag{2.191}$$

Recalling $\sum_n C_n < +\infty$, we then use Theorem 2.G.1 with

$$\begin{aligned}
\tilde{V}_n &:= \lambda_n^2 \|x_n - y_n\|^2 + \frac{1}{8}\frac{n^2}{L\rho_K^2}(f(x_n) - f^*), \\
\tilde{A}_n &:= C_n, \\
\tilde{B}_n &:= \frac{n^2}{2L\rho_K(n+1)}(f(x_n) - f^*) + (2\lambda_n - 1)\|x_n - y_n\|^2, \\
\tilde{\alpha}_n &:= 0.
\end{aligned}$$

Note that $\lambda_n = \frac{1}{1-\alpha_n} = \frac{4+\frac{n^2}{n+1}}{4} \geq 1$, $2\lambda_n - 1 \geq \lambda_n \geq 0$ and $\tilde{B}_n$ is positive. So, we have that $\lim \tilde{V}_n := \tilde{V}_\infty$ exists almost surely, and $\sum_n \tilde{B}_n < +\infty$ almost surely. However, we have

$$\lambda_n \tilde{B}_n \geq \frac{\lambda_n n^2}{2L\rho_K(n+1)}(f(x_n) - f^*) + \lambda_n^2 \left\|x_n - y_n\right\|^2. \tag{2.192}$$

Moreover, we can compute by the parameter setting (Equation 2.169)

$$\frac{\lambda_n n^2}{2L\rho_K(n+1)} = \frac{4+\frac{n^2}{n+1}}{4}\frac{1}{L\rho_K}\frac{1}{2}\frac{n^2}{n+1} = \frac{n^2}{8L\rho_K^2}\frac{\rho_K(4+\frac{n^2}{n+1})}{n+1}. \tag{2.193}$$

As $\frac{\rho_K(4+\frac{n^2}{n+1})}{n+1} = \rho_K \frac{n^2+4n+4}{n^2+2n+1} > 1$, we have $\frac{\lambda_n n^2}{2L\rho_K(n+1)} > \frac{n^2}{8L\rho_K^2}$, and thus

$$\lambda_n \tilde{B}_n \geq \tilde{V}_n. \tag{2.194}$$

We can deduce from the previous inequality

$$\sum_{n\geq 0} \tilde{B}_n = \sum_{n\geq 0} \frac{1}{\lambda_n}\lambda_n \tilde{B}_n \geq \sum_{n\geq 0} \frac{1}{\lambda_n}\tilde{V}_n = 4\sum_{n\geq 0} \frac{\tilde{V}_n}{4+\frac{n^2}{n+1}} \tag{2.195}$$

As $\sum \tilde{B}_n < \infty$ almost surely, we have $\sum_{n\geq 0} \frac{\tilde{V}_n}{4+\frac{n^2}{n+1}} < +\infty$ almost surely, and necessarily $V_\infty = 0$ almost surely. Then, almost surely

$$\frac{1}{8}\frac{n^2}{L\rho_K^2}(f(x_n) - f^*) \overset{a.s.}{\to} 0. \tag{2.196}$$

Finally, we get the result of Proposition 2.G.1

$$f(x_n) - f^* \overset{a.s.}{=} o\left(\frac{1}{n^2}\right). \tag{2.197}$$

□

### 2.G.4 Strongly convex - almost sure

Similarly to Section 2.G.3, we extend statement (2.8) of Theorem 2.3.3 to get a new, asymptotically better, almost sure convergence result.

**Proposition 2.G.2.** *Assume $f$ is $L$-smooth, $\mu$-strongly convex, and that $\tilde{\nabla}_K$ verifies the SGC for $\rho_K \geq 1$. Then SNM (Algorithm 3) with parameter setting $\alpha = \frac{1}{1+\frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}}$, $s = \frac{1}{\rho_K L}$, $\beta = 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}$ and $\eta = \frac{1}{\rho_K\sqrt{\mu L}}$ generates a sequence $(x_n)_{n\in\mathbb{N}}$ such that for all $\varepsilon > 0$, we have*

$$f(x_n) - f^* \overset{a.s.}{=} o\left((q+\varepsilon)^n\right), \tag{2.198}$$

$$\|z_n - x^*\|^2 \overset{a.s.}{=} o\left((q+\varepsilon)^n\right) \tag{2.199}$$

*where $q := 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}$.*

*Proof.* We use the following Lyapunov function

$$E_n := f(x_n) - f^* + \frac{\mu}{2}\|z_n - x^*\|^2. \tag{2.200}$$

We set $q := 1 - \frac{1}{\rho_K}\sqrt{\frac{\mu}{L}}$. We fix $\varepsilon' > 0$. By the Markov inequality and Equation (2.154), we get

$$\mathbb{P}\left(E_n \geq (q+\varepsilon')^n E_0\right) \leq \frac{\mathbb{E}[E_n]}{(q+\varepsilon')^n E_0} \leq \left(\frac{q}{q+\varepsilon'}\right)^n. \tag{2.201}$$

We sum on $n \geq 0$

$$\sum_{n\geq 0} \mathbb{P}\left(E_n \geq (q+\varepsilon')^n E_0\right) \leq \sum_{n\geq 0}\left(\frac{q}{q+\varepsilon'}\right)^n < +\infty. \tag{2.202}$$

Setting $A_n := \{E_n \geq (q+\varepsilon')^n E_0\}$, we have by the Borel Cantelli Lemma that $\mathbb{P}\left(\limsup A_n\right) = 0$, which implies $\mathbb{P}\left(\liminf A_n^c\right) = 1$, where $A_n^c$ is the complementary of $A_n$. In other words, as $A_n^c := \{E_n < (q+\varepsilon')^n E_0\}$, then for almost every $\omega \in \Omega$, $\exists N_0(\omega) \in \mathbb{N}$ such that for all $n \geq N_0(\omega)$, we have

$$E_n(\omega) < (q+\varepsilon')^n E_0. \tag{2.203}$$

Thus, we have

$$\frac{E_n(\omega)}{(q+2\varepsilon')^n} < \left(\frac{q+\varepsilon'}{q+2\varepsilon'}\right)^n E_0 \underset{n\to+\infty}{\to} 0 \tag{2.204}$$

The right term is independent of $\omega$, so almost surely, we have

$$E_n = o\left((q+2\varepsilon')^n\right)$$

Now fix $\varepsilon = 2\varepsilon'$ and we get

$$E_n = o\left((q+\varepsilon)^n\right), \tag{2.205}$$

and thus the result by definition of $E_n$. □

## 2.H Proofs of Section 2.4

In this section, we will prove our results that establish a link between the strong growth condition SGC and the average correlation between gradients, by exploiting the finite sum structure (FS).

### 2.H.1 Proof of Proposition 2.4.1

In order to demonstrate Proposition 2.4.1, we first establish Lemma 2.H.1.

**Lemma 2.H.1.** *Let $\{a_i\}_{i=1}^N$ be a sequence of vectors in $\mathbb{R}^d$ and $K \in \mathbb{N}$. We define $\mathcal{B}(K,N) = \{B \subseteq \{1,\dots,N\} | Card(B) = K\}$. Then, we have*

$$\left\|\sum_{i=1}^N a_i\right\|^2 = \frac{N}{K}\frac{1}{\binom{N}{K}}\sum_{B\in\mathcal{B}(K,N)}\left\|\sum_{i\in B} a_i\right\|^2 + 2\frac{N-K}{N-1}\sum_{1\leq i<j\leq N}\langle a_i, a_j\rangle. \tag{2.206}$$

*Proof.* We fix $\{i_1,\dots,i_K\} \in \mathcal{B}(K,N)$. We have

$$\left\|\sum_{i\in\{i_1,\dots,i_K\}} a_i\right\|^2 = \left\|\sum_{i=1}^N a_i - \sum_{i\notin\{i_1,\dots,i_K\}} a_i\right\|^2 \tag{2.207}$$

$$= \left\|\sum_{i=1}^N a_i\right\|^2 + \left\|\sum_{i\notin\{i_1,\dots,i_K\}} a_i\right\|^2 - 2\sum_{i=1}^N\sum_{j\notin\{i_1,\dots,i_K\}}\langle a_i, a_j\rangle \tag{2.208}$$

$$= \left\|\sum_{i=1}^N a_i\right\|^2 + \sum_{i\notin\{i_1,\dots,i_K\}}\|a_i\|^2 + \sum_{\substack{i,j\notin\{i_1,\dots,i_K\}\\ i\neq j}}\langle a_i, a_j\rangle - 2\sum_{i=1}^N\sum_{j\notin\{i_1,\dots,i_K\}}\langle a_i, a_j\rangle. \tag{2.209}$$

We sum over all the possible $B = \{i_1, \dots, i_K\} \in \mathcal{B}(K, N)$. Note that $|\mathcal{B}(K, N)| = \binom{N}{K}$. We split each term in Equation (2.209), first

$$\sum_{B \in \mathcal{B}(K,N)} \left\| \sum_{i=1}^{N} a_i \right\|^2 = \binom{N}{K} \left\| \sum_{i=1}^{N} a_i \right\|^2 . \tag{2.210}$$

The sum of the second term of Equation (2.209) is

$$\sum_{B \in \mathcal{B}(K,N)} \sum_{i \notin \{i_1, \dots, i_K\}} \|a_i\|^2 = \sum_{B \in \mathcal{B}(N-K,N)} \sum_{i \in \{i_1, \dots, i_{N-K}\}} \|a_i\|^2 \tag{2.211}$$

$$= \binom{N-1}{N-K-1} \sum_{i=1}^{N} \|a_i\|^2 \tag{2.212}$$

$$= \binom{N}{K} \frac{N-K}{N} \sum_{i=1}^{N} \|a_i\|^2 \tag{2.213}$$

$$= \binom{N}{K} \frac{N-K}{N} \left\| \sum_{i=1}^{N} a_i \right\|^2 - \binom{N}{K} \frac{N-K}{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \langle a_i, a_j \rangle . \tag{2.214}$$

The second equality comes from how many times the index $i$ is picked by the sum. We thus count the number of set $\{i_1, \dots, i_{N-K}\} \in \{1, \dots, N\}^{N-K}$ such that $i$ belongs to this set. This amounts to compute the cardinal of the set $\{\{i, i_1, \dots, i_{N-K-1}\}, i_1, \dots, i_{N-K-1} \in \{1, \dots, N\} \backslash \{i\}\}$. This set has the same size as $\mathcal{B}(N-K-1, N-1)$, which is $\binom{N-1}{N-K-1}$.

The sum of the third term of Equation (2.209) is

$$\sum_{B \in \mathcal{B}(K,N)} \sum_{\substack{i,j \notin \{i_1, \dots, i_K\} \\ i \neq j}} \langle a_i, a_j \rangle = \sum_{B \in \mathcal{B}(N-K,N)} \sum_{\substack{i,j \in \{i_1, \dots, i_{N-K}\} \\ i \neq j}} \langle a_i, a_j \rangle \tag{2.215}$$

$$= \binom{N-2}{N-K-2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \langle a_i, a_j \rangle \tag{2.216}$$

$$= \binom{N}{K} \frac{(N-K)(N-K-1)}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \langle a_i, a_j \rangle . \tag{2.217}$$

Here, the second equality comes from the fact that we compute the size of the set $\{\{i, j, i_1, \dots, i_{N-K-2}\}, i_1, \dots, i_{N-K-2} \in \{1, \dots, N\} \backslash \{i, j\}\}$, which is of the same size as $\mathcal{B}(N-K-2, N-2)$, which is $\binom{N-2}{N-K-2}$.

Finally, we compute the sum of the fourth term of Equation (2.209), using Equation (2.213)

$$\sum_{B \in \mathcal{B}(K,N)} \sum_{i=1}^{N} \sum_{j \notin \{i_1, \dots, i_K\}} \langle a_i, a_j \rangle = \sum_{i=1}^{N} \left\langle a_i, \sum_{B \in \mathcal{B}(K,N)} \sum_{j \notin \{i_1, \dots, i_K\}} a_j \right\rangle \tag{2.218}$$

$$= \sum_{i=1}^{N} \left\langle a_i, \binom{N}{K} \frac{N-K}{N} \sum_{j=1}^{N} a_j \right\rangle \tag{2.219}$$

$$= \binom{N}{K} \frac{N-K}{N} \left\| \sum_{i=1}^{N} a_i \right\|^2 . \tag{2.220}$$

Now that we have computed the sum of each term in Equation (2.209), we have

$$\sum_{B\in\mathcal{B}(K,N)}\left\|\sum_{i\in\{i_1,\dots,i_K\}}a_i\right\|^2=\binom{N}{K}\left\|\sum_{i=1}^N a_i\right\|^2+\binom{N}{K}\frac{N-K}{N}\left(\left\|\sum_{i=1}^N a_i\right\|^2-\sum_{\substack{i,j=1\\i\neq j}}^N\langle a_i,a_j\rangle\right) \tag{2.221}$$

$$+\binom{N}{K}\frac{(N-K)(N-K-1)}{N(N-1)}\sum_{\substack{i,j=1\\i\neq j}}^N\langle a_i,a_j\rangle-2\binom{N}{K}\frac{N-K}{N}\left\|\sum_{i=1}^N a_i\right\|^2 \tag{2.222}$$

$$=\binom{N}{K}\left(1-\frac{N-K}{N}\right)\left\|\sum_{i=1}^N a_i\right\|^2+\binom{N}{K}\frac{N-K}{N}\left(\frac{N-K-1}{N-1}-1\right)\sum_{\substack{i,j=1\\i\neq j}}^N\langle a_i,a_j\rangle \tag{2.223}$$

$$=\binom{N}{K}\frac{K}{N}\left\|\sum_{i=1}^N a_i\right\|^2-\binom{N}{K}\frac{K}{N}\frac{N-K}{N-1}\sum_{\substack{i,j=1\\i\neq j}}^N\langle a_i,a_j\rangle\,. \tag{2.224}$$

By rearranging the terms, we obtain the desired result

$$\left\|\sum_{i=1}^N a_i\right\|^2=\frac{N}{K}\frac{1}{\binom{N}{K}}\sum_{B\in\mathcal{B}(K,N)}\left\|\sum_{i\in\{i_1,\dots,i_K\}}a_i\right\|^2+\frac{N-K}{N-1}\sum_{\substack{i,j=1\\i\neq j}}^N\langle a_i,a_j\rangle\,. \tag{2.225}$$

□

The proof the Proposition 2.4.1 is simply an application of Lemma 2.H.1 with $a_i=\frac{1}{N}\nabla f_i(x)$. We obtain

$$\|\nabla f(x)\|^2=\left\|\frac{1}{N}\sum_{i=1}^N\nabla f_i(x)\right\|^2 \tag{2.226}$$

$$=\frac{N}{K}\frac{1}{\binom{N}{K}}\sum_{B\in\mathcal{B}(K,N)}\left\|\frac{1}{N}\sum_{i\in\{i_1,\dots,i_k\}}\nabla f_i(x)\right\|^2+\frac{N-K}{N-1}\frac{1}{N^2}\sum_{\substack{i,j=1\\i\neq j}}^N\langle\nabla f_i(x),\nabla f_j(x)\rangle \tag{2.227}$$

$$=\frac{K}{N}\frac{1}{\binom{N}{K}}\sum_{B\in\mathcal{B}(K,N)}\left\|\frac{1}{K}\sum_{i\in\{i_1,\dots,i_k\}}\nabla f_i(x)\right\|^2+\frac{N-K}{N-1}\frac{1}{N^2}\sum_{\substack{i,j=1\\i\neq j}}^N\langle\nabla f_i(x),\nabla f_j(x)\rangle \tag{2.228}$$

$$=\frac{K}{N}\mathbb{E}\left[\left\|\tilde{\nabla}_K(x)\right\|^2\right]+\frac{N-K}{N-1}\frac{1}{N^2}\sum_{\substack{i,j=1\\i\neq j}}^N\langle\nabla f_i(x),\nabla f_j(x)\rangle \tag{2.229}$$

$$=\frac{K}{N}\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right]+\frac{2}{N^2}\frac{N-K}{N-1}\sum_{1\leq i<j\leq N}\langle\nabla f_i(x),\nabla f_j(x)\rangle. \tag{2.230}$$

The expectation is taken with respect to the uniform low on $\mathcal{B}(K,N)$.

### 2.H.2 Proof of Proposition 2.4.2

In this part, we demonstrate Proposition 2.4.2. The result is a direct consequence of the RACOGA condition. Indeed, considering batch of size 1, by Proposition 2.4.1 we have $\forall x\in\mathbb{R}^d$

$$\|\nabla f(x)\|^2=\frac{1}{N}\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right]+\frac{2}{N^2}\sum_{1\leq i<j\leq N}\langle\nabla f_i(x),\nabla f_j(x)\rangle. \tag{2.231}$$

Now recall the RACOGA condition

$$\forall x \in \mathbb{R}^d,\ \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2} \geq c. \tag{RACOGA}$$

We inject RACOGA in Equation (2.231) to get

$$\|\nabla f(x)\|^2 \geq \frac{1}{N}\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right] + c\frac{2}{N^2}\sum_{i=1}^N \|\nabla f_i(x)\|^2 \tag{2.232}$$

$$= \frac{1}{N}\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right] + c\frac{2}{N}\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right] \tag{2.233}$$

$$= \frac{1}{N}\left(1+2c\right)\mathbb{E}\left[\left\|\tilde{\nabla}_1(x)\right\|^2\right]. \tag{2.234}$$

From Equation (2.234), that holds $\forall x \in \mathbb{R}^d$, we deduce that $f$ satisfy SGC with $\rho_1 \leq \frac{N}{1+2c}$.

### 2.H.3 Proof of Lemma 2.4.1

In this part, we demonstrate Lemma 2.4.1. Assume that for batches of size 1, $f$ verifies a $\rho_1$-SGC, *i.e.*

$$\forall x \in \mathbb{R}^d,\ \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x)\|^2 \leq \rho_1 \|\nabla f(x)\|^2. \tag{2.235}$$

By Proposition 2.4.1, we have

$$\|\nabla f(x)\|^2 = \frac{K}{N}\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] + \frac{2}{N^2}\frac{N-K}{N-1}\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle. \tag{2.236}$$

Moreover, by developing the squared norm of $\nabla f(x)$ and rearranging, we get

$$\frac{1}{N}\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle = \frac{N}{2}\left\|\nabla f(x)\right\|^2 - \frac{1}{2N}\sum_{i=1}^N \left\|\nabla f_i(x)\right\|^2, \tag{2.237}$$

hence, by reinjecting (2.237) into (2.236),

$$\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] = \frac{N}{K}\left(1-\frac{N-K}{N-1}\right)\|\nabla f(x)\|^2 + \frac{1}{NK}\frac{N-K}{N-1}\sum_{i=1}^N \|\nabla f_i(x)\|^2 \tag{2.238}$$

$$\leq \frac{N}{K}\left(1-\frac{N-K}{N-1}+\frac{1}{N}\frac{N-K}{N-1}\rho_1\right)\|\nabla f(x)\|^2 \tag{2.239}$$

$$\leq \frac{1}{K(N-1)}\left(N(K-1)+(N-K)\rho_1\right)\|\nabla f(x)\|^2 \tag{2.240}$$

using the SGC assumption for batches of size 1.

Finally we deduce Lemma 2.4.1: the SGC is verified for every size of batch $K \geq 1$ and we have

$$\rho_K \leq \frac{1}{K(N-1)}\left(\rho_1(N-K)+(K-1)N\right). \tag{2.241}$$

**Remark 2.H.1.** *Lemma 2.4.1 offers an indirect way to demonstrate Corollary 2.4.1. Indeed, the result of Proposition 2.4.1 for the special case of batches of size* 1 *is easily computed, as we have*

$$\|\nabla f(x)\|^2 = \frac{1}{N^2}\sum_{i=1}^N \|\nabla f_i(x)\|^2 + \frac{2}{N^2}\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle. \tag{2.242}$$

*Thus, when PosCorr is verified, with batches of size 1, $f$ verifies SGC with constant $\rho_1 = N$. By applying Lemma 2.4.1 with $\rho_1 = N$, we get that for batches of size $K$, $f$ verifies SGC) with constant*

$$\rho_K \leq \frac{1}{K(N-1)} \left(N(N-K) + (K-1)N\right) = \frac{N}{K}. \tag{2.243}$$

*For the clarity of our presentation, we choose to present Proposition 2.4.1 before Lemma 2.4.1, even if it can be seen as a corollary of this result.*

One can obtain a reciprocal result of Lemma 2.4.1.

**Proposition 2.H.1.** *Assume that for batches of size $K$, $f$ verifies the SGC with constant $\rho_K$. Then, for batches of size 1, $f$ verifies the SGC with a constant $\rho_1$ which satisfies*

$$\rho_1 \leq \frac{K\rho_K(N-1) - N(K-1)}{N-K}. \tag{2.244}$$

*Proof.* From Equality (2.238), we have

$$\mathbb{E}\left[\|\tilde{\nabla}_K(x)\|^2\right] = \frac{1}{KN}\frac{N-K}{N-1}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 + \frac{K-1}{K}\frac{N}{N-1}\|\nabla f(x)\|^2. \tag{2.245}$$

Using the SGC, we obtain

$$\frac{1}{K}\frac{N-K}{N-1}\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 \leq \left(\rho_K - \frac{K-1}{K}\frac{N}{N-1}\right)\|\nabla f(x)\|^2. \tag{2.246}$$

So, the SGC is verified with batches of size 1 and (2.244) is proved. □

**Remark 2.H.2.** *Proposition 2.H.1 indicates that satisfying the SGC for batches of size $K$ implies 2.2.2 if each $f_i$ is convex. Indeed, Proposition 2.H.1 induces that the SGC is verified for batches of size 1 with some $\rho_1 \geq 1$, i.e.*

$$\forall x \in \mathbb{R}^d,\ \frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 \leq \rho_1\|\nabla f(x)\|^2. \tag{2.247}$$

*Interpolation is then a direct consequence of evaluating Inequality (2.247) at some $x^* \in \arg\min f$. Indeed, it implies that each minimizer of $f$ is a critical point of each $f_i$. Convexity of the $f_i$ allow to conclude.*

## 2.H.4 Proof of Theorem 2.4.1

In this section, we demonstrate Theorem 2.4.1. Recall that we assume $f$ is convex, $L$-smooth, and that $\tilde{\nabla}_1$ verify SGC for $\rho_1 \geq 1$. We know that SNM with good choice of parameters (see Theorem 2.3.3) guarantees to reach an $\varepsilon$-precision (2.2) if

$$n \geq \rho_K\sqrt{\frac{2L}{\varepsilon}}\|x_0 - x^*\|. \tag{2.248}$$

Note that by Lemma 2.4.1, we know that $\rho_k$, the SGC constant associated to $\tilde{\nabla}_K$, exists, and that

$$\rho_K \leq \frac{1}{K(N-1)}\left(\rho_1(N-K) + (K-1)N\right). \tag{2.249}$$

So, in particular, we are ensured to reach an $\varepsilon$-precision if

$$n \geq \frac{1}{K(N-1)}\left(\rho_1(N-K) + (K-1)N\right)\sqrt{\frac{2L}{\varepsilon}}\|x_0 - x^*\| \tag{2.250}$$

$$= \underbrace{\left(\frac{N-K}{N-1} + \frac{N}{\rho_1}\frac{K-1}{N-1}\right)}_{:=\Delta_K}\frac{\rho_1}{K}\sqrt{\frac{2L}{\varepsilon}}\|x_0 - x^*\|. \tag{2.251}$$

Now, to translate the result in term of number of $\nabla f_i$ evaluated, note that each iteration requires to evaluate $K$ different $\nabla f_i$, because we consider batches of size $K$. Finally, the number of $\nabla f_i$ to evaluate is

$$\Delta_K \rho_1 \sqrt{\frac{2L}{\varepsilon}} \left\| x_0 - x^* \right\|. \tag{2.252}$$

### 2.H.5 Bound on RACOGA

In this section, we bound the ratio $\frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2}$, that is involved in the RACOGA condition.

**Proposition 2.H.2.** *Let $x \in \mathbb{R}^d \backslash \overline{\mathcal{X}}$. We have*

$$-\frac{1}{2} \leq \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2} \leq \frac{N-1}{2}. \tag{2.253}$$

*where $\overline{\mathcal{X}} = \{x \in \mathbb{R}^d, \forall i \in \{1,\dots,N\}, \|\nabla f_i(x)\| = 0\}$.*

*Proof.* The upper-bound relies on the inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, for $a, b \in \mathbb{R}^d$.

$$\sum_{1\leq i<j\leq N} \langle \nabla f_i(x), \nabla f_j(x)\rangle \leq \frac{1}{2} \sum_{1\leq i<j\leq N} \left( \|\nabla f_i(x)\|^2 + \|\nabla f_j(x)\|^2 \right) \tag{2.254}$$

$$= \frac{N-1}{2} \sum_{i=1}^N \|\nabla f_i(x)\|^2. \tag{2.255}$$

Using $\langle a, b\rangle \geq -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$, for $a, b \in \mathbb{R}^d$, we could get a lower-bound. This lower-bound is not tight when considering a sum of two or more scalar products. This is because the critical equality case occurs when $a = -b$. However considering at least 3 vectors, they cannot be respectively opposite to each other. Interestingly, Proposition 2.4.1 provides a way to get a tighter lower-bound.

By contradiction, assume there exists $x_\varepsilon \in \mathbb{R}^d \backslash \overline{\mathcal{X}}$ such that

$$\frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x_\varepsilon), \nabla f_j(x_\varepsilon)\rangle}{\sum_{i=1}^N \|\nabla f_i(x_\varepsilon)\|^2} = -\frac{1}{2} - \varepsilon \tag{2.256}$$

with $\varepsilon > 0$. Using Proposition 2.4.1 with batches of size 1, we have

$$\|\nabla f(x_\varepsilon)\|^2 = \frac{1}{N^2} \sum_{i=1}^N \|\nabla f_i(x_\varepsilon)\|^2 + \frac{2}{N^2} \sum_{1\leq i<j\leq N} \langle \nabla f_i(x_\varepsilon), \nabla f_j(x_\varepsilon)\rangle \tag{2.257}$$

$$= \frac{1}{N^2} \sum_{i=1}^N \|\nabla f_i(x_\varepsilon)\|^2 - \frac{2}{N^2} \left(\frac{1}{2} + \varepsilon\right) \sum_{i=1}^N \|\nabla f_i(x_\varepsilon)\|^2 \tag{2.258}$$

$$= -\frac{2\varepsilon}{N^2} \sum_{i=1}^N \|\nabla f_i(x_\varepsilon)\|^2 < 0. \tag{2.259}$$

We thus arrive at a contradiction, as $\|\nabla f(x_\varepsilon)\|^2$ is non negative. As a consequence, (2.256) cannot hold. Thus, at worst we have for all $x \in \mathbb{R}^d \backslash \overline{\mathcal{X}}$

$$-\frac{1}{2} \leq \frac{\sum_{1\leq i<j\leq N}\langle \nabla f_i(x), \nabla f_j(x)\rangle}{\sum_{i=1}^N \|\nabla f_i(x)\|^2}. \tag{2.260}$$

So, Proposition 2.H.2 is proved. □

## 2.I RACOGA values in linear regression

In this section, we give deeper insights considering RACOGA values in the case of the linear regression problem. Moreover, we investigate, in this linear regression context, the link between RACOGA values and the curvature.

We have $\{a_i, b_i\}_{i=1}^N$, where each $(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$, and we want to minimize $f$, with

$$f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(\langle a_i, x\rangle - b_i)^2. \tag{LR}$$

As mentioned in Section 2.5.1, in this case the correlation between gradients is directly linked to the correlation between data by

$$\underbrace{\langle \nabla f_i(x), \nabla f_j(x)\rangle}_{\text{gradient correlation}} = (\langle a_i, x\rangle - b_i)(\langle a_j, x\rangle - b_j) \quad \underbrace{\langle a_i, a_j\rangle}_{\text{data correlation}} \quad . \tag{2.261}$$

In particular uncorrelated data, *i.e.* $\langle a_i, a_j\rangle = 0$, will induce uncorrelated gradients, *i.e.* $\forall x \in \mathbb{R}^d$, $\langle \nabla f_i(x), \nabla f_j(x)\rangle = 0$. In this case, RACOGA is verified for $c = 0$. In this section, we will see that outside this special case of orthogonal data, the characterization of RACOGA values is a challenging problem.

### 2.I.1 Two functions in two dimensions

We study in this subsection the simplified case with $d = 2$, $N = 2$ and $b_1 = b_2 = 0$, *i.e.* two functions defined on $\mathbb{R}^2$ such that 0 is their unique minimizer. Formally the function we consider is the following

$$f(x) := \frac{1}{2}\left(\frac{1}{2}\langle a_1, x\rangle^2 + \frac{1}{2}\langle a_2, x\rangle^2\right). \tag{2f}$$

In this special case, we look at the gradient correlation, with same sign of the RACOGA values,

$$\Delta(x) := \langle \nabla f_1(x), \nabla f_2(x)\rangle = \langle a_1, x\rangle \langle a_2, x\rangle \langle a_1, a_2\rangle . \tag{2.262}$$

If $\langle a_1, a_2\rangle \neq 0$, $\Delta(x)$ is not identically equal to zero. Without loss of generality, for the following reasoning we can assume $\langle a_1, a_2\rangle > 0$. Choose $x_0 \in \mathbb{R}^2$ such that $\langle a_1, x_0\rangle > 0$ and $\langle a_2, x_0\rangle \neq 0$. The function $x \to \langle a_1, x\rangle$ has a kernel $a_1^\perp$, the orthogonal of $a_1$. Moving along this kernel, *i.e.* considering $x = x_0 + k$, $k \in a_1^\perp$, if $a_1$ and $a_2$ are not colinear, one can make the scalar product $\langle a_2, x\rangle$ be positive or negative while $\langle a_1, x\rangle > 0$ as it remains equal to $\langle a_1, x_0\rangle$. Therefore necessarily, if $\langle a_1, a_2\rangle \neq 0$ and $a_1$ and $a_2$ are not colinear, we have $\min_x \Delta(x) < 0$. So, if $\langle a_1, a_2\rangle \neq 0$, the minimum of the RACOGA values on the space is necessarily negative.

Figure 2.9 illustrates this behaviour. We observe that non orthogonality of $a_1$ and $a_2$ creates non positive and non negative areas of RACOGA values.

According to Theorem 2.4.1, non positive RACOGA values indicate a bad performance of SNM (Algorithm 3). The example of this section indicates that we can not hope to obtain theoretical results that would ensure high RACOGA values for any linear regression problem, and thus a good performance of SNM. In the next section we see that, nevertheless, we can expect the RACOGA values to be positive over most of the space.

### 2.I.2 RACOGA is high over most of the space

In Section 2.I.1, we considered a 2-dimensional example with 2-functions. Increasing dimension and adding functions, the problem of characterizing RACOGA values becomes harder. In the case of independent data, it is possible to give a theoretical result considering the RACOGA values in expectation over the data.

**Proposition 2.I.1.** *Let* $f(x) = \frac{1}{N}\sum_{i=1}^{N}(\Phi(x, a_i) - b_i)^2$ *where* $\{a_i, b_i\}_{i=1}^N \in \mathbb{R}^p \times \mathbb{R}$ *are i.i.d. and* $\Phi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ *is differentiable. We have*

$$\mathbb{E}[\sum_{1\leq i<j\leq N} \langle \nabla f_i(x), \nabla f_j(x)\rangle] = \frac{N(N-1)}{2}\left\|\mathbb{E}[(\Phi(x, a_1) - b_1)\nabla\Phi(x, a_1)]\right\|^2 \geq 0. \tag{2.263}$$

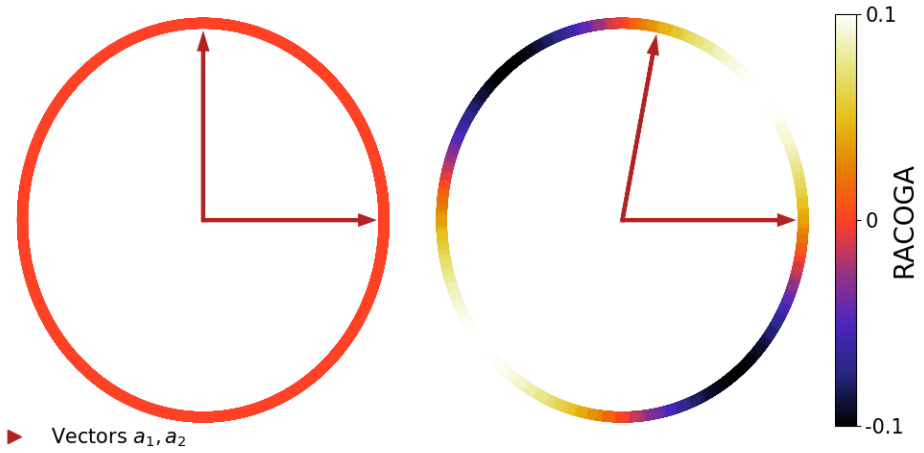



Figure 2.9: Illustration of the RACOGA values for problem (2f), along a circle around the solution. On the left part, $a_1$ and $a_2$ are orthogonal, inducing RACOGA is constant equal to zero. On the right part, $a_1$ and $a_2$ are slightly correlated, inducing positive and negative RACOGA values. Note that the non positive RACOGA areas exactly contain the points $x \in \mathbb{R}^d$ such that $\langle a_1, x\rangle \langle a_2, x\rangle \leq 0$.

*In particular if* $\Phi(x, a_i) = \langle x, a_i\rangle$ *and* $a_1 \sim \mathcal{N}(m, \Gamma)$, $b_1 \sim \mathcal{N}(m_b, \sigma_b^2)$ *with* $a_1 \perp\!\!\!\perp b_1$, *we have*

$$\mathbb{E}\left[\sum_{1\leq i<j\leq N} \langle \nabla f_i(x), \nabla f_j(x)\rangle\right] = \frac{N(N-1)}{2}\left\|\Gamma x + mm^t x - m_b m\right\|^2 \geq 0. \tag{2.264}$$

*In both cases, the expectation is taken with respect to the data* $\{a_i, b_i\}_{i=1}^N$.

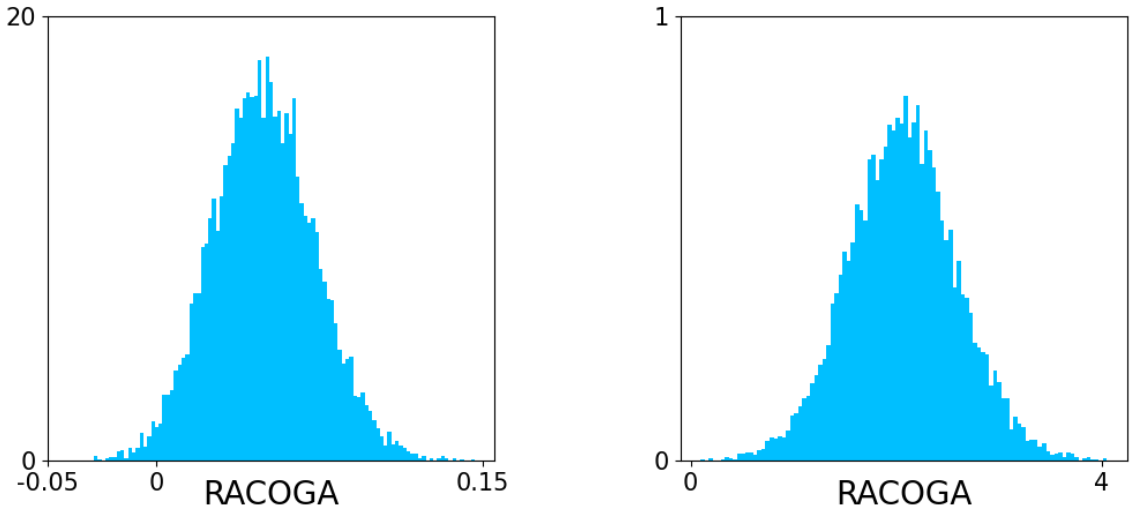
Proposition 2.I.1 indicates that when having a large amount of data, we can expect the RACOGA values to be positive over a large part of the space. The proof of Proposition 2.I.1 is deferred in Appendix 2.I.4.

To confirm this statement empirically, for a fixed dataset $\{a_i, b_i\}_{i=1}^N$, we compute the RACOGA values on a sphere whose center is a minimizer of the function. Note that by the linearity of the gradient, the RACOGA values taken on this sphere are invariant by homothety. We set the bias, *i.e.* the $\{b_i\}_{i=1}^N$, at zero. This forces zero to be a minimizer, without loss of generality. The function we consider is the following

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\langle a_i, x\rangle^2 . \tag{2.265}$$

On Figure 2.10, we run this experiment in the case where $\{a_i\}_{i=1}^N$ are drawn uniformly onto the sphere, inducing low correlation inside data, and also in the case it is drawn following a gaussian mixture law, inducing higher correlation inside data. We set $d = 1000$, $N = 100$. It is the same problem as for Figure 2.1, except that here we set the bias to zero. We sample 10000 points on the sphere. In both cases, RACOGA is almost only non negative. More, all the RACOGA values are positive on Figure 2.10b, *i.e.* for the correlated dataset. Note that the observations we made in Section 2.5.1 are consistent: correlated data induce higher RACOGA values (Figure 2.10b), whereas with uncorrelated data the RACOGA values are smaller (Figure 2.10a).

However, one should not conclude from Figure 2.10b that RACOGA values are positive everywhere, as there could be non positive RACOGA value areas that are so small that our sampled points never fall in. Even more, we should not conclude from the fact that the eventual areas of non positive RACOGA values are small that the optimisation algorithms never cross them. We show in the following section that, actually, these small non positive RACOGA value areas exist and **attract** the optimization algorithms, and that stochasticity prevents the algorithm to get stuck inside.




(a) Uncorrelated data

(b) Correlated data

Figure 2.10: Histogram distribution of the RACOGA values for points sampled uniformly on a sphere centered on a minimizer. On the left plot, the data are fewly correlated and the RACOGA values are mostly positive. On the right plot there is correlation inside data, and all RACOGA values are positive. Note that the RACOGA values are significantly higher on the right plot, because of the higher data correlation.

### 2.I.3 The curvature problem: first-order algorithms are attracted by low RACOGA value areas

In Section 2.I.2, we showed that in the case $d > N$ where $N$ is not too small, one can expect RACOGA values to be high over most of the space. This statement is reinforced in the presence of correlation inside data (Figure 2.10b).

However, if $d$ is high, even though we sample a large amount of points to evaluate RACOGA values, we could miss non positive RACOGA value areas if these areas are too small. On Figure 2.11, we see that these areas indeed exist. Moreover, although they are tiny with respect to the whole space (Section 2.I.2), deterministic algorithms, namely GD (Algorithm 7) and NM (Algorithm 8), dive into these areas and stayed trapped inside. Strikingly, SNM behaves differently and it manages not to get stuck in the same area.

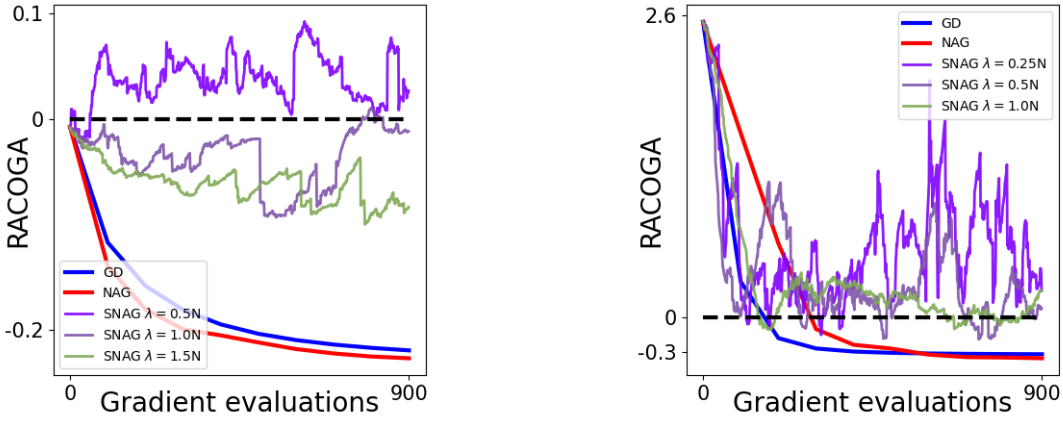
In the remaining of this section, we give an explanation of the behaviour observed on Figure 2.11.

**First-order algorithms fall in low curvature area** In the case of linear regression, which amounts to minimizing a quadratic function, it is well known that first-order algorithms, namely algorithms that use only gradient information, converge faster in the direction of high curvature, *i.e.* directions such that the Hessian matrix has a high eigenvalue. We illustrate this phenomenon on Figure 2.12, where we plot the first iterations of GD (Algorithm 7) and NM (Algorithm 8) applied to the function

$$g(x) = \frac{1}{2} x^T A x \tag{2.266}$$

where $A = \begin{pmatrix} L & 0 \\ 0 & \mu \end{pmatrix}$, $0 < \mu < L$. The algorithms GD and NM dive and stay in a low curvature zone. However, note that the stochasticity of SNM makes it unstable enough to not follow the exact same path.

**The RACOGA-Curvature link** In this paragraph, we connect our observations about curvature and RACOGA values. Intuitively, for a point $x \in \mathbb{R}^d$ such that RACOGA is small, *i.e.* gradients are on average anti correlated, the gradients will compensate each other. Thus, we can expect that around this point, the gradient will have low values. Considering linear regression, this induce that non positive RACOGA areas




(a) Spherical data

(b) Gaussian mixture data

Figure 2.11: Illustration of RACOGA values taken along iterations of GD (Algorithm 7), NM (Algorithm 8) and SNM (Algorithm 3, batch size 1) for the linear regression problem. On the left plot, the data are fewly correlated while on the right plot there is correlation inside data. Note that while deterministic algorithms, *i.e.* GD and NM, dive and stay in a negative RACOGA value area, the stochasticity of SNM enables it to not to be trapped in the same zone.

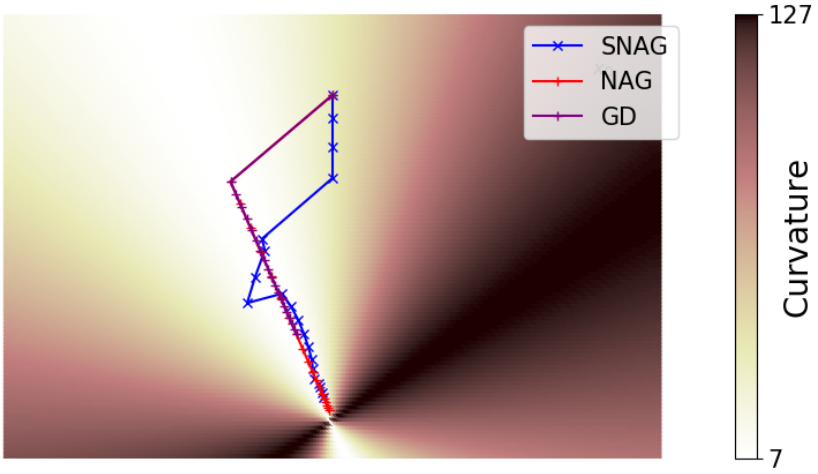



Figure 2.12: Illustration of the iterations of the trajectories of GD (Algorithm 7), NM (Algorithm 8) and SNM (Algorithm 3) applied the function (2.266). We also display the curvature of the function, which we define at $x \in \mathbb{R}^2_*$ as $\frac{x^T A x}{\|x\|^2}$. Note that the deterministic algorithms GD and NM dive in the direction of smallest curvature, and then move following this direction. Note also that the instability of SNM enables itself to follow less strictly this smallest curvature ravine.

tend to produce low curvature areas. We illustrate this phenomenon on Figure 2.13, where we consider problem 2f. Actually if $a_1$ and $a_2$ have the same norm, the lowest RACOGA direction coincides exactly with the lowest curvature direction. As we mentioned in the previous paragraph that deterministic algorithms dive and stay in a low curvature zone, they actually dive and stay in low RACOGA areas. The instability provided by stochasticity allows SNM not to get stuck inside these low RACOGA areas.

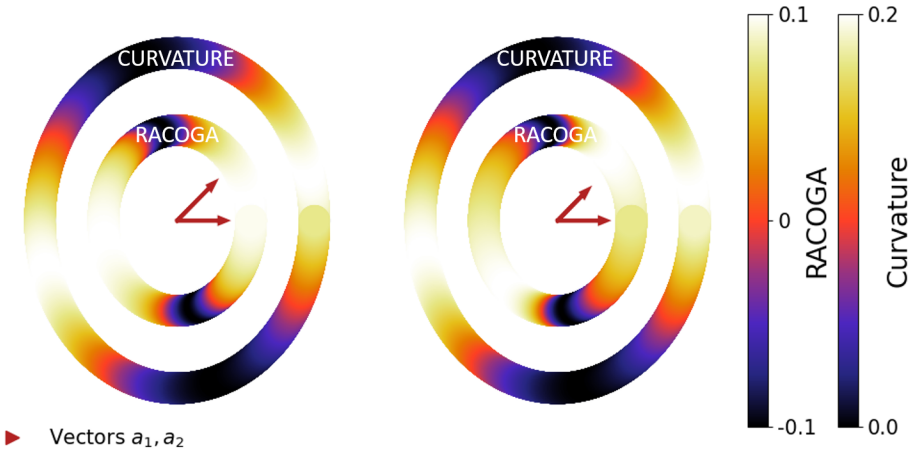



Figure 2.13: Comparison of the RACOGA and curvature values for problem (2f), along a circle around the solution. On the left plot, $a_1$ and $a_2$ have the same norm, which is not the case on the right plot. Note that the low curvature zone are close to the non positive RACOGA areas, and are exactly the same when $a_1$ and $a_2$ have same norm.

### 2.I.4 Proof of Proposition 2.I.1

**Proof of Equation (2.263)** We consider the least square problem defined by

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}(\Phi(x, a_i) - b_i)^2, \tag{2.267}$$

with $\Phi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ differentiable and $\{a_i, b_i\}_{i=1}^N$, random variables drawn i.i.d.

By the independence of the variable, we have for $i \neq j$

$$\begin{aligned}
\mathbb{E}[\langle \nabla f_i(x), \nabla f_j(x)\rangle] &= \langle \mathbb{E}[\nabla f_i(x)], \mathbb{E}[\nabla f_j(x)]\rangle && (2.268)\\
&= \|\mathbb{E}[\nabla f_1(x)]\|^2 && (2.269)\\
&= \|\mathbb{E}[(\Phi(x, a_1) - b_1)\nabla\Phi(x, a_1)]\|^2 \geq 0 && (2.270)
\end{aligned}$$

Finally, we sum over $N$ to get Equation (2.263).

**Proof of Equation (2.264)** First, we compute for $a = (a^{(1)}, \dots, a^{(d)}) \in \mathbb{R}^d$ and $b \in \mathbb{R}$

$$\mathbb{E}[(\langle a, x\rangle - b)a] = (\mathbb{E}[\sum_i a^{(1)}a^{(i)}x_i] - \mathbb{E}[a^{(1)}b], \dots, \mathbb{E}[\sum_i a^{(d)}a^{(i)}x_i] - \mathbb{E}[a^{(d)}b]). \tag{2.271}$$

We have $a \sim \mathcal{N}(m, \Gamma)$, $b \sim \mathcal{N}(m_b, \sigma_b^2)$ with $a \perp\!\!\!\perp b$. So we can deduce that $\mathbb{E}[(a^{(i)})^2] = \Gamma_{i,i} + m_i^2$, and $\mathbb{E}[a^{(i)}a^{(j)}] = \Gamma_{i,j} + m_i m_j$. Thereby, we have for a fixed $x \in \mathbb{R}^d$

$$\begin{aligned}
\mathbb{E}[(\langle a, x\rangle - b)a] &= (\mathbb{E}[\sum_i a^{(1)}a^{(i)}x_i] - \mathbb{E}[a^{(1)}b], \dots, \mathbb{E}[\sum_i a^{(d)}a^{(i)}x_i] - \mathbb{E}[a^{(d)}b]) && (2.272)\\
&= (\sum_i \mathbb{E}[a^{(1)}a^{(i)}]x_i - \mathbb{E}[a^{(1)}]\mathbb{E}[b], \dots, \sum_i \mathbb{E}[a^{(d)}a^{(i)}]x_i - \mathbb{E}[a^{(d)}]\mathbb{E}[b]) && (2.273)\\
&= (\sum_i \Gamma_{1,i}x_i + \sum_i m_1 m_i x_i - m_b m_1, \dots, \sum_i \Gamma_{d,i}x_i + \sum_i m_d m_i x_i - m_b m_d) && (2.274)\\
&= (\sum_i \Gamma_{1,i}x_i, \dots, \sum_i \Gamma_{d,i}x_i) + (m_1, \dots, m_d)\sum_i m_i x_i - m_b(m_1, \dots, m_d) && (2.275)\\
&= \Gamma x + (m^T x - m_b)m && (2.276)
\end{aligned}$$

From the previous computation and Equation (2.263), we deduce Equation (2.264).

# Chapter 3

# Study of Nesterov Momentum for Strongly Quasar Convex Optimization and the Continuous-Discrete Rupture

We recall that the content of this chapter is based on the following paper.

> J. Hermant, J-F. Aujol, C. Dossal, and A. Rondepierre.
> **Study of the behaviour of Nesterov Accelerated Gradient in a non-convex setting: the strongly quasar-convex case**.
> In *Optimization*, 2026

## 3.1 Introduction

In this chapter, we start addressing the limitations raised in Section 1.1.2. We recall that we are interested in the following unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := f^* \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is such that $\arg\min_{x\in\mathbb{R}^d} f(x)$ is nonempty. In this chapter, we begin our investigation of using Nesterov Momentum (NM) to minimize *strongly quasar-convex* functions, which we recall are defined by the following inequality:

$$f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2 \tag{3.1}$$

with $x^*$ a minimizer, $x$ an arbitrary point belonging to $\mathbb{R}^d$, $\tau \in (0, 1]$, $\mu \geq 0$. Strong convexity can be recovered from this equation, fixing $\tau = 1$, and requiring it to hold when replacing $x^*$ by any arbitrary point $y \in \mathbb{R}^d$.

**Contributions** We provide the following contributions.

1. We identify a curvature assumption such that Nesterov's momentum provides an accelerated rate of convergence for the class of smooth strongly quasar-convex functions. We extend this result to composite non differentiable functions, in which case specific difficulties appear.
2. We provide a continuous analysis of the high-resolution ODE associated to NM in the strongly quasar-convex setting. We highlight limitations of the ODE framework as a tool to tune optimization algorithms when it comes to non-convex optimization.
3. We present new properties of strongly quasar-convex functions, together with properties considering more general non-convex functions.

**Related work** We detail existing works about convergence of first-order algorithms applied to strongly quasar-convex functions. In the $L$-smooth and $(1,\mu)-$strongly quasar-convex (3.1) setting, (Zhang et al., 2019b) uses a Runge Kutta discretization procedure of the Heavy Ball ordinary differential equation to get a $1-(\frac{\mu}{L})^\tau$ convergence rate, $\tau = \frac{s+3}{2(s+1)}$ where $s$ is such that the $s$th derivative is Lipschitz. This means that in this case the function must have high-order smoothness to be close to the $1-\sqrt{\frac{\mu}{L}}$ accelerated rate. The authors of (Wang and Wibisono, 2023) apply the *continuized acceleration framework* (Even et al., 2021) to $(\tau,\mu)$-strongly quasar-convex functions, which consists in a continuous stochastic differential equation approach leading to a stochastic version of the Nesterov Momentum. This stochastic version achieves an accelerated rate $1-\tau\sqrt{\frac{\mu}{L}}$ in expectation, and convergence of iterations with high probability can be deduced. Also, (Guminov et al., 2023; Nesterov et al., 2021) show accelerated rates in terms of the number of iterations, but each iteration relies on a low dimensional sub-optimization problem to solve. In (Hinder et al., 2020), the cost of this sub-problem (binary line search) is explicitly computed, and the authors show that their algorithm achieves an almost optimal rate (up to a log factor) for $L$ smooth $(\tau,\mu)$-strongly quasar-convex function in terms of gradient and function evaluations. However, it is not clear whether NM needs to solve a low dimensional sub-optimization problem at each iteration in order to achieve accelerated convergence. Finally, an independent work (Gupta and Wojtowytsch, 2025) studies convergence of smooth strongly quasar-convex functions without assuming that the functions have a unique minimizer, in the restricted case $\tau = 1$.

## 3.2 Preliminaries

We recall and state the assumptions we use throughout the chapter. A function $f:\mathbb{R}^d\to\mathbb{R}$ is said $L$-smooth for some $L>0$ if $f$ is $C^1$ and has an $L$-Lipschitz gradient:

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

Note that $f$ is $L$-smooth if and only if $f$ admits lower and upper quadratic bounds parameterized by $L$ at every point. More precisely:

**Property 3.2.1.** *Let $f:\mathbb{R}^d\to\mathbb{R}$. $f$ is $L$-smooth for some $L>0$ if and only if it verifies for all $x,y$ in $\mathbb{R}^d$:*

$$f(x)+\langle\nabla f(x),y-x\rangle-\frac{L}{2}\|x-y\|^2\leq f(y)\leq f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|x-y\|^2. \tag{3.2}$$

*Proof.* The fact that $L$-smooth implies the inequality (3.2) is just the well-known descent lemma, see for example (Nesterov, 2013). The converse is also well known when dealing with convex functions. Less trivially, it still holds in the non-convex case. To the best of our knowledge the equivalence is not proved in the literature, but a proof has been proposed online[1] which we restate here.
Assume first that $f$ verifies (3.2) with $L=1$. Let $d\in\mathbb{R}^d$. Evaluating (3.2) at 4 different pairs of points:

$$f(y+d)-f(x)-\langle\nabla f(x),y+d-x\rangle\leq\frac{1}{2}\|y+d-x\|^2 \tag{3.3}$$

$$f(x-d)-f(y)-\langle\nabla f(y),x-d-y\rangle\leq\frac{1}{2}\|y+d-x\|^2 \tag{3.4}$$

$$-(f(y+d)-f(y)-\langle\nabla f(y),d\rangle)\leq\frac{1}{2}\|d\|^2 \tag{3.5}$$

$$-(f(x-d)-f(x)-\langle\nabla f(x),-d\rangle)\leq\frac{1}{2}\|d\|^2. \tag{3.6}$$

Adding all these inequalities yields:

$$\langle\nabla f(x)-\nabla f(y),x-y-2d\rangle\leq\|y+d-x\|^2+\|d\|^2 \tag{3.7}$$

Set $g:=\nabla f(x)-\nabla f(y)$ and choose $d$ such that $x-y-2d=g$. Then $d=\frac{1}{2}(x-y-g)$ and $y+d-x=-\frac{1}{2}(x-y+g)$. This results in:

$$\|g\|^2\leq\frac{1}{4}\|x-y+g\|^2+\frac{1}{4}\|x-y-g\|^2=\frac{1}{2}\|x-y\|^2+\frac{1}{2}\|g\|^2 \tag{3.8}$$

[1] Characterization of Lipschitz derivative, `https://math.stackexchange.com/q/4264948` (version: 2021-10-01), Mathematics Stack Exchange

which implies that $f$ has a 1-Lipschitz gradient. It is straightforward to extend it to functions verifying (3.2) with arbitrary $L_0 > 0$, as it implies $\frac{f}{L_0}$ verify (3.2) with $L = 1$, inducing $\frac{\nabla f}{L_0}$ is 1-Lipschitz, thus inducing the result. □

This provides quadratic lower and upper bounds on the function, both parameterized by the constant $L$. However, we will later consider different parameterizations for these bounds. To do so we now introduce the class of $(a, L)$-*curvature-bounded functions*.

**Definition 3.2.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and $(a, L)$ two real constants with $L > 0$ and $a \leq L$. The function $f$ is said to be $(a, L)$-curvature-bounded if it satisfies for all $x, y \in \mathbb{R}^d$:*

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{a}{2}\|x - y\|^2 \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \tag{3.9}$$

This is a generalization of $L$-smoothness as we allow for different characterizations of lower curvature. In particular, observe that a $(-L, L)$-curvature-bounded function is exactly an $L$-smooth function, a $(0, L)$-curvature-bounded function is an $L$-smooth convex function, and a $(\mu, L)$-curvature-bounded function with $\mu > 0$ is an $L$-smooth and $\mu$-strongly convex function. We note that it also corresponds to verifying both $L$-smoothness and $a$-weak convexity.

### 3.2.1 Strong convexity and the question of acceleration

In the chapter, we study a relaxation of a well known property called strong convexity, which we recall below. Note that as we mainly studied the differentiable function case, we state the definitions in this case. In Section 3.3.3 we will consider possibly non differentiable functions.

**Definition 3.2.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. The function $f$ is said $\mu$-strongly convex for some $\mu > 0$ if it satisfies:*

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d,\ f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \leq f(y). \tag{3.10}$$

This is a stronger hypothesis than convexity (which corresponds to the above inequality with $\mu = 0$). Strongly convex functions are not only lower bounded by linear approximations, but by quadratic approximations. In particular, these functions do verify the $\mu$-quadratic growth hypothesis.

**Definition 3.2.3.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $X^* = \operatorname{argmin} f \neq \emptyset$ and $f^* = \min f$. The function $f$ has $\mu$-quadratic growth for some $\mu > 0$ if:*

$$\forall x \in \mathbb{R}^d,\ \frac{\mu}{2} d(x, X^*)^2 \leq f(x) - f^*. \tag{3.11}$$

It is straightforward to see that under the strong convexity assumption, the function $f$ has a unique minimizer. The quadratic growth property ensures that, around this minimizer, the function cannot become flatter than a quadratic. This provides control over how fast the gradient can vanish to zero when approaching the minimizer, allowing to achieve linear convergence speed with gradient descent. More precisely, for $L$-smooth and $\mu$-strongly convex functions, the gradient descent generates a sequence $\{x_n\}_{n\in\mathbb{N}}$ that yields a linear convergence $f(x_n) - f^* \leq \mathcal{O}\left((1 - \frac{\mu}{L})^n\right)$. As mentioned in introduction, using other first-order algorithms, this rate can be improved into a $\mathcal{O}\left((1 - \sqrt{\frac{\mu}{L}})^n\right)$ rate. Notice that we necessarily have $\mu \leq L$. Moreover, in large dimension, the ratio $\frac{\mu}{L}$ (which represents the inverse of the conditioning of the function to minimize) can be very small, this is actually a significant improvement.

**Acceleration for relaxations of strong convexity** In practice however few problems really suit the strong convexity hypothesis. This motivates to consider weaker assumptions. $\mu$-strong convexity provides a lower bound on the function, while $L$-smoothness provides an upper bound. Many relaxations of $\mu$-strongly convex and $L$-smooth functions generalize this fact by defining another pair of assumptions, a lower one parameterized by a constant $\mu \geq 0$ and an upper one parameterized by a constant $L \geq 0$. In many cases, the property $\mu \leq L$ and the convergence rate associated to gradient descent in $\mathcal{O}\left((1 - \frac{\mu}{L})^n\right)$ remains true. This

enables to generalize, in these cases, the characterization of *acceleration* as the exchange of $\frac{\mu}{L}$ for $\sqrt{\frac{\mu}{L}}$. See (Guille-Escuret et al., 2021) for an insightful discussion about lower and upper conditions.

An example of such relaxation is the class of aforementioned $L$-smooth functions having a $\mu$-quadratic growth. However this is a too weak relaxation, since we lose almost all control over the function. In particular critical points can be non-global minima.

This is why we consider slightly stronger hypotheses, such as the class of $\mu$-Polyak-Łojasiewicz functions ($\mu$-PL), which is a non-convex relaxation of the class of $\mu$-strongly convex functions:

**Definition 3.2.4.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with $X^* = \arg\min\ f \neq \emptyset$ and $f^* = \min\ f$. The function $f$ is $\mu$-Polyak-Łojasiewicz ($\mu$-PL) for some $\mu > 0$, if:*

$$\forall x \in \mathbb{R}^d,\ f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{3.12}$$

Note that $\mu$-PL functions have a $\mu$-quadratic growth (Garrigos, 2023). The Łojasiewicz property (Łojasiewicz, 1963, 1993) is a key tool in the mathematical analysis of continuous and discrete dynamical systems, initially introduced to prove the convergence of the trajectories for the gradient flow of analytic functions. The Polyak-Łojasiewicz property is nothing more than the global version of the Łojasiewicz property with an exponent $\frac{1}{2}$, and appears in important practical problems (Liu et al., 2022; Barboni et al., 2022).

Convergence results for $\mu$-PL functions can be deduced from works in the literature that consider more general assumptions (Bolte et al., 2017). Specifically focusing on this class, (Karimi et al., 2016) that gradient descent ensures, for $L$-smooth functions satisfying $\mu$-PL property, a linear convergence: $f(x_n) - f^* \leq (1 - \frac{\mu}{L})^n (f(x_0) - f^*)$. Importantly, in (Yue et al., 2023) is computed a lower bound of the number of gradient queries needed to achieve, with a first-order method, a point $\hat{x}$ such that $f(\hat{x}) - f^* \leq \varepsilon(f(x_0) - f^*)$ for some $\varepsilon > 0$. They show that for every first-order method, there exists a function such that this number of gradient queries is of the order $\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)$. This bound is achieved, up to a constant, by gradient descent. Strikingly, it induces that for these functions, the Nesterov Momentum algorithms are prevented from achieving an accelerated linear convergence of the form $f(x_n) - f^* \leq K_1(1 - K_2\sqrt{\frac{\mu}{L}})^n (f(x_0) - f^*)$, where $K_1 > 0$ and $0 < K_2 \leq 1$ are constants independent of $\mu$ and $L$. This motivates to use stronger assumptions in order to achieve this acceleration, and leads us to consider a stronger hypothesis, namely the strong quasar convexity, which is another relaxation of strong convexity.

### 3.2.2 A relaxation of strong convexity: strong quasar convexity

In this section we define the notion of strong and non-strong quasar convexity. We include a short discussion about its geometrical properties, which will be further developed in section 3.5.

**Definition 3.2.5.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function such that $X^* = \operatorname{argmin} f \neq \emptyset$ and $f^* = \min\ f$. Let $x^*$ be a minimizer of $f$ and $\tau \in (0, 1]$, $\mu > 0$. The function $f$ is said $\tau$-quasar convex with respect to $x^*$ if it satisfies*

$$\forall x \in \mathbb{R}^d,\ f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle, \tag{3.13}$$

*and $(\tau, \mu)$-strongly quasar-convex with respect to $x^*$ if:*

$$\forall x \in \mathbb{R}^d,\ f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2. \tag{3.14}$$

We refer to any minimizer $x^*$ at which (3.13) holds, as a quasar convex point of $f$. The class of (strongly) quasar convex functions was first introduced in 2017 by (Hardt et al., 2018) (with $\tau > 0$) where the authors refer to it as *weak quasi-convexity*. It was then implicitly re-used by (Su et al., 2016) and (Aujol et al., 2019; Apidopoulos et al., 2021) as a *flatness condition* (with $\tau \geq 1$), and revisited more recently in (Hinder et al., 2020) where the quasar convexity name was introduced. A nice property of this class of functions is that any critical point of (strongly) quasar convex function $f$ is a global minimizer of $f$.

Moreover, the set of minimizers $X^*$ of a quasar convex function has some strong regularity: it is a star convex set i.e. there exists $x^* \in X^*$ such that:

$$\forall x \in X^*,\ \forall t \in [0, 1],\ tx^* + (1 - t)x \in X^*, \tag{3.15}$$

and is reduced to a singleton for strongly quasar-convex functions, see (Hinder et al., 2020, Appendix D, Observations 3 and 4).

Lastly, observe that the Polyak-Łojasiewicz and the quadratic growth properties can be seen as relaxations of strong quasar convexity, as stated by the following result:

**Proposition 3.2.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $(\tau, \mu)$-strongly quasar-convex function for some $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+$ and $x^*$ its minimizer. Let $f^* = \min f$. Then:*

1. *$f$ is $\mu\tau^2$-PL, i.e.*

$$\forall x \in \mathbb{R}^d, \ \frac{1}{2\tau^2\mu}\|\nabla f(x)\|^2 \geq f(x) - f^*. \tag{3.16}$$

2. *(Hinder et al., 2020, Corollary 1) $f$ has a $\frac{\tau\mu}{2-\tau}$-quadratic growth, i.e*

$$\forall x \in \mathbb{R}^d, \ f(x) - f^* \geq \frac{\tau\mu}{2(2-\tau)}\|x - x^*\|^2. \tag{3.17}$$

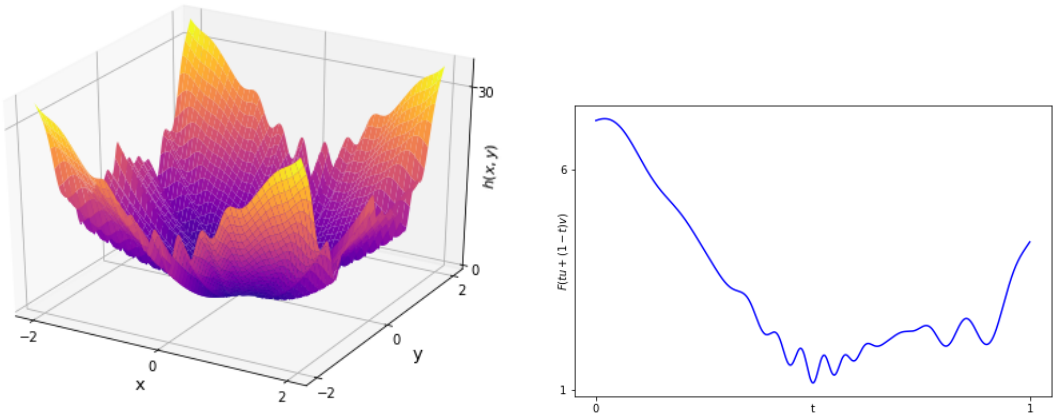



Figure 3.1: An example of strongly quasar-convex function built as (3.18), whose explicit expression is given in section 3.6. On the left, the graph of this function. On the right, a cut of this graph along a segment of $\mathbb{R}^2$, such that the minimizer does not belong to this segment.

Strongly quasar convex functions can be highly oscillating. Let us take the construction described in (Lee and Valiant, 2016; Hinder et al., 2020) to highlight the non-convexity of strongly quasar-convex functions. Let $f : \mathbb{R} \to \mathbb{R}$ such that it is $(\tau, \mu)$-strongly quasar-convex, with $f^* = f(0) = 0$. Let $g : S^{d-1} \to \mathbb{R}$ be an arbitrary continuous function defined on the unit circle of $\mathbb{R}^d$ such that $g \geq 1$. Consider

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right), \ x \in \mathbb{R}^d. \tag{3.18}$$

This function is $(\tau, \mu)$-strongly quasar-convex independently of the choice of $g$ (see (Hinder et al., 2020, Appendix D.3) for the non strongly quasar-convex case, and see Appendix 3.A.2 for the strongly quasar-convex case). An example of such a function is displayed on the left side of figure 3.1. Radially this function behaves like $cf(\|x\|)$ where $c$ is constant. Restricted to this direction, the function is unimodal and critical points are minimizers (Hinder et al., 2020, Observation 1). However since $g$ may be extremely non-convex, we see that taking the segment between two arbitrary points $x_0$ and $x_1$, smoothness aside we will have no control over the behaviour of the function (*e.g.* right side of figure 3.1).
This lack of local regularity may not be a big deal with gradient descent, as it follows a descent direction at each iteration. However, in the case of NM, the presence of inertia prevents us from controlling the direction of the trajectory. We will see later that this lack of local regularity complicates considerably the potential acceleration of NM.

## 3.3 Acceleration with curvature assumption for strongly quasar-convex functions

### 3.3.1 The gradient descent convergence

As we are interested in faster algorithms than gradient descent, we first establish the convergence results associated with this algorithm. We recall it is defined for some $x_0 \in \mathbb{R}^d$, $s \geq 0$ by this recursive formula:

$$x_{n+1} = x_n - s\nabla f(x_n). \tag{GD}$$

Let us first recall a known result for smooth strongly convex functions.

**Proposition 3.3.1** ((Nesterov, 2013))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $L$-smooth and $\mu$-strongly convex function for some $0 < \mu \leq L$, and $x^*$ its minimizer. Let $f^* = \min f$. Let $(x_n)_{n\in\mathbb{N}}$ be generated by (GD) with stepsize $s = \frac{1}{L}$. Then:*

$$\forall n \in \mathbb{N},\ \|x_n - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^n \|x_0 - x^*\|^2. \tag{3.19}$$

Proposition 3.3.1 indicates that for $\mu$-strongly convex and $L$-smooth functions, (GD) achieves a linear decrease. Proposition 3.3.1 can be extended to strongly quasar-convex functions. To the best of our knowledge, the following proposition is not clearly stated in literature. The case $\tau = 1$ is proved in (Guille-Escuret et al., 2021). A result can be found in the stochastic case in (Gower et al., 2021), from which a result for the deterministic case can be deduced. Their stepsize is however lower than ours, so it yields a slower rate of convergence.

**Proposition 3.3.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $L$-smooth and $(\tau, \mu)$-strongly quasar-convex function for some $0 < \mu \leq L$, $\tau \in (0, 1]$, and let $x^*$ be its minimizer. Let $(x_n)_{n\in\mathbb{N}}$ be generated by (GD) with stepsize $s \leq \frac{1}{L}$. Then:*

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq \frac{2}{\tau}(1 - \tau\mu s)^n (f(x_0) - f^*). \tag{3.20}$$

*Proof.* Let $x^*$ be the quasar convex point of $f$ and:

$$E_n = f(x_n) - f^* + \frac{\mu}{2}\|x_n - x^*\|^2. \tag{3.21}$$

We compute

$$E_{n+1} - E_n = f(x_{n+1}) - f(x_n) + \frac{\mu}{2}\|x_{n+1} - x^*\|^2 - \frac{\mu}{2}\|x_n - x^*\|^2 \tag{3.22}$$

$$\overset{(GD)}{=} f(x_{n+1}) - f(x_n) - \mu s\langle x_n - x^*, \nabla f(x_n)\rangle + s^2\frac{\mu}{2}\|\nabla f(x_n)\|^2. \tag{3.23}$$

The $L$-smooth inequality (3.2) implies that: $f(x_{n+1}) - f(x_n) \leq -\frac{s}{2}\|\nabla f(x_n)\|^2$ provided that $s \leq \frac{1}{L}$. Combined with the $(\tau, \mu)$-strongly quasar-convexity to control the scalar product, we get:

$$E_{n+1} - E_n \leq -\frac{s}{2}\|\nabla f(x_n)\|^2 - \tau\mu s(f(x_n) - f^*) - \tau s\frac{\mu^2}{2}\|x_n - x^*\|^2 + s^2\frac{\mu}{2}\|\nabla f(x_n)\|^2 \tag{3.24}$$

$$= \frac{s}{2}\left(\mu s - 1\right)\|\nabla f(x_n)\|^2 - \tau\mu s\left(f(x_n) - f^* + \frac{\mu}{2}\|x_n - x^*\|^2\right) \tag{3.25}$$

as $s \leq \frac{1}{L}$ the first term is negative, inducing

$$E_{n+1} - E_n \leq -\tau\mu s E_n \Rightarrow E_{n+1} \leq (1 - \tau\mu s)E_n. \tag{3.26}$$

By induction, we then deduce:

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq (1 - \tau\mu s)^n (f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2). \tag{3.27}$$

Using the $\frac{\tau\mu}{2-\tau}$-quadratic growth induced by the $(\tau, \mu)$-quasar strong convexity of $f$ (see Corollary 1 (Hinder et al., 2020)), we finally get the expected convergence rate. □

### 3.3.2 The Nesterov Momentum convergence

The Nesterov Momentum algorithm used to optimize $L$-smooth and $\mu$-strongly convex function (NM-SC) is often written in the following way (Nesterov, 2013, Algorithm (2.2.22)):

$$\begin{cases} y_n = x_n + \frac{1-\sqrt{\frac{\mu}{L}}}{1+\sqrt{\frac{\mu}{L}}}(x_n - x_{n-1}) \\ x_{n+1} = y_n - \frac{1}{L}\nabla f(y_n) \end{cases} \tag{NM-SC 2 POINTS}$$

Introducing the auxiliary variable:

$$z_n = \left(1 + \sqrt{\frac{L}{\mu}}\right) y_n - \sqrt{\frac{L}{\mu}} x_n,$$

this algorithm can be rewritten as a 3-points scheme, see (Nesterov, 2013, Algorithm (2.2.19) with $\tau_0 = \mu$):

$$\begin{cases} y_n = \left(\frac{1}{1+\sqrt{\frac{\mu}{L}}}\right) x_n + (1 - \frac{1}{1+\sqrt{\frac{\mu}{L}}}) z_n \\ x_{n+1} = y_n - \frac{1}{L}\nabla f(y_n) \\ z_{n+1} = (1 - \sqrt{\frac{\mu}{L}}) z_n + \sqrt{\frac{\mu}{L}}(y_n - \frac{1}{\mu}\nabla f(y_n)) \end{cases} \tag{NM-SC 3 POINTS}$$

We have the following well known result.

**Theorem 3.3.1** ((Nesterov, 2013))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $L$-smooth and $\mu$-strongly convex function for some $0 < \mu \leq L$, and $x^*$ its unique minimizer. Let $f^* = \min f$. Let $(x_n)_{n\in\mathbb{N}}$ be the sequence generated by (NM-SC 3 POINTS) with $x_0 = z_0$. Then:*

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^n \left(f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right). \tag{3.28}$$

Observe that this is a considerable improvement over the $\mathcal{O}\left((1 - \frac{\mu}{L})^n\right)$ rate given by gradient descent (stepsize $\frac{1}{L}$), as $\frac{\mu}{L} \leq 1$ may be extremely low for high-dimensional functions.

Let us introduce a generalization of the classical (NM-SC 3 POINTS) algorithm, see Algorithm 13 (see *e.g.* (Hinder et al., 2020)).

**Algorithm 13** Nesterov Momentum (three-point form)

Let $z_0 = x_0$
**for** n = 0,. . . , **do**
  $y_n = \alpha_n x_n + (1 - \alpha_n) z_n$
  $x_{n+1} = y_n - s\nabla f(y_n)$
  $z_{n+1} = \beta_n z_n + (1 - \beta_n) y_n - \eta_n \nabla f(y_n)$
**end for**

We want to use Algorithm 13 to achieve an accelerated convergence similar to Theorem 3.3.1 with strongly quasar-convex functions. We do so using the notion of *curvature function* (3.9):

**Theorem 3.3.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $(\tau, \mu)$-strongly quasar-convex function for some $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$ and $f^* = \min f$. Assume additionally that $f$ is a $(\rho, L)$-curvature-bounded function for some $L > 0$ and $\rho \leq L$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of iterates generated by Algorithm 1 with parameters:*

$$s \leq \tfrac{1}{L},\ \alpha_n = \tfrac{1}{1+\sqrt{\mu s}},\ \beta_n = 1 - \tau\sqrt{\mu s},\ \eta_n = \tfrac{\sqrt{s}}{\sqrt{\mu}}.$$

*If $\rho \geq -\tau\sqrt{\frac{\mu}{s}}$, then:*

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq \frac{2}{\tau}\left(1 - \tau\sqrt{\mu s}\right)^n \left(f(x_0) - f^*\right). \tag{3.29}$$

See the proof in Appendix 3.B.1, where we prove linear decrease of the following Lyapunov function

$$E_n = f(x_n) - f^* + \frac{\mu}{2}\|z_n - x^*\|^2. \tag{3.30}$$

More precisely, we show $E_{n+1} \leq (1 - \tau\sqrt{\mu s})E_n$, $\forall n \in \mathbb{N}$. The $(\rho, L)$-curvature assumption with $\rho < 0$ allows us to replace convexity by a weaker bound control, namely:

$$\forall n \in \mathbb{N},\ \langle \nabla f(y_n), x_n - y_n\rangle + f(y_n) - f(x_n) \leq -\frac{\rho}{2}\|x_n - y_n\|^2.$$

Fixing $s = \frac{1}{L}$, the curvature bound becomes $\rho \geq -\tau\sqrt{\mu L} = -\tau\sqrt{\frac{\mu}{L}}L$. Parameterizing the algorithm with an arbitrary step size $s \leq \frac{1}{L}$ allows us to highlight that we can trade restriction on the curvature with convergence speed. Formally:

**Corollary 3.3.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $L$-smooth and $(\tau, \mu)$-strongly quasar-convex function for some $0 < \mu \leq L$, $\tau \in (0, 1]$, and let $f^* = \min f$. Let $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 with parameters $s = \tau^2\frac{\mu}{L^2}$, $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \tau\sqrt{\mu s}$ and $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$. Then:*

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq \frac{2}{\tau}\left(1 - \tau^2\frac{\mu}{L}\right)^n (f(x_0) - f^*). \tag{3.31}$$

*Proof.* Just solve $-\tau\sqrt{\frac{\mu}{s}} = -L$, we get $s = \tau^2\frac{\mu}{L^2}$. □

We see that we can delete the curvature assumption, but at the cost of acceleration. The negative curvature problem has already been observed under other non-convex hypothesis. The authors of (Carmon et al., 2017; Jin et al., 2018) prove accelerated convergence to a critical point in a Hessian Lipschitz setting, where they use NM when curvature between iterates is not too negative, otherwise an alternative step using Hessian Lipschitz property is performed. In our case, alternative step using Hessian Lipschitz property is hardly useful as our Lyapunov function, designed to obtain linear convergence, is less adapted to this argument.

**Comparison with the algorithm of (Hinder et al., 2020, Algorithm 3) (binary search)** The convergence bound of Theorem 3.3.2 is similar to the one of (Hinder et al., 2020). They achieve this bound using Algorithm 13 where, at each iteration, a line search procedure to compute one of the parameters (see Algorithm 3 in (Hinder et al., 2020)). This results in a $\log\left(\tau^{-1}\frac{L}{\mu}\right)$ factor added to the number of gradient and function evaluations needed, that does not appear in our convergence rate. Getting rid of this line search procedure allows to get deeper insights about NM using ODEs modeling, as we see in Section 3.4.

**Two-point form of Algorithm 1** Following arguments of (Lee et al., 2021), we can rewrite our algorithm in a two-point version, as stated with this result:

**Proposition 3.3.3.** *The algorithm 1 with parameters $s \leq \frac{1}{L}$, $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \tau\sqrt{\mu s}$ and $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$ can be written as the following two-point algorithm:*

$$\begin{cases} y_n = x_n + \frac{1-\tau\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\tau - 1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla f(y_n) \end{cases} \tag{NM-SQC 2 POINTS}$$

When $\tau = 1$, we recover the classical Nesterov algorithm for minimizing strongly convex functions. Note that two-point scheme versions are widely present in the literature. Considering the two-point or the three-point version gives different interpretations of the algorithm, and switching from a version to another is not obvious. The proof of Proposition 3.3.3 is in Appendix 3.B.1.

### 3.3.3 Composite non differentiable case

A natural way to extend the differential case to the non differentiable one is to consider the class of composite functions $F = f+g$, where $f$ is assumed to be differentiable and strongly quasar-convex, and $g$ convex, proper lower semi-continuous, and such that its proximal operator (Combettes and Pesquet, 2011) can be computed:

$$\text{prox}_g(x) = \arg\min_y \left( g(y) + \frac{1}{2}\|x-y\|^2 \right). \tag{3.32}$$

A typical example of such functions is the composite functions with a $\ell_1$ penalization term: $F = f + \|x\|_1$. The idea is then to replace the gradient in Algorithm 1 by the composite gradient mapping:

$$\nabla f(y_n) \to \frac{1}{s}\left(y_n - \text{prox}_{sg}(y_n - s\nabla f(y_n))\right) \tag{3.33}$$

which is a generalization of the gradient (we recover it when $g \equiv 0$). It leads us to Algorithm 2.

**Algorithm 14** Proximal Nesterov Momentum (three-point form)

Let $z_0 = x_0$
**for** k = 0,. . . , **do**
  $y_n = \alpha_n x_n + (1-\alpha_n) z_n$
  $x_{n+1} = \text{prox}_{sg}(y_n - s\nabla f(y_n)) := T_s(y_n)$
  $z_{n+1} = \beta_n z_n + (1-\beta_n) y_n - \frac{\eta_n}{s}(y_n - T_s(y_n))$
**end for**

Extending Theorem 3.3.2 to the non-differentiable case introduces an additional technicality since a minimizer/quasar point of $f$ is generally not a minimizer of the composite function $F$. In other words, if $f$ is strongly quasar-convex with respect to its minimizer $x_f^*$, there is clearly no guarantee that this assumption holds at a minimizer $x_F^*$ of $F$. Therefore, we propose an extension of strong quasar convexity with respect to another point than a minimizer as suggested in (Hinder et al., 2020, Appendix D.2):

**Definition 3.3.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\hat{x} \in \mathbb{R}^d$. The function $f$ is said $(1,\mu)$-strongly quasar-convex with respect to $\hat{x}$ if:*

$$\forall x \in \mathbb{R}^d,\ \forall t \in [0,1],\ f(t\hat{x} + (1-t)x) + t(1-t)\frac{\mu}{2}\|\hat{x}-x\|^2 \le tf(\hat{x}) + (1-t)f(x),$$

*or equivalently, when $f$ is additionally assumed to be differentiable:*

$$\forall x \in \mathbb{R}^d,\ f(\hat{x}) \ge f(x) + \langle \nabla f(x), \hat{x} - x\rangle + \frac{\mu}{2}\|\hat{x}-x\|^2. \tag{3.34}$$

The equivalence is shown in (Hinder et al., 2020, Lemma 11), whose proof works for $\hat{x}$ not being a minimizer, if we consider $(\tau,\mu)$ strong convexity with $\tau = 1$.

**The $\tau < 1$ case** Taking $\tau < 1$ does not cope well with quasar convexity with respect to an arbitrary point. Assume that for some $f \in C^1$, $\mu > 0$ and $\tau \in (0,1]$ the following holds:

$$\forall x \in \mathbb{R}^d,\ f(\hat{x}) \ge f(x) + \frac{1}{\tau}\langle \nabla f(x), \hat{x} - x\rangle + \frac{\mu}{2}\|\hat{x}-x\|^2. \tag{3.35}$$

with $\hat{x}$ such that $\nabla f(\hat{x}) \neq 0$. Set $x_h = \hat{x} - h\nabla f(\hat{x})$, $h > 0$. Using a first-order Taylor expansion:

$$f(\hat{x}) = f(x_h) + \langle \nabla f(x_h), \hat{x} - x_h\rangle + o(\|\hat{x}-x_h\|) = f(x_h) + h\langle \nabla f(x_h), \nabla f(\hat{x})\rangle + o(h) \tag{3.36}$$

Combining this with (3.35) evaluated in $x = x_h$, we have:

$$\langle \nabla f(x_h), \nabla f(\hat{x})\rangle \left(1 - \frac{1}{\tau}\right) + \frac{o(h)}{h} \ge \frac{h\mu}{2}\|\nabla f(\hat{x})\|^2 \tag{3.37}$$

As $h$ goes to 0, the right hand side goes to 0 while the left hand side goes to $\|\nabla f(\hat{x})\|\left(1-\frac{1}{\tau}\right)$ because $f$ is $C^1$. As the latter is strictly negative for $\tau \in (0,1)$, then in our definition $\tau$ must be equal to 1.
Observe now that Definition 3.3.1 is not an empty definition, and that such functions can easily be defined. A possible construction consists in adding some convex function to a $(1,\mu)$-strongly quasar-convex function.

**Lemma 3.3.1.** *Let $f:\mathbb{R}^d \to \mathbb{R}$ be $(1,\mu)$-strongly quasar-convex with respect to $\hat{x}$ and $g:\mathbb{R}^d \to \mathbb{R}$ be convex, neither necessarily differentiable. Then $F := f+g$ is $(1,\mu)$-strongly quasar-convex with respect to $\hat{x}$.*

*Proof.* The proof is straightforward by summing the respective inequalities of $(1,\mu)$-strong quasar convexity of $f$ and convexity of $g$ written between any $x \in \mathbb{R}^d$ and $\hat{x}$. □

As a special case of Lemma 3.3.1, observe that if $f$ is $(1,\mu)$-strongly quasar-convex with respect to its minimizer $x_f^*$, then $x_f^*$ is not necessarily a minimizer of $F$, but it is a quasar convex point of $F$.

**Convergence result for Algorithm 14** Using the extended definition of $(1,\mu)$-strongly quasar-convexity, we are ready to state our result.

**Theorem 3.3.3.** *Let $F = f+g$ where $f:\mathbb{R}^d \to \mathbb{R}$ is an $L$-smooth function for some $L>0$ and $g:\mathbb{R}^d \to \mathbb{R}$ is convex, proper, lower semi-continuous. Assume that $F$ has a nonempty set of minimizers and that $f$ is $(1,\mu)$-strongly quasar-convex with respect to $x_F^* \in \arg\min F$ for some $0<\mu\leq L$ and $(\rho,L)$-curvature-bounded for some $\rho \leq L$. Let $(x_n)_{n\in\mathbb{N}}$ be the sequence of iterates generated by Algorithm 14 with parameters:*

$$s \leq \tfrac{1}{L},\ \alpha_n = \tfrac{1}{1+\sqrt{\mu s}},\ \beta_n = 1-\sqrt{\mu s},\ \eta_n = \tfrac{\sqrt{s}}{\sqrt{\mu}}.$$

*If $\rho \geq -\sqrt{\frac{\mu}{s}}$, then:*

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leq 2\left(1-\sqrt{\mu s}\right)^n \left(F(x_0)-F^*\right). \tag{3.38}$$

Proof is in Appendix 3.B.1. It uses the same Lyapunov function as in Theorem 3.3.2 and it makes use of the Prox-Grad theorem from (Beck, 2017).
It might seem odd, for practical purposes, to ask strong quasar convexity at a point which is not a minimizer of the aforementioned function. It allows to show that theoretically, convexity of $f$ is not necessary to extend the accelerated result in the differentiable case to the composite case. Note that together with difficulties of defining (strongly) quasar convex functions with respect to an arbitrary point, arguments used in our proof do not work if $\tau < 1$. In that case, the $\tau < 1$ relaxation is non trivial, while in some cases it merely requires adjusting the calculations.

## 3.4 Continuous analysis

Considering first-order algorithms as a discretization of an ODE is a powerful tool, helping to gain intuition, finding and generalizing new convergence results (Su et al., 2016; Siegel, 2019). The classic ODE version of NM (Algorithm 13) is written as the following:

$$\ddot{X}(t) + \tau(t)\dot{X}(t) + \nabla f(X(t)) = 0,$$

where $\tau(t)$ is a non negative function. The solution of this ODE can be thought as the continuous analogue of the NM algorithm. Using Lyapunov strategies, one can deduce convergence results on the solution of this system. This has been extensively done in the case of convex or strongly convex functions, *e.g.* see (Su et al., 2016; Siegel, 2019; Aujol et al., 2023a, 2024c; Attouch et al., 2022; Shi et al., 2022; Li et al., 2024).

### 3.4.1 High-resolution ordinary differential equation

High-resolution ODEs are introduced in (Shi et al., 2022) to study first-order optimization algorithms. The interest of these ODEs is to describe more accurately the behaviour of the corresponding algorithms compared to the classical versions of these ODEs, thus renamed low-resolution ODEs. This is made by keeping $\mathcal{O}(\sqrt{s})$ terms during the discretization process, inducing that the ODE depends on the stepsize $s$ of NM (Algorithm 13).

**$f$ $\mu$-strongly convex** Considering a $\mu$-strongly convex function $f$, Polyak's Heavy Ball (Polyak, 1964) and NM-SC (see Section 3.3.2) can be seen as discretizations of the same low resolution ODE:

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \nabla f(X(t)) = 0.$$

However, these algorithms correspond to two different high-resolution ODEs (see (Shi et al., 2022) for more details):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1+\sqrt{\mu s})\nabla f(X(t)) = 0 \tag{HB-ODE}$$

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1+\sqrt{\mu s})\nabla f(X(t)) = 0 \tag{NM-SC-ODE}$$

As Polyak's Heavy ball cannot achieve an accelerated rate for smooth strongly convex functions, while NM-SC can, understanding how the two high-resolution ODEs differ is thus interesting. The only differing term is $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$, which results in the corresponding algorithm to what the authors refer to as a gradient correction term. In (NM-SC-ODE), factorizing $\dot{X}$ terms, we see that the damping coefficient becomes $2\sqrt{\mu}+\sqrt{s}\nabla^2 f(X(t))$, which is now adaptive to the position of $X$. In particular if $\dot{X}$ is collinear with the eigenvector corresponding to the highest eigenvalue of $\nabla^2 f$ (possibly $L$), the new damping rate increases, thus reducing oscillations.
In the $\mu$-strongly convex case, one can obtain the following convergence result for the solution of (NM-SC-ODE).

**Theorem 3.4.1.** *[Theorem 1, (Shi et al., 2022)] Let $f$ be $C^2$, $\mu$-strongly convex and $L$-smooth for some $0 < \mu \leq L$. Assume $X$ is solution of (NM-SC-ODE) with $0 < s \leq \frac{1}{L}$, $X(0) = X_0$ and $\dot{X}(0) = \frac{-2\sqrt{s}\nabla f(X_0)}{1+\sqrt{\mu s}}$. Then:*

$$f(X(t)) - f^* \leq \frac{2\|X_0 - x^*\|^2}{s} e^{-\frac{\sqrt{\mu}}{4}t}. \tag{3.39}$$

Theorem 3.4.1 states that $\mu$-strong convexity allows for an exponential convergence with rate $\mathcal{O}\left(e^{-\sqrt{\mu}t/4}\right)$. We extend this result when relaxing strong convexity with strong quasar convexity.

**$f$ $(\tau, \mu)$-strongly quasar-convex** We first derive the high-resolution ODE associated to Algorithm 13. Recall that with the choice of parameters of Theorem 3.3.2, Algorithm 1 is

$$\begin{cases} y_n = \frac{1}{1+\sqrt{\mu s}} x_n + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}} z_n \\ x_{n+1} = y_n - s\nabla f(y_n) \\ z_{n+1} = 1 - \tau\sqrt{\mu s} z_n + \tau\sqrt{\mu s} y_n - \sqrt{\frac{s}{\mu}}\nabla f(y_n) \end{cases} \tag{NM-SQC-DISCRETE}$$

The high-resolution ODE associated with (NM-SQC-DISCRETE) can be written as follows:

$$\left(1 + \frac{1-\tau}{2}\sqrt{\mu s}\right)\ddot{X}(t) + (1+\tau)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1+\tau\sqrt{\mu s})\nabla f(X(t)) = 0 \tag{NM-SQC-ODE}$$

The derivation of this ODE is detailed in Appendix 3.C. Observe that taking $\tau = 1$, we recover (NM-SC-ODE).

**Remark** The Hessian term will induce specific behaviour in a non-convex case. If $f$ is $L$-smooth, and $\dot{X}$ is collinear to the eigenvector of $\nabla^2 f(X(t))$ associated with eigenvalue $-L$, the damping rate becomes

$$((1+\tau)\sqrt{\mu} - L\sqrt{s})\dot{X}(t).$$

Take $s = \frac{1}{L}$ and it is negative (if $\mu$ is not too close to $L$). The case of negative damping rate is unusual. In particular, it induces the increase of the total mechanical energy $(f(X(t)) - f^* + \frac{1}{2}\|\dot{X}(t)\|^2)$.

Despite the eventual negative friction of (NM-SQC-ODE), one can get convergence results that are similar to the one obtained in the strongly convex setting (Siegel, 2019), up to a $\tau$ factor.

**Theorem 3.4.2.** *Let $f$ be $C^2$, $(\tau,\mu)-$strongly quasar-convex and $L$-smooth for some $0<\mu\leq L$, $\tau\in(0,1]$. Assume $X$ is solution of (NM-SQC-ODE) with $0\leq s\leq \frac{1}{L}$, $X(0)=X_0$ and $\dot{X}(0)=0$. Then:*

$$f(X(t))-f^*\leq K_0(\tau,\mu,L,s)\frac{1}{\tau}(f(X_0)-f^*)e^{-\tau\frac{\sqrt{\mu}}{2}t} \tag{3.40}$$

*where $K_0(\tau,\mu,L,s)$ can be uniformly bounded by 7.*

To prove Theorem 3.4.2, we aim to show the linear decrease of a Lyapunov function of the form:

$$\mathcal{E}(t)=\delta(f(X(t))-f^*)+\frac{1}{2}\left\|\left(1+\frac{1-\tau}{2}\sqrt{\mu s}\right)\dot{X}(t)+\lambda(X(t)-x^*)+\sqrt{s}\nabla f(X(t))\right\|^2. \tag{3.41}$$

where $\delta$, $\lambda$ belong in $\mathbb{R}$ and are well chosen parameters. See Appendix 3.C for the proof.
The result of Theorem 3.4.2 is quite surprising. It indicates that despite the occurrence of negative damping along the trajectory, we can still achieve similar convergence bounds as for the strongly convex case. In particular, we do not need assumptions such as $(a,L)$-curvature (Definition 3.2.1). Thus, while non-convexity of strongly quasar-convex functions may impact negatively the convergence of Algorithm 13, it is not the case for the associated continuous system. This indicates that non-convexity impacts the discretization process. We discuss this intuition in the next section.

**Remark:** Taking $s=0$ in (NM-SQC-ODE), we can automatically deduce the low resolution ODE associated with (NM-SQC-DISCRETE).

### 3.4.2 The continuous/discrete rupture

As shown in the previous section, one can achieve accelerated convergence with the solution of the high-resolution ODE (NM-SQC-ODE) associated to the discrete version, without the need to of curvature assumption.
As the algorithm and the discrete Lyapunov function we use are discretization of continuous counterparts, one expect these to be similar. More precisely, ordering term by their dependence on $s$ ($1$, $\mathcal{O}(\sqrt{s})$, $\mathcal{O}(s),\dots$), one expects the *main terms*, namely the one with lower dependence on $s$, to appear both in continuous and discrete setting, while eventually, the discretization process will make appear new terms with higher dependency on $s$. In this section, we see why it is not necessarily the case when non-convexity steps in.

**Continuous/discrete Lyapunov comparison** Recall the continuous Lyapunov used to prove Theorem 3.4.2.

$$\mathcal{E}(t)=\delta(f(X(t))-f^*)+\frac{1}{2}\|\upsilon\dot{X}(t)+\lambda(X(t)-x^*)+\sqrt{s}\nabla f(X(t))\|^2 \tag{3.42}$$

where, depending on the values of $\tau,\mu,L$, we have $\upsilon\in[1,\frac{3}{2}]$, $\delta\in[1,3]$ and $\lambda\in[\sqrt{\mu},\frac{9}{8}\sqrt{\mu}]$ (see the exact values in Appendix 3.C). The discrete Lyapunov used to prove Theorem 3.3.2 can be written the following way:

$$E_n=f(x_n)-f^*+\frac{1}{2}\|\frac{(y_n-y_{n-1})}{\sqrt{s}}+\sqrt{\mu}(y_n-x^*)+\sqrt{s}\nabla f(y_{n-1})\|^2. \tag{3.43}$$

The fact that $\{E_n\}_{n\in\mathbb{N}}$ is a discretization of $\mathcal{E}$ appears here clearly.

**Difference between derivatives and finite differences** In the continuous case, we studied $\dot{\mathcal{E}}$, where we aimed to show $\dot{\mathcal{E}}(t)\leq-\tau\frac{\sqrt{\mu}}{2}\mathcal{E}(t)$. We emphasize the two following facts:

- Differentiation of $\delta(f(X(t))-f^*)$ leads to $\delta\langle\nabla f(X),\dot{X}\rangle$.
- Differentiation of the norm term leads, among other terms, to $-\langle\nabla f(X),\dot{X}\rangle$ (up to a parameter).

Appropriate parameter tuning allows to cancel these terms, see proof of Theorem 3.4.2.
In the discrete case, we study a finite difference $E_{n+1}-E_n$. Importantly, the corresponding terms, which agree in the continuous setting, differ here.

- Finite difference of $f(x_n) - f^*$ leads to $f(x_{n+1}) - f(x_n)$.
- Finite difference of the norm term brings, among other terms, to $-\langle y_n - x_n, \nabla f(y_n)\rangle$.

More precisely, with slight modifications of the proof of Theorem 3.3.2, we get:

$$E_{n+1} \leq \underbrace{(1-\tau\sqrt{\mu s})E_n}_{\text{Aimed linear decrease}} + (1-\tau\sqrt{\mu s})\underbrace{(f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle)}_{\text{Finite-difference residual}}$$
$$\underbrace{-\tau(1-\tau\sqrt{\mu s})\sqrt{\frac{\mu}{s}}\|y_n - x_n\|^2}_{\leq 0}. \tag{3.44}$$

In (3.44), apart from the negative kinetic term, that can be matched with a continuous counterpart ($-\tau\frac{\sqrt{\mu}v^2}{2}\|\dot{X}(t)\|^2$ term in (3.177)), it remains the difference of the two aforementioned quantities, *i.e.*

$$f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle. \tag{3.45}$$

In contrast to the continuous case, we cannot just use parameter tuning to control this term. Convexity is exactly the assumption we need to get rid of this finite-difference residual term. However, if $f$ is $(a, L)$-curvature-bounded with $a < 0$, the best control we have is the following:

$$f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle \leq -\frac{a}{2}\|x_n - y_n\|^2. \tag{3.46}$$

The smaller $a < 0$ is, and thus the more non-convexity we allow, the more this extra term can get large, up to a $+\frac{L}{2}\|x_n - y_n\|^2$ term in the $L$-smooth case ($a = -L$). If this term was appearing in the continuous case, the same restriction on the curvature as for Theorem 3.3.2 would be necessary. This highlights that when considering properties satisfied by the solution of a continuous system, the transfer of these properties to the discrete case via discretization can be hurt by non-convexity.

## 3.5 Geometrical considerations

In this section, we present some new properties of strongly quasar-convex functions, together with properties concerning more general non-convex functions.

**How weaker than strong convexity is strong quasar convexity ?** Compared with strong convexity, the main difference is that we lose lots of *local information*. While for $C^2$ $\mu$-strongly convex functions we have $\langle \nabla^2 f(x)y, y\rangle \geq \mu\|y\|^2$ for all $x, y \in \mathbb{R}^d$ (or equivalently, all eigenvalues of the Hessian matrix are above $\mu$), we lose this regularity with strongly quasar-convex functions. Actually, we only have, on average, a similar regularity on the segment joining points $x \in \mathbb{R}^d$ and the minimizer $x^*$.

**Proposition 3.5.1.** *Let $f$ be $C^2$ and $(\tau, \mu)$-strongly quasar-convex for some $\tau \in (0, 1]$, $\mu > 0$. Let $x \neq x^*$ and $t > 0$. Then:*

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 f(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geq \tau\frac{\mu}{2}. \tag{3.47}$$

See appendix 3.A.1 for the proof.

**Local strong convexity for $C^2$ functions** One may think that the property of uniqueness of the minimizer of strongly quasar-convex functions induces that when we are close enough to the minimizer, the function is bowl-shaped and we avoid negative curvature, *i.e.* the function is locally convex around the minimizer. This is true if the function is $C^2$. The following two results hold under the assumption of a quadratic growth function with a unique minimizer, which is a weaker statement than strong quasar convexity (see Section 3.2.2).

**Proposition 3.5.2.** *Let $f$ be $C^2$, with a unique minimizer $x^*$, and a $\mu$-quadratic growth. Then, there exists $\eta > 0$ such that $f$ is strongly convex on $B(x^*, \eta)$.*

*Sketch of Proof.* The quadratic growth property around $x^*$ implies that $\nabla^2 f(x^*)$ is definite positive. By continuity of the Hessian we get the result. Rigorous proof is deferred in appendix 3.A.3. □

**Pathological behaviour for non $C^2$ functions** When dropping the $C^2$ assumption, one cannot ensure the local convexity property anymore, even when considering segments joining the minimizer (on which we have the stronger structure assumption). This is stated in the following result.

**Proposition 3.5.3.** *One can construct $f : [0,1] \to \mathbb{R}$ strongly quasar-convex with minimizer $x^*$, $L$-smooth, such that for all $x \neq x^*$, there exists $x_0 \neq x^*$ such that:*

$$|x^* - x_0| \leq |x^* - x| \text{ and } f''(x_0) = -L \tag{3.48}$$

*Sketch of Proof.* The idea behind the construction is the following: the function $f$ is built such that its curvature alternates $L$ and $-L$, given that the $L$ curvature will occur more often than $-L$ to ensure that the function grows enough to be strongly quasar-convex.
To create such a function, we define the following set

$$E := \bigcup_{n\geq 1} \left( \underbrace{\left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right)}_{\text{partition of } [0,1]} \cap \underbrace{\left[\frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right]}_{\text{subpart of partition}} \right) = \bigcup_{n\geq 1} \left[\frac{7}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right). \tag{3.49}$$

Then we define $f$ on $[0,1]$, such that:

$$f''(x) = \begin{cases} -L & \text{if } x \in E \\ L & \text{else.} \end{cases}$$

with $f'(0) = f(0) = 0$. $f$ is clearly such that it will reach its lower curvature $-L$ in every neighborhood of its minimizer. It remains to show that it is strongly quasar-convex, which is done in Appendix 3.A.4.

□

**Other pathological examples** The construction used to show Proposition 3.5.3 can be adapted to make other pathological behaviours. For example, consider $\mu$-quadratic growth functions, *i.e.* which satisfy for all $x \in \mathbb{R}^d$, $f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2$. This defines functions lower bounded by a quadratic, but it may have critical points that are not (global) minima. However due to the local regularity around a global minimizer $x^*$ offered by the quadratic growth, one may ask the following question:

**Can we ensure that a smooth function with unique minimizer $x^*$ satisfying $\mu$-quadratic growth has no other critical points when sufficiently close to $x^*$ ?**

We address this question in appendix 3.A.4, answering negatively by constructing a counter example based on the aforementioned construction.

## 3.6 Numerical experiments details

### 3.6.1 Algorithm performance

We tested our algorithm on a function $h : \mathbb{R}^d \to \mathbb{R}$, $d = 100$, where

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{3.50}$$

with $f(t) = t^2$, and

$$g(x_1, \ldots, x_{100}) = \sum_{i=1}^{100} \left(a_i \sin(b_i x_i)^2\right) + 1 \tag{3.51}$$

where $\{a_i\}_i$ are independently and uniformly distributed on $[0,1]$ and the $\{b_i\}_i$ are independently and uniformly distributed on $[-2.5, 2.5]$. This function is $(1,2)$-strongly quasar-convex, as $f$ is $(1,2)$-strongly quasar-convex (see Proposition 3.A.1). Recall that this type of functions exhibits, by construction, lots of negative curvature. Hence, importantly, on these functions the assumption on the curvature of Theorem 3.3.2 is not satisfied.

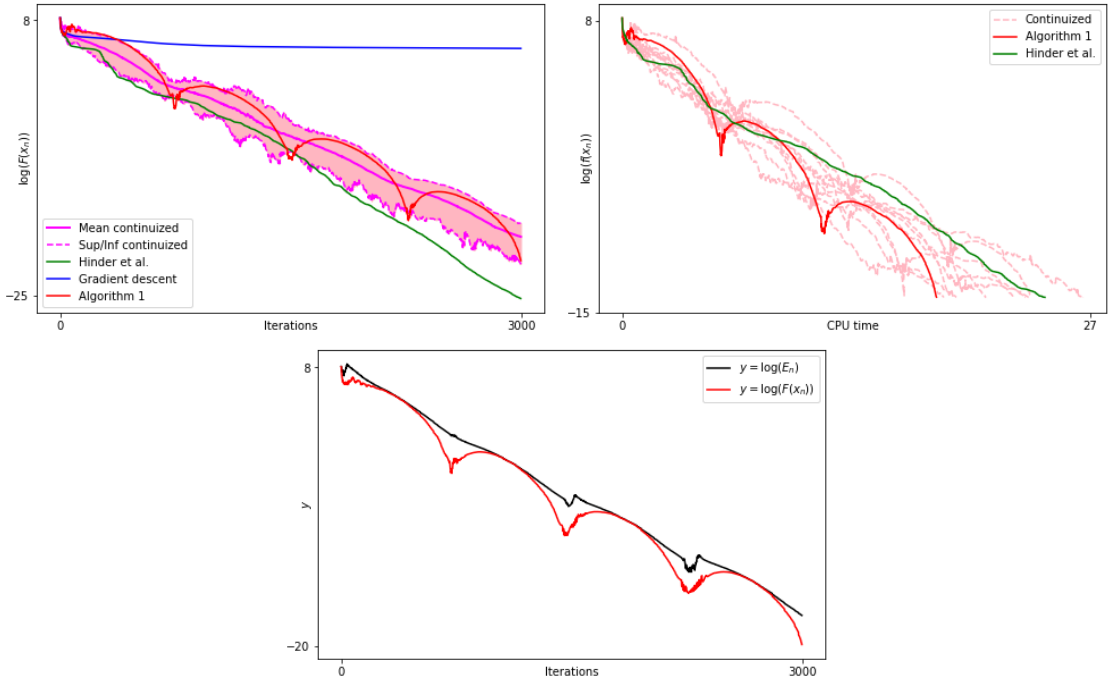



Figure 3.2: We compare the performance of the algorithm using a line search procedure from (Hinder et al. (Hinder et al., 2020)), the continuized version (Even et al., 2021), and Algorithm 13. It is done iteration wise on the top left plot, while the top-right plot compares the time needed to achieve an $\varepsilon$-solution. On the lowest plot we show the behaviour of Algorithm 13 with our choice of parameter in the presence of strong negative curvature regions.

**Description of the experiments** We performed numerical experiments on this function. We computed $L$ at each iteration with the same backtracking process as in (Hinder et al., 2020). We compared the performance of Algorithm 1 using our choice of parameters with two other methods:

1. Algorithm 1 with line search computing the $(\alpha_n)_n$ (Hinder et al., 2020).
2. Algorithm 1 with stochastic coefficients obtained using continuized framework (Even et al., 2021; Wang and Wibisono, 2023).

As the line search procedure induces more computational complexity, we compared the performance in two different ways: firstly iteration wise, and secondly we compared the CPU time needed to achieve an $\varepsilon$-precision, *i.e.* a point $\hat{x}$ such that $h(\hat{x}) - h^* < \varepsilon$, $\varepsilon > 0$.
The continuized framework leads to a stochastic algorithm and therefore to stochastic convergence results. Hence, to make the comparison with deterministic algorithms more relevant, we ran several times the algorithm. We make the following precision:

- for the iteration wise comparison, displayed on top left of figure 3.2, we ran 50 times the algorithm. We plot the mean trajectory (Mean continuized), as well as the infimum and supremum of all the trajectories along the iterations (Sup/Inf continuized). The pink zone between these two plots thus contains all the trajectories.
- for the CPU time needed to attain an $\varepsilon$-precision, displayed on top right of figure 3.2, we ran 10 times the algorithm. Here we simply plot 10 trajectories, corresponding each to a different run of the algorithm. We set $\varepsilon = 10^{-6}$.

**A note on the use of backtracking** When using this backtracking, there is necessarily a divergence with our theoretical background. The $L_n$ is computed such that $f(y_n - \frac{1}{L_n}\nabla f(y_n)) \leq f(y_n) - \frac{1}{2L_n}\|\nabla f(y_n))\|^2$. However to compute $y_n$, we need $\alpha_n$, which needs the $L_n$ we are willing to compute. We thus chose to compute $\alpha_n$ with the previous $L_{n-1}$.

**Observations** This precision being made, we can now state our observations.

1. The two top plots displayed in figure 3.2 are not very surprising: iteration wise, the binary search procedure offers a better speed. However when considering the CPU time needed to achieve an $\varepsilon$-precision, doing without line search allows for better performance as the iterations are less computationally heavy.
2. We observe empirically that a large amount of strong negative curvature encountered during the running of Algorithm 1 correlates with "bad behaviour" of the algorithm. The lowest display of figure 3.2 is characteristic of the non monotone decreasing behaviour of the Lyapunov function $E_n$ we can witness when the algorithm crosses strong negative curvature regions.

### 3.6.2 Explicit expression of the function displayed in figure 3.1

The function displayed in figure 3.1 is $h : \mathbb{R}^2 \to \mathbb{R}$, and is built in the following way:

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{3.52}$$

where $f(t) = t^2$ and

$$g(x_1, x_2) = \frac{1}{4N}\sum_{i=1}^{N}\left(a_i \sin(b_i x_1)^2 + c_i \cos(d_i x_2)^2\right) + 1 \tag{3.53}$$

with $N = 10$ and the $\{a_i\}_i$, $\{c_i\}_i$ are independently and uniformly distributed on $[0, 20]$, and the $\{b_i\}_i$, $\{d_i\}_i$ are independently and uniformly distributed on $[-25, 25]$.

## 3.A Some properties on strongly quasar-convex functions

In this section, we prove our claims about properties of strong quasar convex functions.

### 3.A.1 Strong convexity on average on segments joining the minimizer

**Proposition.** *Let $f$ be $C^2$ and $(\tau, \mu)$-strongly quasar-convex. Let $x \neq x^*$ and $t > 0$. Then we have*

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 f(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geq \tau\frac{\mu}{2} \tag{3.54}$$

*Proof.* Let $f : \mathbb{R}^d \to \mathbb{R}$ $(\tau, \mu)$-strongly quasar-convex. Define, for $t \in \mathbb{R}_+$, the function $g(t) = f(x^* + t(x - x^*))$. We have $g'(t) = \langle \nabla f(x^* + t(x - x^*)), x - x^*\rangle$. By strong quasar convexity of $f$, we have

$$f(x^* + t(x - x^*)) + \frac{1}{\tau}\langle \nabla f(x^* + t(x - x^*)), x^* - (x^* + t(x - x^*))\rangle + \frac{\mu}{2}\|x^* - (x^* + t(x - x^*))\|^2 \leq f^* \tag{3.55}$$

$$\Rightarrow g(t) - \frac{t}{\tau}g'(t) + \frac{\mu t^2}{2}\|x - x^*\|^2 \leq g(0) \tag{3.56}$$

$$\Rightarrow \tau\frac{g(t) - g(0)}{t} + \frac{\tau\mu t}{2}\|x - x^*\|^2 \leq g'(t) = \int_0^t g''(s)ds = \int_0^t \langle \nabla^2 f(x^* + s(x - x^*))x^* - x, x^* - x\rangle ds \tag{3.57}$$

Where for the last line we suppose $t > 0$. From this we deduce:

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 f(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geq \tau\frac{\mu}{2} \tag{3.58}$$

□

In particular, the above reasoning remains true for non strongly quasar-convex functions taking $\mu = 0$, inducing a on average convexity on segments joining minimizers:

$$\frac{1}{t}\int_0^t \frac{\langle \nabla^2 f(x^* + s(x - x^*))(x^* - x), x^* - x\rangle}{\|x - x^*\|^2} ds \geq 0 \tag{3.59}$$

### 3.A.2 Synthetic strongly quasar-convex example proof

Let $f : \mathbb{R} \to \mathbb{R}$ such that it is $(\tau, \mu)$-strongly quasar-convex and $f(0) = 0 = f^*$. Let $g : S^{d-1} \to \mathbb{R}$ differentiable and such that $g(x) \geq 1$ for all $x \in S^{d-1}$. We define

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right) \tag{3.60}$$

We have $h(x) \to 0$ as $x \to 0$, we extend $h$ to 0 by continuity defining $h(0) = 0$. We clearly have $h^* = h(0) = 0$.

**Proposition 3.A.1.** *$h$ is $(\tau, \mu)$-strongly quasar-convex.*

*Proof.* We have

$$h(x) - f(\|x\|) = f(\|x\|)(g\left(\frac{x}{\|x\|}\right) - 1).$$

As $g\left(\frac{x}{\|x\|}\right) - 1$ is nonnegative and $C^1$, following Appendix D.3 (Hinder et al., 2020), we have that $h(x) - f(\|x\|)$ is $\tau$-quasar convex. Thus $h(x) = h(x) - f(\|x\|) + f(\|x\|)$ is a sum of a $\tau$-quasar convex and a $\mu$-strongly quasar-convex function, which is $\mu$-strongly quasar-convex (see (Hinder et al., 2020) Appendix D.3).

□

### 3.A.3 Local strong convexity around the minimizer for $C^2$ functions

**Proposition.** *Let $f$ be $C^2$, with a unique minimizer $x^*$, with $\mu$-quadratic growth. There exists $\eta > 0$ such that $f$ is strongly convex on $B(x^*, \eta)$,*

*Proof.* **Step 1: $\nabla^2 f(x^*)$ is definite positive** We start by showing that $\nabla^2 f(x^*)$ is definite positive, *i.e.* for all $x \in \mathbb{R}^d \backslash \{0\}$ we have $\langle \nabla^2 f(x^*)x, x\rangle > 0$. Let $g(h) = f(x^* + h(x - x^*))$, $h \geq 0$ and $x \neq x^*$. We perform a second-order Taylor expansion at 0 of $g$:

$$g(h) = g(0) + hg'(0) + \frac{h^2}{2}g''(0) + o(h^2) \tag{3.61}$$

$$\Leftrightarrow f(x^* + h(x - x^*)) = f^* + h\underbrace{\langle \nabla f(x^*), x - x^*\rangle}_{=0} + \frac{h^2}{2}\langle \nabla^2 f(x^*)(x - x^*), x - x^*\rangle + o(h^2) \tag{3.62}$$

$$\Leftrightarrow f(x^* + h(x - x^*)) - f^* = \frac{h^2}{2}\langle \nabla^2 f(x^*)(x - x^*), x - x^*\rangle + o(h^2) \tag{3.63}$$

As $f$ is with $\mu$-quadratic growth, we have:

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2 \tag{3.64}$$

We thus have

$$\frac{h^2}{2}\langle \nabla^2 f(x^*)(x - x^*), x - x^*\rangle + o(h^2) \geq \frac{\mu h^2}{2}\|x - x^*\|^2 \tag{3.65}$$

$$\Rightarrow \frac{\langle \nabla^2 f(x^*)(x - x^*), x - x^*\rangle}{\|x - x^*\|^2} + \frac{2}{\|x - x^*\|^2}\frac{o(h^2)}{h^2} \geq \mu \tag{3.66}$$

Taking $h \to 0$, we get

$$\frac{\langle \nabla^2 f(x^*)(x - x^*), x - x^*\rangle}{\|x - x^*\|^2} \geq \mu \tag{3.67}$$

Taking $x = y + x^*$, we get that for all $y \in \mathbb{R}^d \backslash \{0\}$, we have $\langle \nabla^2 f(x^*)y, y\rangle \geq \mu\|y\|^2 > 0$. We showed that $\nabla^2 f(x^*)$ is definite positive. **Step 2: extension in a local vicinity** We showed in step 1 that all eigenvalues of the Hessian matrix evaluated at $x^*$ are strictly positive. As we assumed $\nabla^2 f$ is continuous, for all $\varepsilon$ such that $0 < \varepsilon < \mu$, there exists $\eta > 0$ such that for all $x \in B(x^*, \eta)$, the eigenvalues of $\nabla^2 f(x)$ are above $\mu - \varepsilon$. This means that on this ball $f$ is strongly convex. □

### 3.A.4 No local convexity for non $C^2$ functions: a non-convex pathological construction

We construct a function that exhibits pathological non-convex behaviour around its minimizer. This function is defined on $\mathbb{R}$, so this pathological behaviour does not need several dimensions to happen.

**Proof of proposition 3.5.3**

**Proposition.** *One can construct $f : [0, 1] \to \mathbb{R}$ strongly quasar-convex with minimizer $x^*$, L-smooth, such that for all $x \neq x^*$, there exists $x_0 \neq x^*$ such that:*

$$|x^* - x_0| \leq |x^* - x| \text{ and } f''(x_0) = -L \tag{3.68}$$

*Proof.* Let

$$E := \bigcup_{n\geq 1}\left(\underbrace{\left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right[}_{\text{partition of [0,1]}} \cap \underbrace{\left[\frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right]}_{\text{subpart of partition}}\right) = \bigcup_{n\geq 1}\left[\frac{7}{4}\frac{1}{2^n}, \frac{1}{2^{n-1}}\right[ \tag{3.69}$$

and

$$E_n := \bigcup_{k\geq n+1}\left[\frac{7}{4}\frac{1}{2^k}, \frac{1}{2^{k-1}}\right[ \tag{3.70}$$

Let $f$ be a function defined on $[0,1]$, such that:

$$f''(x) = \begin{cases} -L & \text{if } x \in E \\ L & \text{else.} \end{cases}$$

We suppose $f'(0) = f(0) = 0$. $f$ is clearly such that it will reach its lower curvature $-L$ in every neighborhood of its minimizer. We now want to show that $f$ is strongly quasar-convex.
Suppose $x$ is such that $\exists k \geq 1$, $x = \frac{1}{2^k}$. Then,

$$f'(x) = f'(x) - f'(0) = \int_{[0,x]} f''(s)ds = \int_{E_n} f''(s)ds + \int_{[0,x]\setminus E_n} f''(s)ds \tag{3.71}$$

By definition of $f''$, this simply becomes

$$f'(x) = -L\lambda(E_n) + L\lambda([0,x] \setminus E_n) = L(x - 2\lambda(E_n)) \tag{3.72}$$

Where $\lambda(.)$ is the Lebesgue measure. But by construction $\lambda(E_n) = \frac{1}{4}x$, which means

$$f'(x) = \frac{L}{2}x \tag{3.73}$$

Now suppose $x = \frac{1}{2^k} + \varepsilon$, where $k \geq 1$ and $0 < \varepsilon \leq \frac{3}{4}\frac{1}{2^k}$. This time we get

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^k}]} f''(s)ds + \int_{[\frac{1}{2^k},x]} f''(s)ds = \frac{L}{2}\frac{1}{2^k} + L\varepsilon \geq \frac{L}{2}(\frac{1}{2^k} + \varepsilon) = \frac{L}{2}x \tag{3.74}$$

finally, suppose $x = \frac{1}{2^n} + \frac{3}{4}\frac{1}{2^n} + \varepsilon$, where $n \geq 1$ and $0 < \varepsilon < \frac{1}{4}\frac{1}{2^n}$.

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^n}+\frac{3}{4}\frac{1}{2^n}]} f''(s)ds + \int_{[\frac{1}{2^n}+\frac{3}{4}\frac{1}{2^n},x]} f''(s)ds = \frac{L}{2}\frac{1}{2^n} + \frac{3L}{4}\frac{1}{2^n} - L\varepsilon \geq \frac{L}{2^n} \tag{3.75}$$

But here $x \leq \frac{1}{2^{n-1}}$, which gives

$$f'(x) \geq \frac{L}{2}\frac{1}{2^{n-1}} = \frac{L}{2}x \tag{3.76}$$

finally, for all $x \in [0,1]$, we have $f'(x) \geq \frac{L}{2}x$. To prove that this function is strongly quasar-convex, remark by definition of $f''$ that $f(x) \leq \frac{L}{2}x^2$, and using $f'(x) \geq \frac{L}{2}x$, we have that for any $\mu > 0$:

$$f(x) - \frac{\mu + L}{L}f'(x)x + \frac{\mu}{2}x^2 \leq \frac{L}{2}x^2 - \frac{\mu + L}{2}x^2 + \frac{\mu}{2}x^2 = 0 = f^* \tag{3.77}$$

In words we showed that $f$ is $(\frac{L}{\mu+L}, \mu)-$strongly quasar-convex. □

**Quadratic growth and local uniqueness of critical points ?**

We recall that a function satisfies $\mu$-quadratic growth hypothesis if it satisfies

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2 \tag{3.78}$$

Suppose such a function has a unique minimizer. Regarding the local regularity around $x^*$ provided by this hypothesis, it is not obvious whether there exists a vicinity around this minimizer such that there is no other critical points. We provide a counter example.

**Proposition 3.A.2.** *There exists a quadratic growth function $f : [0,1] \to \mathbb{R}$ such that in every neighborhood of its minimizer, $f$ has an infinite amount of critical points.*

*Proof.* Let

$$E := \bigcup_{n\geq 1}\left(\left[\frac{1}{2^n},\frac{1}{2^{n-1}}\right[ \cap \left[\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n},\frac{1}{2^{n-1}}\right]\right) = \bigcup_{n\geq 1}\left[\frac{3}{2}\frac{1}{2^n},\frac{1}{2^{n-1}}\right[ \tag{3.79}$$

and

$$E_n := \bigcup_{k\geq n+1}\left[\frac{3}{2}\frac{1}{2^k},\frac{1}{2^{k-1}}\right[ \tag{3.80}$$

Let $f$ be a function defined on $[0,1]$, such that:

$$f''(x) = \begin{cases} -1 & \text{if } x \in E \\ 1 & \text{else.} \end{cases}$$

We suppose $f'(0) = f(0) = 0$. Suppose $x$ is such that $\exists n \geq 0$, $x = \frac{1}{2^n}$. Then,

$$f'(x) = f'(x) - f'(0) = \int_{[0,x]} f''(s)ds = \int_{E_n} f''(s)ds + \int_{[0,x]\setminus E_n} f''(s)ds \tag{3.81}$$

By definition of $f''$, this simply becomes

$$f'(x) = -\lambda(E_n) - \lambda([0,x]\setminus E_n) = x - 2\lambda(E_n) \tag{3.82}$$

Where $\lambda(.)$ is the Lebesgue measure. But by construction $\lambda(E_n) = \frac{1}{2}x$, which means

$$f'(x) = 0 \tag{3.83}$$

This means our function has infinite amount of critical point in all vicinity of minimizer. We want now to show that $f$ is a quadratic growth function.
We suppose $x = \frac{1}{2^k} + \varepsilon$, where $k \geq 1$ and $0 < \varepsilon \leq \frac{1}{2}\frac{1}{2^k}$. This time we get

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^k}]} f''(s)ds + \int_{[\frac{1}{2^k},x]} f''(s)ds = \varepsilon \tag{3.84}$$

Now suppose $x = \frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n} + \varepsilon$, where $n \geq 1$ and $0 < \varepsilon < \frac{1}{2}\frac{1}{2^n}$.

$$f'(x) = \int_{[0,x]} f''(s)ds = \int_{[0,\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n}]} f''(s)ds + \int_{[\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n},x]} f''(s)ds = \frac{1}{2}\frac{1}{2^n} - \varepsilon \tag{3.85}$$

We have then

$$f\left(\frac{1}{2^{n-1}}\right) - f\left(\frac{1}{2^n}\right) = 2\int_{[\frac{1}{2^n},\frac{1}{2^n}+\frac{1}{2}\frac{1}{2^n}]} f'(s)ds = 2\int_{[\frac{1}{2^n},\frac{3}{2}\frac{1}{2^n}]} sds = \frac{5}{4}\frac{1}{2^{2n}} \tag{3.86}$$

Summing over $n$ we get

$$f\left(\frac{1}{2^{n-1}}\right) = \frac{5}{4}\sum_{i\geq n}\frac{1}{2^{2n}} = \frac{5}{4}\frac{\frac{1}{4^n}}{1-\frac{1}{4}} = \frac{5}{12}\frac{1}{2^{2n-2}} = \frac{5}{12}\left(\frac{1}{2^{n-1}}\right)^2 \tag{3.87}$$

If $x = \frac{1}{2^k} + \varepsilon$, where $k \geq 1$ and $0 < \varepsilon \leq \frac{1}{2}\frac{1}{2^k}$, We have

$$f(x) - f\left(\frac{1}{2^n}\right) = \int_{[\frac{1}{2^n},x]} f'(s)ds = \frac{1}{2}\left(x^2 - \frac{1}{2^{2n}}\right) \tag{3.88}$$

$$\Rightarrow f(x) = \frac{5}{12}\frac{1}{2^{2n}} + \frac{1}{2}x^2 - \frac{6}{12}\frac{1}{2^{2n}} = \frac{5}{12}x^2 + \frac{1}{12}x^2 - \frac{1}{12}\frac{1}{2^{2n}} \tag{3.89}$$

However as $x \geq \frac{1}{2^n}$, we get that $f(x) \geq \frac{5}{12}x^2$.
Now suppose $x = \frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n} + \varepsilon$, where $n \geq 1$ and $0 < \varepsilon < \frac{1}{2}\frac{1}{2^n}$. We show by contradiction that $f(x) \geq \frac{5}{12}x^2$. Suppose $f(x) < \frac{5}{12}x^2$. Note that as we showed

$$f'(x) = \frac{1}{2}\frac{1}{2^n} - \varepsilon \tag{3.90}$$

we have

$$\frac{5}{6}x - f'(x) = \frac{15}{12}\frac{1}{2^n} + \frac{5}{6}x_2 - \frac{1}{2}\frac{1}{2^n} + \varepsilon = \frac{6}{12}\frac{1}{2^n} + \frac{9}{12} \geq 0 \tag{3.91}$$

So we have both:

$$f(x) < \frac{5}{12}x^2 \text{ and } f'(x) \leq \left(\frac{5}{12}x^2\right)' = \frac{5}{6}x \tag{3.92}$$

The second inequality holds for all $y$ in $[\frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n}, \frac{1}{2^{n-1}}]$, inducing that $f(y) < \frac{5}{12}y^2$. However, we already showed that $f(\frac{1}{2^{n-1}}) = \frac{5}{12}\left(\frac{1}{2^{n-1}}\right)^2$, leading to a contradiction. We conclude that $f(x) \geq \frac{5}{12}x^2$.
We then showed that

$$f(x) = f(x) - f^* \geq \frac{\frac{5}{12}}{2}x^2 = \frac{\frac{5}{12}}{2}(x - x^*)^2 \tag{3.93}$$

or in words, $f$ has infinite amount of critical point in each vicinity of minimizer and is $\frac{5}{12}$-quadratic growth. □

Finally, we show that strong quasar convexity implies the PL condition.

**Proposition 3.A.3.** *Let $f$ be $(\tau, \mu)$-strongly quasar-convex. It is then $\mu\tau^2$-PL.*

*Proof.* We have

$$\frac{1}{2}\|\frac{1}{\tau\sqrt{\mu}}\nabla f(x) + \sqrt{\mu}(x^* - x)\|^2 = \frac{1}{2\tau^2\mu}\|\nabla f(x)\|^2 + \frac{1}{\tau}\langle\nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2 \tag{3.94}$$

Writing the definition of $(\tau, \mu)$-strong quasar convexity, we have

$$f^* \geq f(x) + \frac{1}{\tau}\langle\nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2 \tag{3.95}$$

$$= f(x) + \frac{1}{2}\|\frac{1}{\tau\sqrt{\mu}}\nabla f(x) + \sqrt{\mu}(x^* - x)\|^2 - \frac{1}{2\tau^2\mu}\|\nabla f(x)\|^2 \tag{3.96}$$

$$\Rightarrow \frac{1}{2\tau^2\mu}\|\nabla f(x)\|^2 \geq f(x) - f^* \tag{3.97}$$

□

# 3.B Proofs of section 3

## 3.B.1 Differentiable strongly quasar-convex

### Proof of Theorem 3.3.2

In this section, we detail the proof of Theorem 3.3.2 whose statement is recalled here: let $f$ be a $(\tau, \mu)$-strongly quasar-convex function for some $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$ having a $(\rho, L)$ curvature for some $L > 0$ and $\rho \leq L$. Let $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 13:

$$\begin{aligned} y_n &= \alpha_n x_n + (1 - \alpha_n)z_n \\ x_{n+1} &= y_n - s\nabla f(y_n) \\ z_{n+1} &= \beta_n z_n + (1 - \beta_n)y_n - \eta_n\nabla f(y_n) \end{aligned}$$

with parameters

$$s \leq \frac{1}{L}, \quad \alpha_n = \frac{1}{1 + \sqrt{\mu s}} := \alpha, \quad \beta_n = 1 - \tau\sqrt{\mu s} := \beta, \quad \eta_n = \frac{\sqrt{s}}{\sqrt{\mu}} := \eta.$$

Assuming that $\rho \geq -\tau\sqrt{\frac{\mu}{s}}$ we want to prove that:

$$\forall n \in \mathbb{N},\ f(x_n) - f^* \leq \frac{2}{\tau}\left(1 - \tau\sqrt{\mu s}\right)^n \left(f(x_0) - f^*\right) \tag{3.98}$$

where $f^* = \min\ f$. Let $x^*$ be the unique minimizer of $f$. Let be the following Lyapunov energy:

$$E_n = f(x_n) - f^* + \frac{\mu}{2}\|z_n - x^*\|^2 \tag{3.99}$$

The main idea of the proof consists in finding parameters such that the following inequality holds

$$E_{n+1} - E_n \leq cE_n \tag{3.100}$$

with $c < 0$ being as small as possible. We will then deduce the convergence rate (3.98) by induction.

**Step 1.** Since:

$$E_{n+1} - E_n = f(x_{n+1}) - f(x_n) + \frac{\mu}{2}\|z_{n+1} - x^*\|^2 - \frac{\mu}{2}\|z_n - x^*\|^2, \tag{3.101}$$

let us start by considering the right term:

$$\begin{aligned}
\Delta_n &= \|z_{n+1} - x^*\|^2 - \|z_n - x^*\|^2 \\
&= \|\beta z_n + (1-\beta)y_n - \eta\nabla f(y_n) - x^*\|^2 - \|z_n - x^*\|^2 \\
&= (\beta^2 - 1)\|z_n - x^*\|^2 + (1-\beta)^2\|y_n - x^*\|^2 + \eta^2\|\nabla f(y_n)\|^2 \\
&\quad +2\beta\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla f(y_n)\rangle - 2(1-\beta)\eta\langle \nabla f(y_n), y_n - x^*\rangle
\end{aligned}$$

by construction of Algorithm 13. The tricky part here is to control the first scalar product: using the definition of Algorithm 13, we can rewrite $z_n = y_n + \frac{\alpha}{1-\alpha}(y_n - x_n)$, and thus

$$\begin{aligned}
&\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla f(y_n)\rangle \\
&= \langle y_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla f(y_n)\rangle + \frac{\alpha}{1-\alpha}\langle y_n - x_n, (1-\beta)(y_n - x^*) - \eta\nabla f(y_n)\rangle \\
&= (1-\beta)\|y_n - x^*\|^2 - \eta\langle y_n - x^*, \nabla f(y_n)\rangle - \frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \nabla f(y_n)\rangle \\
&+\frac{\alpha}{1-\alpha}(1-\beta)\langle y_n - x_n, y_n - x^*\rangle
\end{aligned}$$

Applying the relation $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ to $a = y_n - x^*$ and $b = \frac{\alpha}{1-\alpha}(y_n - x_n)$, we get:

$$\frac{\alpha}{1-\alpha}\langle y_n - x_n, y_n - x^*\rangle = \frac{1}{2}\|z_n - x^*\|^2 - \frac{1}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{1}{2}\|y_n - x^*\|^2, \tag{3.102}$$

so that:

$$\begin{aligned}
&\langle z_n - x^*, (1-\beta)(y_n - x^*) - \eta\nabla f(y_n)\rangle \\
&= \frac{1-\beta}{2}\left(\|z_n - x^*\|^2 + \|y_n - x^*\|^2 - \left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2\right) - \eta\langle y_n - x^*, \nabla f(y_n)\rangle \\
&-\frac{\alpha}{1-\alpha}\eta\langle y_n - x_n, \nabla f(y_n)\rangle
\end{aligned}$$

and

$$\begin{aligned}
\Delta_n &= -(1-\beta)\|z_n - x^*\|^2 + (1-\beta)\|y_n - x^*\|^2 + \eta^2\|\nabla f(y_n)\|^2 \\
&\quad -\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - 2\frac{\alpha\beta\eta}{1-\alpha}\langle\nabla f(y_n), y_n - x_n\rangle - 2\eta\langle\nabla f(y_n), y_n - x^*\rangle.
\end{aligned}$$

Reinjecting $\Delta_n$ in the expression of $E_{n+1} - E_n$ and by definition of the Lyapunov energy $E_n$, we get:

$$\begin{aligned}
E_{n+1} - E_n = &-(1-\beta)E_n + f(x_{n+1}) - f^* - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x^*\|^2 + \frac{\mu}{2}\eta^2\|\nabla f(y_n)\|^2 \\
&- \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle\nabla f(y_n), y_n - x_n\rangle - \mu\eta\langle\nabla f(y_n), y_n - x^*\rangle.
\end{aligned} \tag{3.103}$$

**Step 2.** Let us now prove that for any $n \in \mathbb{N}$, we have: $E_{n+1} - E_n \leq -(1-\beta)E_n$ for some well-chosen values of the parameters $\beta$, $\eta$ and $\alpha$.

Remember that $f$ is assumed strongly quasar-convex, hence:

$$\forall n \in \mathbb{N},\ \langle \nabla f(y_n), y_n - x^* \rangle \geq \tau(f(y_n) - f^*) + \frac{\tau\mu}{2}\|y_n - x^*\|^2 \tag{3.104}$$

and $L$-smooth which induces that:

$$\forall s \leq \frac{1}{L},\ \forall n \in \mathbb{N},\ \frac{s}{2}\|\nabla f(y_n)\|^2 \leq f(y_n) - f(x_{n+1}) \tag{3.105}$$

Reinjecting these two inequalities into $E_{n+1} - E_n$, we then get:

$$\begin{aligned}
E_{n+1} - E_n &\leq -(1-\beta)E_n + f(x_{n+1}) - f^* - \beta\left(f(x_n) - f^*\right) + \frac{\mu}{s}\eta^2(f(y_n) - f(x_{n+1})) \\
&\quad - \frac{\mu\beta(1-\beta)}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle \nabla f(y_n), y_n - x_n\rangle - \eta\mu\tau(f(y_n) - f^*) \\
&\quad + \frac{\mu}{2}(1 - \beta - \tau\eta\mu)\|y_n - x^*\|^2 \\
&\leq -(1-\beta)E_n + (\frac{\mu}{s}\eta^2 - \tau\mu\eta)(f(y_n) - f^*) + (1 - \frac{\mu}{s}\eta^2)(f(x_{n+1}) - f^*) \\
&\quad - \beta(f(x_n) - f^*) - \frac{\alpha\beta\eta\mu}{1-\alpha}\langle \nabla f(y_n), y_n - x_n\rangle - \frac{\mu\beta(1-\beta)}{2}\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \\
&\quad + \frac{\mu}{2}(1 - \beta - \tau\eta\mu)\|y_n - x^*\|^2
\end{aligned} \tag{3.106}$$

Choosing now $\eta = \sqrt{\frac{s}{\mu}}$ and $\beta = 1 - \tau\eta\mu = 1 - \tau\sqrt{\mu s}$ to cancel out the terms in $f(x_{n+1}) - f^*$ and $\|y_n - x^*\|^2$, we deduce:

$$\begin{aligned}
E_{n+1} - E_n &\leq -\tau\sqrt{\mu s}E_n + (1 - \tau\sqrt{\mu s})\left(f(y_n) - f(x_n)\right) \\
&\quad + (1 - \tau\sqrt{\mu s})\sqrt{\mu s}\frac{\alpha}{1-\alpha}\left(\langle \nabla f(y_n), x_n - y_n\rangle - \frac{\alpha}{1-\alpha}\frac{\tau\mu}{2}\|y_n - x_n\|^2\right).
\end{aligned} \tag{3.107}$$

Suppose additionally that the lower curvature is bounded from below by $\rho$, hence:

$$\forall n \in \mathbb{N},\ f(y_n) + \langle \nabla f(y_n), x_n - y_n\rangle + \frac{\rho}{2}\|x_n - y_n\|^2 \leq f(x_n),$$

or equivalently:

$$\forall n \in \mathbb{N},\ \langle \nabla f(y_n), x_n - y_n\rangle \leq f(x_n) - f(y_n) - \frac{\rho}{2}\|x_n - y_n\|^2.$$

Injecting the very last inequality into (3.107) we get:

$$\begin{aligned}
E_{n+1} - E_n &\leq -\tau\sqrt{\mu s}E_n + (1 - \tau\sqrt{\mu s})\left(1 - \sqrt{\mu s}\frac{\alpha}{1-\alpha}\right)\left(f(y_n) - f(x_n)\right) \\
&\quad + (1 - \tau\sqrt{\mu s})\frac{\sqrt{\mu s}}{2}\frac{\alpha}{1-\alpha}\left(-\rho - \frac{\alpha}{1-\alpha}\tau\mu\right)\|y_n - x_n\|^2
\end{aligned}$$

Lastly, choose $\alpha = \frac{1}{1+\sqrt{\mu s}}$ to cancel out the term in $f(y_n) - f(x_n)$, so that we finally get:

$$E_{n+1} - E_n \leq -\tau\sqrt{\mu s}E_n + \frac{1 - \tau\sqrt{\mu s}}{2}\left(-\rho - \tau\frac{\sqrt{\mu}}{\sqrt{s}}\right)\|y_n - x_n\|^2 \leq -\tau\sqrt{\mu s}E_n$$

provided that the lower curvature satisfies: $\rho \geq -\tau\frac{\sqrt{\mu}}{\sqrt{s}}$.

**Step 3.** To conclude, we proved that:

$$E_{n+1} - E_n \leq -\tau\sqrt{\mu s}E_n \Rightarrow E_{n+1} \leq (1 - \tau\sqrt{\mu s})E_n \tag{3.108}$$

By induction:

$$\forall n \in \mathbb{N},\ E_{n+1} \leq (1 - \tau\sqrt{\mu s})^{n+1} E_0 = (1 - \tau\sqrt{\mu s})^{n+1} \left(f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) \tag{3.109}$$

In the last equality we use that by definition of Algorithm 13 $x_0 = z_0$. We then use the fact that $(\tau, \mu)$-strong quasar-convexity implies $\frac{\tau\mu}{2-\tau}$-quadratic growth (corollary 1 (Hinder et al., 2020)), *i.e.*

$$\forall x \in \mathbb{R}^d,\ f(x) - f^* \geq \frac{\tau\mu}{2(2-\tau)}\|x - x^*\|^2 \tag{3.110}$$

which finally yields:

$$f(x_n) - f^* \leq \frac{2}{\tau}\left(1 - \tau\sqrt{\mu s}\right)^n (f(x_0) - f^*). \tag{3.111}$$

**two-point scheme version of algorithm 1**

Here we build upon the work (Lee et al., 2021), where they show there exists an equivalence between a three-point and two-point scheme version of Nesterov Momentum. We want to deduce a two-point scheme from our three-point one, but we cannot directly apply their result because the three-point algorithm they consider is slightly different.

**Proposition 3.B.1.** *The algorithm*

$$\begin{cases} y_n = \alpha_n x_n + (1 - \alpha_n) z_n \\ x_{n+1} = y_n - s\nabla f(y_n) \\ z_{n+1} = \beta_n z_n + (1 - \beta_n) y_n - \eta_n \nabla f(y_n) \end{cases} \tag{3.112}$$

*can be written as the following two-point scheme*

$$\begin{cases} y_n = x_n + \frac{1-\alpha_n}{1-\alpha_{n-1}}\alpha_{n-1}\beta_{n-1}(x_n - x_{n-1}) + (1 - \alpha_n)\left(\frac{\eta_{n-1}}{s} - \frac{\alpha_{n-1}\beta_{n-1}}{1-\alpha_{n-1}} - 1\right)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla f(y_n) \end{cases} \tag{3.113}$$

*Proof.* We adapt the proof of Lemma 2 from (Lee et al., 2021), which is mainly based on Thales theorem. In their result, the $\alpha_n$ is $\frac{\varphi_n}{\varphi_{n-1}}$ and $\beta_n z_n + (1 - \beta_n) y_n$ is $z_n$. The main idea is that if a vector $u$ is collinear to $v$, *i.e.* $\exists \lambda \in \mathbb{R}\ u = \lambda v$, then $\lambda = \frac{\|u\|}{\|v\|}$, which we can find using Thales theorem. As in the original proof, let us rewrite (3.112) as:

$$\begin{cases} x_n = \alpha_n x_{n-1}^+ + (1 - \alpha_n) z_n \\ z_{n+1} = \beta_n z_n + (1 - \beta_n) x_n - \eta_n \nabla f(x_n) \end{cases} \tag{3.114}$$

where $x_{n-1}^+ = x_{n-1} - s\nabla f(x_{n-1})$. We suppose $\nabla f(x_{n-1}) \neq 0$ (non degenerate case). We set $v_n = \beta_n z_n + (1 - \beta_n) x_n$, and let's set A on the $[x_{n-1}^+ x_n]$ segment such that $Ax_{n+1} \parallel x x_n^+$. Let $B$ on $x_{n+1}^+ x_n^+ \cap v_n z_{n+1}$. Since $Ax_{n+1} \parallel Bz_{n+1}$, we have by Thales theorem:

$$\frac{\|B - A\|}{\|B - x_n^+\|} = \frac{\|z_{n+1} - x_{n+1}\|}{\|z_{n+1} - x_n^+\|} := (\star) \tag{3.115}$$

Then, by definition:

$$z_{n+1} - x_{n+1} = \alpha_{n+1}(z_{n+1} - x_n^+) \Rightarrow \alpha_{n+1} = (\star) \tag{3.116}$$

As $B - A = \lambda(B - x_n^+)$ for some $\lambda \in \mathbb{R}$ (collinearity), we have $\lambda = \alpha_{n+1}$ and then

$$B - A = \alpha_{n+1}(B - x_n^+) \tag{3.117}$$

Also, collinearity of $Bx_n^+$ and $x_n^+x_{n-1}^+$ together with $x_nx_n^+ \parallel v_nB$ leads to $B - x_n^+ = \lambda(x_n^+ - x_{n-1}^+)$, with

$$\lambda = \frac{\|B - x_n^+\|}{\|x_n^+ - x_{n-1}^+\|} = \frac{\|v_n - x_n\|}{\|x_n - x_{n-1}^+\|} \tag{3.118}$$

where we have

$$v_n - x_n = \beta_n z_n + (1-\beta_n)x_n - x_n = \beta_n(z_n - x_n) = \beta_n\alpha_n(z_n - x_{n-1}^+) \tag{3.119}$$

$$x_n - x_{n-1}^+ = (1-\alpha_n)(z_n - x_{n-1}^+) \tag{3.120}$$

such that

$$B - x_n^+ = \beta_n \frac{\alpha_n}{1-\alpha_n}(x_n^+ - x_{n-1}^+) \tag{3.121}$$

Combining (3.117) and (3.121), we get

$$A - x_n^+ = (B - x_n^+) - (B - A) = (B - x_n^+) = (B - x_n^+) - \alpha_{n+1}(B - x_n^+) \tag{3.122}$$

$$= (1-\alpha_{n+1})(B - x_n^+) = \beta_n \frac{(1-\alpha_{n+1})\alpha_n}{1-\alpha_n}(x_n^+ - x_{n-1}^+) \tag{3.123}$$

Then, we study $x_{n+1} - A$. As $Ax_{n+1} \parallel Bz_{n+1}$ we have by Thales theorem

$$\frac{\|x_{n+1} - A\|}{\|z_{n+1} - B\|} = \frac{\|x_{n+1} - x_n^+\|}{\|z_{n+1} - x_n^+\|} = 1 - \alpha_{n+1} \tag{3.124}$$

Last equality because $x_{n+1} - x_n^+ = (1-\alpha_{n+1})(z_{n+1} - x_n^+)$. We then have

$$x_{n+1} - A = (1-\alpha_{n+1})(z_{n+1} - B) = (1-\alpha_{n+1})((z_{n+1} - v_n) - (B - v_n)) \tag{3.125}$$

Where we recall $v_n = \beta_n z_n + (1-\beta_n)x_n$. We have:

$$z_{n+1} - v_n = (x_n^+ - x_n)\frac{\eta_n}{s} \tag{3.126}$$

Then we use $x_nx_n^+ \parallel v_nB$ to us Thales theorem once again:

$$\frac{\|B - v_n\|}{\|x_n^+ - x_n\|} = \frac{\|v_n - x_{n-1}^+\|}{\|x_n - x_{n-1}^+\|} := (\star\star) \tag{3.127}$$

We have $v_n - x_{n-1}^+ = \beta_n z_n + (1-\beta_n)x_n - x_{n-1}^+$. We have also

$$x_n = \alpha_n x_{n-1}^+ + (1-\alpha_n)z_n \tag{3.128}$$

$$\Rightarrow \beta_n z_n - \beta_n x_n = \alpha_n\beta_n(z_n - x_{n-1}^+) \tag{3.129}$$

$$\Rightarrow \beta_n z_n + (1-\beta_n)x_n - x_{n-1}^+ = \alpha_n\beta_n(z_n - x_{n-1}^+) + x_n - x_{n-1}^+ \tag{3.130}$$

$$\Rightarrow v_n - x_{n-1}^+ = (\frac{\alpha_n\beta_n}{1-\alpha_n} + 1)(x_n - x_{n-1}^+) \tag{3.131}$$

This induces that $(\star\star) = (\frac{\alpha_n\beta_n}{1-\alpha_n} + 1)$ and then

$$B - v_n = \left(\frac{\alpha_n\beta_n}{1-\alpha_n} + 1\right)(x_n^+ - x_n) \tag{3.132}$$

finally, injecting (3.126) and (3.132) in (3.125), we get

$$x_{n+1} - A = (1-\alpha_{n+1})\left(\frac{\eta_n}{s}(x_n^+ - x_n) - (\frac{\alpha_n\beta_n}{1-\alpha_n} + 1)(x_n^+ - x_n)\right) \tag{3.133}$$

$$= (1-\alpha_{n+1})\left(\frac{\eta_n}{s} - \frac{\alpha_n\beta_n}{1-\alpha_n} + 1\right)(x_n^+ - x_n) \tag{3.134}$$

We can conclude by combining (3.122) and (3.133) that

$$x_{n+1} = x_n^+ + \beta_n\frac{(1-\alpha_{n+1})\alpha_n}{1-\alpha_n}(x_n^+ - x_{n-1}^+) + (1-\alpha_{n+1})\left(\frac{\eta_n}{s} - \frac{\alpha_n\beta_n}{1-\alpha_n} + 1\right)(x_n^+ - x_n) \tag{3.135}$$

□

**Corollary 3.B.1.** *The algorithm 1 with parameters* $s \leq \frac{1}{L}$, $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \tau\sqrt{\mu s}$ *and* $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$ *can be written as the following two-point scheme*

$$\begin{cases} y_n = x_n + \frac{1-\tau\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\tau - 1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla f(y_n) \end{cases} \tag{3.136}$$

*Proof.* We just apply previous result with $\alpha_n = \frac{1}{1+\sqrt{\mu s}}$, $\beta_n = 1 - \tau\sqrt{\mu s}$ and $\eta_n = \frac{\sqrt{s}}{\sqrt{\mu}}$. □

**Proof of theorem 3.3.3**

In this section, we detail the proof of Theorem 3.3.3 whose statement is: let $F = f + g$ where $f : \mathbb{R}^d \to \mathbb{R}$ is an $L$-smooth function for some $L > 0$ and $g : \mathbb{R}^d \to \mathbb{R}$ is convex, proper, lower semi-continuous. Assume that $F$ has a nonempty set of minimizers and that $f$ is $(1, \mu)$-strongly quasar-convex with respect to $x_F^* \in \arg\min\ F$ and $(\rho, L)$-curvature-bounded for some $\rho \leq L$. Let $(x_n)_{n\in\mathbb{N}}$ be generated by Algorithm 14:

$$\begin{aligned} y_n &= \alpha_n x_n + (1-\alpha_n) z_n \\ x_{n+1} &= \operatorname{prox}_{sg}(y_n - s\nabla f(y_n)) := T_s(y_n) \\ z_{n+1} &= \beta_n z_n + (1-\beta_n) y_n - \frac{\eta_n}{s}(y_n - T_s(y_n)) \end{aligned}$$

with parameters

$$s \leq \frac{1}{L}, \quad \alpha_n = \frac{1}{1+\sqrt{\mu s}} := \alpha, \quad \beta_n = 1 - \sqrt{\mu s} := \beta, \quad \eta_n = \frac{\sqrt{s}}{\sqrt{\mu}} := \eta.$$

Assuming that $\rho \geq -\sqrt{\frac{\mu}{s}}$ we want to prove that:

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leq 2\left(1 - \sqrt{\mu s}\right)^n \left(F(x_0) - F^*\right) \tag{3.137}$$

where $F^* = \min\ F$. Let $x_F^*$ be the unique minimizer of $F$. As for the proof of Theorem 3.3.2, we introduce the following Lyapunov energy:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x_F^*\|^2 \tag{3.138}$$

and we seek the parameters and conditions for which the following inequality holds:

$$E_{n+1} - E_n \leq cE_n \tag{3.139}$$

with $c < 0$ being as small as possible. We will then deduce the convergence rate (3.137) by induction.

*Proof.* The proof is very similar to Theorem 3.3.2. Let us consider the same Lyapunov energy:

$$E_n = F(x_n) - F^* + \frac{\mu}{2}\|z_n - x_F^*\|^2 \tag{3.140}$$

**Step 1.** We start by calculating $E_{n+1} - E_n$, and the exact same computations as in Step 1 for Theorem 3.3.2 leads to:

$$\begin{aligned} E_{n+1} - E_n = &-(1-\beta)E_n + F(x_{n+1}) - F^* - \beta\left(F(x_n) - F^*\right) + \frac{\mu}{2}(1-\beta)\|y_n - x_F^*\|^2 \\ &+ \frac{\mu}{2s^2}\eta^2\|y_n - T_s(y_n)\|^2 - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 \\ &+ \frac{\mu\eta}{s}\langle y_n - T_s(y_n), x_F^* - y_n\rangle + \frac{\alpha\beta\eta\mu}{(1-\alpha)s}\langle y_n - T_s(y_n), x_n - y_n\rangle, \end{aligned} \tag{3.141}$$

replacing $\nabla f(y_n)$ by the composite gradient $\frac{1}{s}(y_n - T_s(y_n))$, where $T_s(y_n) := \operatorname{prox}_{sg}(y_n - s\nabla f(y_n))$.

**Step 2.** Let us now prove that for any $n \in \mathbb{N}$, we have: $E_{n+1} - E_n \leq -(1-\beta)E_n$ for some well-chosen values of the parameters $\beta$, $\eta$ and $\alpha$.

To control the scalar products, first note that:

$$2\langle y_n - T_s(y_n), x_F^* - y_n\rangle = \|T_s(y_n) - x_F^*\|^2 - \|y_n - T_s(y_n)\|^2 - \|y_n - x_F^*\|^2. \tag{3.142}$$

Combining the prox-grad inequality ((Beck, 2017, Theorem 10.16)): for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$F(x) - F(T_s(y)) \geq \frac{1}{2s}\|x - T_s(y)\|^2 - \frac{1}{2s}\|x - y\|^2 + f(x) - f(y) - \langle \nabla f(y), x - y\rangle, \tag{Prox-Grad}$$

applied at $x = x_F^*$ and $y = y_n$, and the definition of strong quasar convexity in the sense of Definition 3.3.1:

$$\forall n \in \mathbb{N},\ f(x_F^*) - f(y_n) - \langle \nabla f(y_n), x_F^* - y_n\rangle \geq \frac{\mu}{2}\|x_F^* - y_n\|^2,$$

we prove that for any $n \in \mathbb{N}$,

$$\begin{aligned} 2s\left(F^* - F(T_s(y_n))\right) &\geq \|x_F^* - T_s(y_n)\|^2 - \|x_F^* - y_n\|^2 + 2s\left(f(x_F^*) - f(y_n) - \langle \nabla f(y_n), x_F^* - y_n\rangle\right) \\ &\geq \|x_F^* - T_s(y_n)\|^2 - \|x_F^* - y_n\|^2 + \mu s\|x_F^* - y_n\|^2. \end{aligned}$$

Hence, reinjecting into (3.142) and remembering $x_{n+1} = T_s(y_n)$, we get:

$$\forall n \in \mathbb{N},\ \langle y_n - T_s(y_n), x_F^* - y_n\rangle \leq -s\left(F(x_{n+1}) - F^*\right) - \frac{\mu s}{2}\|y_n - x_F^*\|^2 - \frac{1}{2}\|y_n - x_{n+1}\|^2. \tag{3.143}$$

Similarly, consider then the second scalar product:

$$\begin{aligned} 2\langle x_n - y_n,\ y_n - T_s(y_n)\rangle &= \|T_s(y_n) - x_n\|^2 - \|y_n - T_s(y_n)\|^2 - \|y_n - x_n\|^2 && (3.144) \\ &\leq 2s\left(F(x_n) - F(T_s(y_n))\right) + 2s(f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle) \\ &\quad -\|y_n - T_s(y_n)\|^2 && (3.145) \\ &\leq 2s\left(F(x_n) - F(x_{n+1})\right) + 2s(f(y_n) - f(x_n) + \langle \nabla f(y_n), x_n - y_n\rangle) \\ &\quad -\|y_n - x_{n+1}\|^2 && (3.146) \end{aligned}$$

using again (Prox-Grad) evaluated at $x = x_n$ and $y = y_n$ and $x_{n+1} = T_s(y_n)$.

Reinjecting (3.143) and (3.146) into the expression of $E_{n+1} - E_n$ obtained at the end of Step 1, we get:

$$\begin{aligned} E_{n+1} - E_n \leq\ & -(1-\beta)E_n + \left(1 - \mu\eta - \mu\beta\frac{\alpha}{1-\alpha}\eta\right)\left(F(x_{n+1}) - F^*\right) + \beta\left(\frac{\alpha\eta\mu}{1-\alpha} - 1\right)\left(F(x_n) - F^*\right) \\ & + \frac{\mu}{2}\left(1 - \beta - \mu\eta\right)\|y_n - x_F^*\|^2 + \frac{\mu\eta}{2s}\left(\frac{\eta}{s} - 1 - \frac{\alpha\beta}{1-\alpha}\right)\|y_n - x_{n+1}\|^2 \\ & - \frac{\mu}{2}\beta(1-\beta)\left(\frac{\alpha}{1-\alpha}\right)^2\|y_n - x_n\|^2 - \frac{\alpha\beta\eta\mu}{1-\alpha}\left(f(x_n) - f(y_n) - \langle \nabla f(y_n), y_n - x_n\rangle\right) \end{aligned}$$

As for Theorem 3.3.2, choose: $\eta = \frac{\sqrt{s}}{\sqrt{\mu}}$, $\beta = 1 - \eta\mu = 1 - \sqrt{\mu s}$ and $\alpha = \frac{1}{1+\eta\mu} = \frac{1}{1+\sqrt{\mu s}}$ to cancel out the terms in $F(x_{n+1}) - F^*$, $\|y_n - x_F^*\|^2$, $F(x_n) - F^*$ and $\|y_n - x_{n+1}\|^2$. We then get:

$$E_{n+1} - E_n \leq -(1-\beta)E_n + \beta\left(f(y_n) + \langle \nabla f(y_n), x_n - y_n\rangle - f(x_n)\right) - \beta\frac{\sqrt{\mu}}{2\sqrt{s}}\|x_n - y_n\|^2. \tag{3.147}$$

finally, assuming additionally that $f$ is $(\rho, L)$-curvature-bounded for some $\rho \leq L$, observe that:

$$\forall n \in \mathbb{N}, f(y_n) + \langle \nabla f(y_n), x_n - y_n\rangle - f(x_n) \leq -\frac{\rho}{2}\|x_n - y_n\|^2,$$

which induces:

$$\forall n \in \mathbb{N},\ E_{n+1} - E_n \leq -(1-\beta)E_n - \frac{\beta}{2}\left(\rho + \frac{\sqrt{\mu}}{\sqrt{s}}\right)\|x_n - y_n\|^2$$

Provided that $\rho \geq -\sqrt{\frac{\mu}{s}}$, we finally obtain the expected inequality, namely: $E_{n+1} - E_n \leq -\sqrt{\mu s}E_n$ for all $n \in \mathbb{N}$, and we can conclude the proof in exactly the same way as for Theorem 3.3.2, Step 3. □

## 3.C Continuous analysis through High-resolution ODEs

**Derivation of ODE** Recall that the algorithm we prove convergence in Theorem 3.3.2 can be written

$$\begin{cases} y_n = x_n + \frac{1-\tau\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_n - x_{n-1}) + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\tau-1)(x_n - y_{n-1}) \\ x_{n+1} = y_n - s\nabla f(y_n) \end{cases} \tag{3.148}$$

Writing only with respect to $y_n$, we get

$$y_{n+1} = y_n + \frac{1-\tau\sqrt{\mu s}}{1+\sqrt{\mu s}}(y_n - y_{n-1}) - s\left(1 + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(\tau-1)\right)\nabla f(y_n) - s\frac{1-\tau\sqrt{\mu s}}{1+\sqrt{\mu s}}(\nabla f(y_n) - \nabla f(y_{n-1})) \tag{3.149}$$

The following development will be very close to the one introduced in (Shi et al., 2022). We assume there exists a smooth curve $X$ such that $X(t_n) = y_n$, where $t_n = n\sqrt{s}$. By Taylor expansion, we have

$$y_{n+1} = X(t_{n+1}) = X(t_n) + \sqrt{s}\dot{X}(t_n) + \frac{s}{2}\ddot{X}(t_n) + \frac{\sqrt{s}^3}{6}\dddot{X}(t_n) + \mathcal{O}(s^2) \tag{3.150}$$

$$y_{n-1} = X(t_{n-1}) = X(t_n) - \sqrt{s}\dot{X}(t_n) + \frac{s}{2}\ddot{X}(t_n) - \frac{\sqrt{s}^3}{6}\dddot{X}(t_n) + \mathcal{O}(s^2) \tag{3.151}$$

Another Taylor expansion gives

$$\nabla f(y_n) - \nabla f(y_{n-1}) = \nabla^2 f(X(t_n))\dot{X}(t_n)\sqrt{s} + \mathcal{O}(s) \tag{3.152}$$

Multiplying both sides of (3.149) by $\frac{1+\sqrt{\mu s}}{1-\tau\sqrt{\mu s}}\frac{1}{s}$, we get

$$\frac{y_{n+1} + y_{n-1} - 2y_n}{s} + \frac{(1+\tau)\sqrt{\mu s}}{1-\tau\sqrt{\mu s}}\frac{y_{n+1} - y_n}{s} + \nabla f(y_n) - \nabla f(y_{n-1}) + \frac{1+\tau\sqrt{\mu s}}{1-\tau\sqrt{\mu s}}\nabla f(y_{n-1}) = 0 \tag{3.153}$$

Using Taylor expansions above, we have

$$\ddot{X}(t_n) + \mathcal{O}(s) + \frac{(1+\tau)\sqrt{\mu}}{1-\tau\sqrt{\mu s}}\left[\dot{X}(t_n) + \frac{1}{2}\ddot{X}(t_n)\sqrt{s} + \mathcal{O}(s)\right] \tag{3.154}$$

$$+ \nabla^2 f(X(t_n))\dot{X}(t_n)\sqrt{s} + \mathcal{O}(s) + \left(\frac{1+\tau\sqrt{\mu s}}{1-\tau\sqrt{\mu s}}\right)\nabla f(X(t_n)) = 0 \tag{3.155}$$

Multiplying both sides by $1-\tau\sqrt{\mu s}$ and ignoring $\mathcal{O}(s)$ terms, we get that (3.149) is a discretization of:

$$(1 + \frac{1-\tau}{2}\sqrt{\mu s})\ddot{X}(t) + (1+\tau)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X} + (1+\tau\sqrt{\mu s})\nabla f(X(t)) = 0 \quad \text{(NM-SQC-ODE)}$$

**Theorem.** *Let $f$ be $C^2$, $(\tau,\mu)-$strongly quasar-convex and $L$-smooth. Assume $X$ is solution of (NM-SQC-ODE) with $0 \le s \le \frac{1}{L}$, $X(0) = X_0$ and $\dot{X}(0) = 0$. Then:*

$$f(X(t)) - f^* \le K_0(\tau,\mu,L,s)\frac{1}{\tau}(f(X_0) - f^*)e^{-\tau\frac{\sqrt{\mu}}{2}t} \tag{3.156}$$

*where $K_0(\tau,\mu,L,s) \le 7$.*

*Proof.* We rewrite the ODE (NM-SQC-ODE) the following way.

$$\upsilon\ddot{X}(t) + (1+\tau)\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1+\tau\sqrt{\mu s})\nabla f(X(t)) = 0 \tag{3.157}$$

Where $\upsilon = 1 + \frac{1-\tau}{2}\sqrt{\mu s}$. Set the following Lyapunov function:

$$\mathcal{E}(t) = \delta(f(X(t)) - f^*) + \frac{1}{2}\|\upsilon\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla f(X(t))\|^2 \tag{3.158}$$

To simplify the following computations, we write $X(t)$ as $X$, and we do the same for the first and second derivatives of $X$. We have

$$\dot{\mathcal{E}}(t) = \delta\langle \dot{X}, \nabla f(X)\rangle + \langle \upsilon\dot{X}(t) + \lambda(X(t) - x^*) + \sqrt{s}\nabla f(X(t)), \upsilon\ddot{X} + \lambda\dot{X} + \sqrt{s}\nabla^2 f(X)\dot{X}\rangle. \tag{3.159}$$

Injecting (NM-SQC-ODE), we get

$$\begin{aligned}
\dot{\mathcal{E}}(t) &= \delta\langle \dot{X}, \nabla f(X)\rangle + \langle \upsilon\dot{X}(t) + \lambda(X(t) - x^*) && (3.160)\\
&+ \sqrt{s}\nabla f(X(t)), (\lambda - (1+\tau)\sqrt{\mu})\sqrt{\mu}\dot{X} - (1+\tau\sqrt{\mu s})\nabla f(X)\rangle && (3.161)\\
&= \delta\langle \dot{X}, \nabla f(X)\rangle + \upsilon(\lambda - (1+\tau)\sqrt{\mu})(\|\dot{X}\|^2) && (3.162)\\
&- \upsilon(1+\tau\sqrt{\mu s})\langle \dot{X}, \nabla f(X)\rangle + \lambda(\lambda - (1+\tau)\sqrt{\mu})\langle X - x^*, \dot{X}\rangle && (3.163)\\
&- \lambda(1+\tau\sqrt{\mu s})\langle X - x^*, \nabla f(X)\rangle + (\lambda - (1+\tau)\sqrt{\mu})\sqrt{s}\langle \nabla f, \dot{X}\rangle - \sqrt{s}(1+\tau\sqrt{\mu s})\|\nabla f(X)\|^2 && (3.164)
\end{aligned}$$

We set $(\lambda - (1+\tau)\sqrt{\mu}) = -\upsilon\tau\sqrt{\mu}$. Then to cancel $\langle \dot{X}, \nabla f(X)\rangle$ terms, we set

$$\delta = (\upsilon(1+\tau\sqrt{\mu s}) + \upsilon\tau\sqrt{\mu s}) = \upsilon(1+2\tau\sqrt{\mu s}) \tag{3.165}$$

Then we get

$$\begin{aligned}
\dot{\mathcal{E}}(t) &\leq -\tau\sqrt{\mu}\left(\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda}{\tau\sqrt{\mu}}(1+\tau\sqrt{\mu s})\langle X - x^*, \nabla f(X)\rangle + s\|\nabla f(X)\|^2\right) && (3.166)\\
&- \sqrt{s}\|\nabla f(X)\|^2 && (3.167)
\end{aligned}$$

We use strong quasar convexity.

$$\begin{aligned}
\dot{\mathcal{E}}(t) &\leq -\tau\sqrt{\mu}\left(\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda}{\sqrt{\mu}}(1+\tau\sqrt{\mu s})(f(X) - f^*)\right. && (3.168)\\
&\left. + \frac{\lambda\sqrt{\mu}}{2}(1+\tau\sqrt{\mu s})\|X - x^*\|^2 + s\|\nabla f(X)\|^2\right) - \sqrt{s}\|\nabla f(X)\|^2 && (3.169)\\
&= -\tau\sqrt{\mu}\left(\frac{1}{2}\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda^2}{2}\|X - x^*\|^2\right) && (3.170)\\
&+ \frac{\upsilon}{2}(1+2\tau\sqrt{\mu s})(f(X) - f^*) + s\|\nabla f(X)\|^2 - \sqrt{s}\|\nabla f(X)\|^2\Big) && (3.171)\\
&- \tau\sqrt{\mu}\left(\frac{\lambda\sqrt{\mu}}{2}(1+\tau\sqrt{\mu s}) - \frac{\lambda^2}{2}\right)\|X - x^*\|^2 && (3.172)\\
&- \tau\sqrt{\mu}\left(\frac{\lambda}{\sqrt{\mu}}(1+\tau\sqrt{\mu s}) - \frac{\upsilon}{2}(1+2\tau\sqrt{\mu s})\right)(f(X) - f^*) - \tau\frac{\sqrt{\mu}\upsilon^2}{2}\|\dot{X}\|^2 && (3.173)
\end{aligned}$$

We have $\frac{1}{2}\upsilon^2\|\dot{X}\|^2 + \lambda\upsilon\langle X - x^*, \dot{X}\rangle + \frac{\lambda^2}{2}\|X - x^*\|^2 = \|\upsilon\dot{X} + \lambda(X - x^*)\|^2$. Then, we use:

$$\begin{aligned}
\frac{1}{2}\|\upsilon\dot{X} + \lambda(X - x^*) + \sqrt{s}\nabla f(X)\|^2 &\leq \|\upsilon\dot{X} + \lambda(X - x^*)\|^2 + s\|\nabla f(X)\|^2 && (3.174)\\
\Rightarrow -\frac{1}{4}\|\upsilon\dot{X} + \lambda(X - x^*) + \sqrt{s}\nabla f(X)\|^2 &\geq -\frac{1}{2}\|\upsilon\dot{X} + \lambda(X - x^*)\|^2 - \frac{s}{2}\|\nabla f(X)\|^2 && (3.175)
\end{aligned}$$

This leads to:

$$\begin{aligned}
\dot{\mathcal{E}}(t) &\leq -\tau\frac{\sqrt{\mu}}{2}\mathcal{E}(t) - \sqrt{s}\|\nabla f(X)\|^2 - \tau\sqrt{\mu}\left(\frac{\lambda\sqrt{\mu}}{2}(1+\tau\sqrt{\mu s}) - \frac{\lambda^2}{2}\right)\|X - x^*\|^2 && (3.176)\\
&- \tau\sqrt{\mu}\left(\frac{\lambda}{\sqrt{\mu}}(1+\tau\sqrt{\mu s}) - \frac{\upsilon}{2}(1+2\tau\sqrt{\mu s})\right)(f(X) - f^*) - \tau\frac{\sqrt{\mu}\upsilon^2}{2}\|\dot{X}\|^2 && (3.177)
\end{aligned}$$

We have to check that some terms are negatives. We have $\lambda = \sqrt{\mu}(1+\tau(1-\upsilon))$, and

$$\frac{\lambda\sqrt{\mu}}{2}(1+\tau\sqrt{\mu s}) - \frac{\lambda^2}{2} = \frac{\lambda}{2}\left(\sqrt{\mu}(1+\tau\sqrt{\mu s}) - \lambda\right) = \frac{\lambda\mu}{2}\left(\tau\sqrt{\mu s} - \tau(1-\upsilon)\right) = \frac{\lambda\tau\mu\sqrt{s}}{2}\left(\frac{3-\tau}{2}\right) \geq 0 \tag{3.178}$$

and

$$\frac{\lambda}{\sqrt{\mu}}(1+\tau\sqrt{\mu s})-\frac{\upsilon}{2}(1+2\tau\sqrt{\mu s})=\left(1-\frac{\tau(1-\tau)}{2}\sqrt{\mu s}\right)(1+\tau\sqrt{\mu s})-\frac{1}{2}(1+\frac{1-\tau}{2}\sqrt{\mu s})(1+2\tau\sqrt{\mu s}) \tag{3.179}$$

$$=1+\tau\sqrt{\mu s}-\frac{1}{2}(1+2\tau\sqrt{\mu s})-\frac{\tau(1-\tau)}{2}\sqrt{\mu s}(1+\tau\sqrt{\mu s}) \tag{3.180}$$

$$-\frac{1}{2}\frac{1-\tau}{2}\sqrt{\mu s}(1+2\tau\sqrt{\mu s}) \tag{3.181}$$

$$=\frac{1}{2}-\frac{1-\tau}{2}\sqrt{\mu s}\left(\tau(1+\tau\sqrt{\mu s})+\frac{1}{2}+\tau\sqrt{\mu s}\right) \tag{3.182}$$

We want to be sure that this quantity is positive. To do so, we will maximize the right term with respect to $\tau\in[0,1]$. First, note that supposing $s\leq\frac{1}{L}$, we have:

$$\frac{1-\tau}{2}\sqrt{\mu s}\left(\tau(1+\tau\sqrt{\mu s})+\frac{1}{2}+\tau\sqrt{\mu s}\right)\leq\frac{1-\tau}{2}\left(\tau(1+\tau)+\frac{1}{2}+\tau\right)=\frac{1}{4}(1-\tau)(1+4\tau+2\tau^2):=g(\tau) \tag{3.183}$$

We have

$$4g(\tau)=(1+3\tau-2\tau^2-2\tau^3) \tag{3.184}$$

We now want to find critical points of $g$.

$$4g'(\tau)=3-4\tau-6\tau^2 \tag{3.185}$$

We calculate the discriminant $\Delta=4^2+4*6*3=88$, inducing that the roots of $g'$ are

$$x_1=-\frac{4+2\sqrt{22}}{12},\quad x_2=\frac{-4+2\sqrt{22}}{12} \tag{3.186}$$

We clearly have $x_1<0$. $x_2$ however belongs to $[0,1]$. We evaluate numerically $g(x_2)\approx 0.44<\frac{1}{2}$. We conclude that for all $\tau\in[0,1]$ we have

$$\frac{1}{2}-\frac{1-\tau}{2}\sqrt{\mu s}\left(\tau(1+\tau\sqrt{\mu s})+\frac{1}{2}+\tau\sqrt{\mu s}\right)>0 \tag{3.187}$$

All the out of parenthesis terms are negative, so we conclude that:

$$\dot{\mathcal{E}}(t)\leq-\tau\frac{\sqrt{\mu}}{2}\mathcal{E}(t)\Rightarrow\mathcal{E}(t)\leq\mathcal{E}(0)e^{-\tau\frac{\sqrt{\mu}}{2}t} \tag{3.188}$$

**Deducing rate on $f(X(t))-f^*$** Using initial conditions, we have

$$\mathcal{E}(0)=\delta(f(X_0)-f^*)+\frac{1}{2}\|\lambda(X_0-x^*)-\sqrt{s}\nabla f(X_0)\|^2 \tag{3.189}$$

$$\leq\delta(f(X_0)-f^*)+\lambda^2\|X_0-x^*\|^2+s\|\nabla f(X_0)\|^2 \tag{3.190}$$

$$\leq\left(\delta+\frac{2\lambda^2(2-\tau)}{\tau\mu}+2Ls\right)(f(X_0)-f^*) \tag{3.191}$$

Where the third inequality uses that $f$ is $\frac{\mu\tau}{2-\tau}$-quadratic growth (Proposition 3.2.1) and that $f$ is $L$-Smooth. Then, we have

$$f(X(t))-f^*\leq\underbrace{\left(\tau+\frac{2\lambda^2(2-\tau)}{\mu\delta}+2\tau Ls\right)}_{:=K_0(\tau,\mu,L,s)}\frac{1}{\tau}(f(X_0)-f^*)e^{-\tau\frac{\sqrt{\mu}}{2}t} \tag{3.192}$$

We need to check that $K_0(\tau,\mu,L,s)$ is uniformly bounded. Note first that as $0<s\leq\frac{1}{L}$, we have

$$sL\leq 1,\quad \mu s\leq 1 \tag{3.193}$$

We now bound $\upsilon, \delta, \lambda$, that we already fixed in the proof.

$$1 \leq \upsilon := 1 + \frac{1-\tau}{2}\sqrt{\mu s} \leq \frac{3}{2} \tag{3.194}$$

$$1 \leq \delta := \upsilon(1 + 2\tau\sqrt{\mu s}) \leq \frac{9}{2} \tag{3.195}$$

$$\sqrt{\mu}\left(1 - \frac{1}{8}\right) \leq \lambda := \sqrt{\mu}\left(1 + \frac{\tau(\tau-1)}{2}\sqrt{\mu s}\right) \leq \sqrt{\mu} \tag{3.196}$$

In the last inequality, we used the well known fact that $0 \leq p(1-p) \leq \frac{1}{4}$, for all $p \in [0,1]$.
We thus can explicitly compute that

$$K_0(\tau, \mu, L, s) \leq (1 + 4 + 2) = 7. \tag{3.197}$$

□

# Chapter 4

# Continuized Nesterov Acceleration for Non-Convex Optimization: Application to Strongly Quasar Convex Functions

We recall that the content of this chapter is based on the following paper.

J. Hermant, J-F. Aujol, C. Dossal, L. Huang and A. Rondepierre.
**Continuized Nesterov Acceleration for Non-Convex Optimization**.
*arXiv preprint*, 2025

## 4.1 Introduction

We recall that we consider the following Nesterov Momentum algorithm (Nesterov, 2013)

$$\left\{ \begin{array}{ll} \tilde{y}_k & = \alpha_k \tilde{x}_k + (1-\alpha_k)\tilde{z}_k \\ \tilde{x}_{k+1} & = \tilde{y}_k - s\nabla f(\tilde{y}_k) \\ \tilde{z}_{k+1} & = \beta_k \tilde{z}_k + (1-\beta_k)\tilde{y}_k - \eta_k \nabla f(\tilde{y}_k) \end{array} \right. \tag{4.1}$$

In Section 1.2, we discussed that a common approach to studying optimization algorithms is to consider their continuous-time counterparts. By letting the stepsize $s$ in (4.1) tend to zero, the continuous-time limit of the above algorithm is

$$\left\{ \begin{array}{ll} \dot{x}_t & = \eta_t(z_t - x_t) - \gamma_t \nabla f(x_t) \\ \dot{z}_t & = \eta'_t(x_t - z_t) - \gamma'_t \nabla f(x_t) \end{array} \right. \tag{4.2}$$

with $t \in \mathbb{R}_+$ for some choices of parameters $\eta_t, \eta'_t, \gamma_t$ and $\gamma'_t$, see the derivation in (Kim and Yang, 2023, Appendix A.1). We discussed that this continuous time object can be used to find Lyapunov functions, namely a suitable function $\varphi : \mathbb{R}^d \to \mathbb{R}$, typically depending on the function we want to minimize, such that the Lyapunov function $E_t := \varphi(x_t, z_t, t)$ will decrease with $t$. A reason is that working in continuous time is often more convenient, since one can analyze the derivative $\dot{E}_t$. The computations carried out in continuous time can serve as an inspiration for studying the corresponding algorithm. Importantly, in Section 1.2, we stated some results in the literature where discrepancies arise between results obtained for the ODE and results obtained for the algorithm, specifically when considering non-convex functions. In Chapter 3, Section 3.4.2, we gave a new instance of such discrepancies, which is the case of strongly quasar convex functions. We observed how non-convexity is precisely what causes discretization discrepancies.

### 4.1.1 A Promising Alternative: Continuized Nesterov

A modification of (4.2) is introduced in (Even et al., 2021), motivated by the study of asynchronous algorithms, and drawing on ideas originating from the analysis of Markov chains (Aldous and Fill, 2002).

Intuitively, instead of making the gradient act continuously in $t \in \mathbb{R}_+$, gradient steps will occur at discrete random times $T_1, T_2, \dots$ where for all $i$, $T_{i+1} - T_i$ are independent and follow an exponential law of parameter 1. Precisely, between these random times, the dynamics of $(x_t, z_t)_{t\geq 0}$ smoothly interpolate the two processes, namely

$$\begin{cases} dx_t &= \eta_t(z_t - x_t)dt \\ dz_t &= \eta'_t(x_t - z_t)dt \end{cases} \tag{4.3}$$

At $t = T_i$, the process *jumps* by performing a gradient step

$$x_{T_i} = x_{T_i^-} - \gamma_{T_i^-}\nabla f(x_{T_i^-}), \quad z_{T_i} = z_{T_i^-} - \gamma'_{T_i^-}\nabla f(x_{T_i^-}), \tag{4.4}$$

where $x_{T_i^-}$ denotes the left-limit $\lim_{t\to 0^-} x_{T_i+t}$. This process can be written as a stochastic differential equation that we call the *continuized Nesterov equations*

$$\begin{cases} dx_t &= \eta_t(z_t - x_t)dt - \gamma_{t^-}\nabla f(x_{t^-})N(dt) \\ dz_t &= \eta'_t(x_t - z_t)dt - \gamma'_{t^-}\nabla f(x_{t^-})N(dt) \end{cases} \tag{4.5}$$

where $N(dt) = \sum_{k\geq 0}\delta_{T_k}(dt)$ is a Poisson point measure. Intuitively, since $\mathbb{E}[dN_t] = dt$, the mean process $(\mathbb{E}[x_t], \mathbb{E}[z_t])_{t\geq 0}$ is smooth, see Figure 4.1. In particular, if $\nabla f$ is linear, it can be written exactly in the form of (4.2). Consequently, individual realizations of the jump-driven process (4.5) provide a stochastic approximation of a much smoother expected trajectory. It implies that, for a given Lyapunov function $E_t := \varphi(x_t, z_t, t)$, the analysis of $\mathbb{E}[E_t]$ is close to that arising in the standard continuous setting, such as in the study of (4.2).

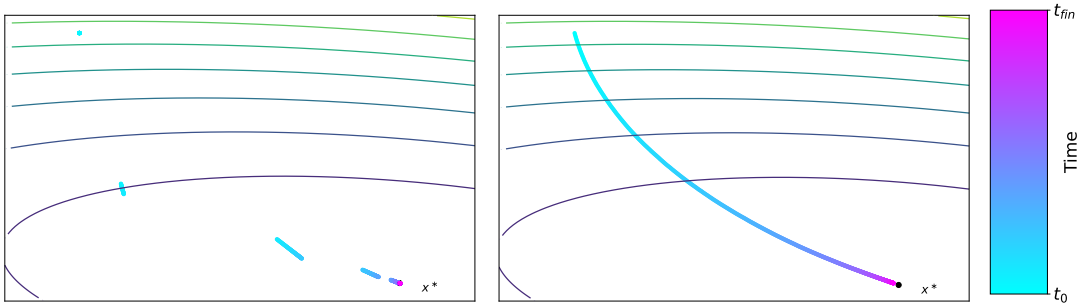



Figure 4.1: Left: Visualization of a single trajectory of $(x_t)_{t\geq 0}$ satisfying (4.5), between $t_0$ and $t_{\text{fin}}$. Right: Visualization of the average over $10^5$ such trajectories, providing an estimate of the mean process $(\mathbb{E}[x_t])_{t\geq 0}$ for which the jumps are smoothed out.

Strikingly, a core property of (4.5) is that the discretization $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$, $k \in \mathbb{N}$, can be **computed exactly**. The resulting discretization writes as a Nesterov momentum algorithm of the form (4.1), with stochastic parameters involving the random times $\{T_k\}_{k\geq 0}$. The convergence results obtained for the solution of (4.5) thus almost **automatically transfer** to the aforementioned discretization. It totally differs from the classic ODE framework, where (4.1) is a discrete **approximation** of (4.2). In their seminal work, (Even et al., 2021) show that applied to (strongly) convex functions, the continuized version of the Nesterov momentum algorithm recovers, in expectation, similar results to the classical Nesterov algorithm. Their analysis has been extended to strongly quasar convex functions (Wang and Wibisono, 2023). If this method has known some extension in the specific field of decentralized asynchronous algorithms (Nabli and Oyallon, 2025; Nabli et al., 2023), its use in a more classic optimization setting remains limited.

### 4.1.2 Limits of the Framework and Contributions

Compared with the classical ODE framework, a key difference is that the continuized approach yields an exact discretization whose convergence properties inherit almost directly those of the continuous-time process. Because of this conceptual difference, the continuized method may at first feel unfamiliar or non-intuitive to optimizers, which might explain why it is still largely unexploited. Moreover, it has some limitations:

first, because of the inherent stochasticity of the continuized process, most of the existing results hold in expectation even if using the exact gradient. This leads to results that are weaker in nature than those of classical momentum algorithms, which provide deterministic guarantees. Results holding trajectory-wise with a given probability are given in (Wang and Wibisono, 2023), but the derived probability can be trivial for a large amount of the first iterations (Section 4.4.2). Also, the current framework only allows to use smooth Lyapunov functions, preventing for example its application to non-smooth functions.

**Contributions** (i) We present a clear methodology for employing the continuized method and extend its fundamental tools. We illustrate the method on the class of smooth Polyak-Łojasiewicz functions, deducing a new, though non-accelerated, convergence result, see Section 4.3.

(ii) We provide an Itô formula (Proposition 4.5.1) that allows to handle non-differentiable Lyapunov functions. In the case of exponentially fast convergence rates, we significantly tighten the existing bound on the probability to have a trajectory-wise convergence result, and further derive asymptotic almost sure convergence rates.

(iii) We exploit the method to address two limitations raised in Section 1.1.2. Adapting results coming from the classic ODE framework literature, we improve the best known convergence rate by a constant factor for smooth strongly quasar convex functions. We further relax the existing assumption on the set of minimizers for this class of functions. Our results hold when considering unbiased stochastic estimates of the gradients, as long as a strong growth condition is verified.

## 4.2 Assumptions and Notations

### 4.2.1 Geometrical Assumptions

We state the geometrical assumptions used throughout the chapter. For discussions about these assumptions, see Section 1.1.2, Section 1.1.1 or Chapter 3.

**Assumption 4.2.1** (L-smooth)**.** *$f$ is such that $\forall x, y \in \mathbb{R}^d,\ f(x) - f(y) - \langle \nabla f(y), x - y\rangle \leq \frac{L}{2}\|x-y\|^2$, for some $L > 0$.*

**Assumption 4.2.2** ($\mathrm{PL}_\mu$)**.** *$f$ is such that $\forall x \in \mathbb{R}^d$, $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$, for some $\mu > 0$.*

We recall gradient descent is known to ensure linear convergence to the minimum for functions satisfying Assumptions (L-smooth) and ($\mathrm{PL}_\mu$), with a rate of order $\mathcal{O}((1-\mu/L)^k)$ after $k$ iterations (Bolte et al., 2017; Karimi et al., 2016). We also recall that momentum algorithms cannot significantly improve over this bound (Yue et al., 2023), in the sense that the dependence on $\mu/L$ cannot be improved to $\sqrt{\mu/L}$, as in the strongly convex setting.

**Assumption 4.2.3** ($\mathrm{SQC}_{\tau,\mu}$)**.** *$f$ verifies $f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2$ for all $x \in \mathbb{R}^d$, and some $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$.*

We recall that functions satisfying Assumptions (L-smooth) and ($\mathrm{SQC}_{\tau,\mu}$) achieve similar rates as strongly convex functions, namely $\mathcal{O}((1-\tau\mu/L)^k)$ for gradient descent and $\mathcal{O}((1-\tau\sqrt{\mu/L})^k)$ when using momentum, see Section 4.4. As discussed in Section 1.1.2, a limitation of this class is the unique minimizer property induced by Assumption ($\mathrm{SQC}_{\tau,\mu}$) (Hinder et al., 2020, Observation 4), which prevents several machine-learning problems from fitting directly. We will further relax this uniqueness property in Section 4.5.

### 4.2.2 Stochastic Gradients

We consider using stochastic estimators of the gradient $\nabla f$.

**Assumption 4.2.4** ($\mathrm{SGC}_\rho$)**.** *We assume we have access to stochastic approximations $\nabla f(x, \xi)$ of the real gradient $\nabla f(x)$, where $\xi \sim \mathcal{P}(\Xi)$, such that*

*(i) (Unbiased estimator) $\forall x \in \mathbb{R}^d,\ \mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla f(x)$.*

*(ii) (Strong Growth Condition) For $\rho \geq 1, \forall x \in \mathbb{R}^d,\ \mathbb{E}\left[\|\nabla f(x, \xi)\|^2\right] \leq \rho\|\nabla f(x)\|^2$.*

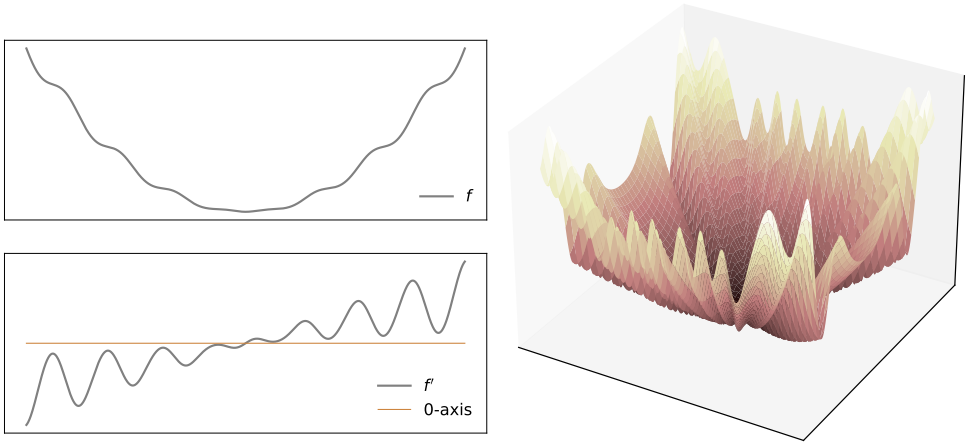



Figure 4.2: Visualization of functions verifying Assumption ($\mathrm{SQC}_{\tau,\mu}$). Left: 1-dimensional example $f(t) = 0.5 \cdot (t + 0.15\sin(5t))^2$ on top, its derivative on the bottom. Right: 2-dimensional example of the form $h(x,y) = f(\|(x,y)\|)g\left(\frac{(x,y)}{\|(x,y)\|}\right)$, where $f$ verifies Assumption ($\mathrm{SQC}_{\tau,\mu}$) and $g \geq 1$. It clearly appears in both cases that this class of functions permits high oscillations.

We provided extensive discussions on this assumption in Section 1.1.3 and Chapter 2. We will simply emphasize that from an optimization point of view, under Assumption (L-smooth), one can deduce a stochastic version of the descent lemma, whose proof is deferred in Appendix 4.C.1.

**Lemma 4.2.1.** *Under Assumption (L-smooth) and ($SGC_\rho$), if $\gamma \leq \frac{1}{L\rho}$, we have*

$$\forall x \in \mathbb{R}^d,\ \mathbb{E}\left[f(x - \gamma\nabla f(x,\xi)) - f(x)\right] \leq \gamma\left(\frac{L\rho}{2}\gamma - 1\right)\|\nabla f(x)\|^2.$$

Lemma 4.2.1 bounds, in expectation, the decrease of the function $f$ when performing a stochastic gradient step. For deterministic gradient, *i.e.* $\rho = 1$, we recover the classical descent lemma (Nesterov, 2013, eq. (1.2.19)).

## 4.3 How to Use the Continuized Nesterov Framework

In this section, we detail the methodology for leveraging the continuized Nesterov dynamics to derive convergence results. It can be outlined through the three following steps: *(i)* continuous-time analysis via stochastic Lyapunov functions, *(ii)* derivation of an exact discrete algorithm, and *(iii)* automatic transfer of convergence guarantees to this algorithm, using stopping-time arguments.

First, we note that the continuized equations naturally handle stochastic gradients. Following the discussion of Section 4.1.1, consider the same random times $\{T_k\}_{k\in\mathbb{N}}$. Assume $\xi_1, \xi_2, \ldots$ are i.i.d random variables of law $\mathcal{P}$. At $t = T_k$, a stochastic gradient step is performed

$$x_{T_k} = x_{T_k^-} - \gamma\nabla f(x_{T_k^-}, \xi_k), \quad z_{T_k} = z_{T_k^-} - \gamma'\nabla f(x_{T_k^-}, \xi_k).$$

Between these random times, the dynamics followed by $(x_t, z_t)_{t\geq 0}$ is still

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt \\ dz_t &= \eta'(x_t - z_t)dt \end{cases}$$

These *continuized Nesterov equations* write as ((15)-(16) in (Even et al., 2021))

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma\int_\Xi \nabla f(x_{t^-}, \xi)N(dt, d\xi) \\ dz_t &= \eta'(x_t - z_t)dt - \gamma'\int_\Xi \nabla f(x_{t^-}, \xi)N(dt, d\xi) \end{cases} \tag{CNE}$$

with the Poisson point measure $N(dt, d\xi) = \sum_{k\geq 0} \delta_{(T_k,\xi_k)}(dt, d\xi)$ on $\mathbb{R}_+ \times \Xi$, which has intensity $dt \otimes \mathcal{P}$. By definition, the process is càdlàg, *i.e.* right-continuous with left limits. We assume $x_0 = z_0$ is deterministic. Note that compared to the discussion of Section 4.1.1, we consider a constant set of parameters $(\eta, \eta', \gamma, \gamma')$.

**Remark 4.3.1.** *Because gradient steps intervene at discrete times, the continuized Nesterov equations are highly flexible to modifications. For instance, we could consider using proximal updates (Rockafellar, 1976) in a composite setting where we minimize $F = f+g$ with $g$ potentially non-smooth, in a FISTA-style system Beck and Teboulle (2009); or using different stochastic estimators of the gradients, e.g. using variance reduction techniques (Johnson and Zhang, 2013). Although we believe these extensions are interesting directions for future research, they fall outside the scope of the present work.*

Our main interest is to analyze the convergence properties of $(x_t, z_t)_{t\geq 0}$. In the following section, we provide an analytical tool that replaces differentiation in the presence of jumps.

### 4.3.1 Lyapunov Analysis with Stochastic Calculus

To deduce convergence results using (CNE), a classical strategy is to consider a well chosen Lyapunov function, *e.g.* $t \to \mathbb{E}[f(x_t)]$, and to show it is decreasing. Although $(x_t)_{t\geq 0}$ is indexed by continuous time, it cannot be differentiated with respect to time because of the jumps introduced by the Poisson component. Fortunately, the theory of stochastic calculus provides analogous tools, such that we can use a Lyapunov-style approach.

**Proposition 4.3.1** ($C^1$-Itô Formula)**.** *Let $x_t \in \mathbb{R}^d$ be a solution of*

$$dx_t = \zeta(x_t)dt + \int_{\Xi} G(x_{t^-}, \xi)N(dt, d\xi)$$

*where $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ is a locally Lipschitz function and $G : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ is measurable. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be with Lipschitz gradient. Then,*

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla\varphi(x_s), \zeta(x_s)\rangle \, ds + \int_{[0,t]} \mathbb{E}_\xi[\varphi(x_s + G(x_s, \xi)) - \varphi(x_s)]ds + M_t, \tag{4.6}$$

*where $M_t$ is a martingale such that $\mathbb{E}[M_t] = 0$, $\forall t \geq 0$.*

This Itô formula is a close formulation of (Even et al., 2021, Proposition 2). For readers who are not familiar with martingales or stochastic differential equations, standard references include (Williams, 1991; Protter, 2012). Proposition 4.3.1 can be applied to a well chosen Lyapunov function $\varphi$. In particular, it suffices to show that $\langle \nabla\varphi(x_s), \zeta(x_s)\rangle + \mathbb{E}_\xi\varphi(x_s + G(x_s, \xi)) - \varphi(x_s) \leq 0$ to have that $t \to \varphi(x_t)$ is a supermartingale, which in turn induces

$$\mathbb{E}[\varphi(x_t)] \leq \mathbb{E}[\varphi(x_0)].$$

We further relax this $C^1$ condition in Proposition 4.5.1.

**Remark 4.3.2.** *The $\int_0^t \langle \nabla\varphi(x_s), \zeta(x_s)\rangle \, ds$ term appearing in (4.6) is purely of continuous nature. Computations involving this term will be similar to the one involving the classic deterministic system (4.2). The term $\int_{[0,t]} \mathbb{E}_\xi[\varphi(x_s + G(x_s, \xi)) - \varphi(x_s)]ds$ is induced by the Poisson process, and will involve some terms leading to computations that are closer to the one occurring when analyzing (4.1). As mentioned in (Even et al., 2021), the continuized framework involves a continuous-discrete mix.*

### 4.3.2 Almost Automatic Discretization

Because the solution to (4.2) requires knowledge of the gradient at all times $t \in \mathbb{R}_+$, it is generally not available in closed form. Thus, convergence guarantees for the continuous system cannot be applied directly in practice, and one typically proceeds by establishing a corresponding result for a discretization, *e.g.* (4.1). A remarkable property of (CNE) is that the gradient acts only at discrete times. Defining $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$ as evaluations of the process $(x_t, z_t)_{t\geq 0}$ defined by (CNE), one can explicitly integrate the trajectory between consecutive times $\{T_k\}_{k\in\mathbb{N}}$.

**Proposition 4.3.2.** *Let $(x_t, z_t)_{t\geq 0}$ follow* (CNE) *with $\eta + \eta' > 0$ and with underlying jumping times $\{T_k\}_{k\in\mathbb{N}^*}$. Define $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_{k+1} := x_{T_{k+1}}$ and $\tilde{z}_{k+1} := z_{T_{k+1}}$ as evaluations of this process. Then, $(\tilde{y}_k, \tilde{x}_k, \tilde{z}_k)$ writes as a Nesterov algorithm in the form of* (4.1) *with stochastic parameters*

$$\left\{\begin{array}{ll} \tilde{y}_k & = (\tilde{z}_k - \tilde{x}_k)\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + \tilde{x}_k \\ \tilde{x}_{k+1} & = \tilde{y}_k - \gamma\nabla f(\tilde{y}_k, \xi_{k+1}) \\ \tilde{z}_{k+1} & = \tilde{z}_k + \eta'\frac{\left(1-e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta'+\eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(\tilde{y}_k - \tilde{z}_k) - \gamma'\nabla f(\tilde{y}_k, \xi_{k+1}) \end{array}\right. \tag{CNA}$$

Proposition 4.3.2 is a generalization of (Even et al., 2021; Wang and Wibisono, 2023) in the case of general constant parameters, see its proof in Appendix 4.A. Note that the random times $\{T_k\}_{k\in\mathbb{N}}$ can be easily simulated, such that the algorithm is implementable. We also note that $\eta' + \eta e^{-(\eta+\eta')(T_{k+1}-T_k)} \neq 0$, almost surely, which avoids undefined expressions in practice. We refer to (CNA) as the *continuized Nesterov algorithm*. Compared to the classic ODE framework, (CNA) is not an approximation of a continuous counterpart, it is an exact evaluation of a realization of (CNE). As $\tilde{x}_k$ is exactly $x_t$ evaluated at the stopping time $T_k$, one can transfer a convergence result to the sequence $\{\tilde{x}_k\}_{k\in\mathbb{N}}$, which we formalize as follows.

**Theorem 4.3.1** (Stopping theorem)**.** *Let $(\varphi_t)_{t\in\mathbb{R}_+}$ be a non-negative process with càdlàg trajectories, such that it verifies*

$$\varphi_t \leq K_0 + M_t,$$

*for some positive random variable $K_0$, some martingale $(M_t)_{t\geq 0}$ with $\mathbb{E}[M_0] = 0$. Then, for any almost surely finite stopping time $\tau$, one has*

$$\mathbb{E}[\varphi_\tau] \leq \mathbb{E}[K_0].$$

*Proof.* Let $\tau$ be an almost surely finite time, *i.e.* $\tau(\omega) < +\infty$ almost surely. The stopped martingale $\{M_{t\wedge\tau}\}_{t\in\mathbb{R}_+}$ is a martingale, with $\mathbb{E}[M_{t\wedge\tau}] = \mathbb{E}[M_{0\wedge\tau}] = 0$ for all $t \in \mathbb{R}_+$. Then, the stopped process $V_t := \varphi_{\tau\wedge t}$ is such that

$$\forall t \geq 0,\ \mathbb{E}[V_t] \leq \mathbb{E}[K_0].$$

As we assume $\inf_t \varphi_t \geq 0$, we also have $\inf_t V_t \geq 0$. We thus apply Fatou's Lemma

$$\mathbb{E}\left[\liminf_{t\to+\infty} V_t\right] \leq \liminf_{t\to+\infty} \mathbb{E}[V_t] \leq \mathbb{E}[K_0]. \tag{4.7}$$

$\tau$ being almost surely finite implies that $\lim_{t\to+\infty} t\wedge\tau(\omega) = \tau(\omega)$, almost surely. Then, for almost all $\omega \in \Omega$, we have

$$\lim_{t\to+\infty} V_t(\omega) = \lim_{t\to+\infty} \varphi_{t\wedge\tau(\omega)}(\omega) = \varphi_{\tau(\omega)}(\omega), \tag{4.8}$$

namely $V_t$ converges almost surely to $\varphi_\tau$. Plugging (4.8) into (4.7), we obtain

$$\mathbb{E}[\varphi_\tau] \leq \mathbb{E}[K_0].$$

□

The stopping theorem used in (Even et al., 2021) assumes the process is a uniformly integrable supermartingale, which might be difficult to verify. We state a result that holds for a broader range of $\{\varphi_t\}_{t\in\mathbb{R}_+}$, and is more convenient to use. Finally, we note that this automatic transfer from the continuous to the discrete setting comes at the cost of yielding non-deterministic convergence results, even in the presence of deterministic gradients.

### 4.3.3 Application to Smooth, Polyak-Łojasiewicz Functions

We illustrate the previously presented methodology in the context of minimizing $L$-smooth, $\mu$-Polyak-Łojasiewicz functions. We start by deriving a convergence result associated to the continuous process $(x_t)_{t\geq 0}$, using the Itô formula introduced in Section 4.3.1.

**Proposition 4.3.3.** *If $(x_t, z_t)_{t\geq 0}$ is a solution of (CNE) with $\gamma \leq \frac{1}{\rho L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2\rho}}$, $\eta = \sqrt{\frac{\gamma}{2\rho}}$ and $\eta' = \frac{\mu\gamma}{2} - \sqrt{\frac{\gamma}{2\rho}}$ under Assumptions (L-smooth), ($PL_\mu$) and ($SGC_\rho$), we have*

$$\mathbb{E}\left[f(x_t) - f^*\right] \leq e^{-\frac{\mu\gamma}{2}t}(f(x_0) - f^*).$$

*Proof.* **Formalism Setting** We start to set the formalism such that we can apply Proposition 4.3.1. Let $\overline{x}_t = (t, x_t, z_t)$ where $(x_t, z_t)_{t\geq 0}$ satisfy (CNE). It satisfies $d\overline{x}_t = \zeta(\overline{x}_t)dt + G(\overline{x}_t, \xi)N(dt, d\xi)$, with

$$\zeta(\overline{x}_t) = \begin{pmatrix} 1 \\ \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\overline{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma\nabla f(x_t, \xi) \\ -\gamma'\nabla f(x_t, \xi). \end{pmatrix}$$

We apply Proposition 4.3.1 to $\varphi(\overline{x}_t)$, where

$$\varphi(t, x, z) = a_t(f(x) - f^*) + \frac{b_t}{2}\left\|x - z\right\|^2, \tag{4.9}$$

such that for a martingale $M_t$, we have

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \langle\nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s, \xi)) - \varphi(\overline{x}_s)]ds + M_t. \tag{4.10}$$

**Computations** We compute $\langle\nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle$ and $\mathbb{E}_\xi\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)$. We have

$$\frac{\partial\varphi}{\partial s} = \frac{da_s}{ds}(f(x) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|x - z\right\|^2, \quad \frac{\partial\varphi}{\partial x} = a_s\nabla f(x) + b_s(x - z), \quad \frac{\partial\varphi}{\partial z} = b_s(z - x).$$

So,

$$\begin{aligned}
&\langle\nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle \\
&= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|x_s - z_s\right\|^2 \\
&+ \langle a_s\nabla f(x_s) + b_s(x_s - z_s), \eta(z_s - x_s)\rangle + \langle b_s(z_s - x_s), \eta'(x_s - z_s)\rangle \\
&= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|x_s - z_s\right\|^2 + a_s\eta\left\langle\nabla f(x_s), z_s - x_s\right\rangle - b_s\left(\eta + \eta'\right)\left\|x_s - z_s\right\|^2.
\end{aligned} \tag{4.11}$$

Also,

$$\begin{aligned}
&\mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)] \\
&= a_s\mathbb{E}_\xi[f(x_s - \gamma\nabla f(x_s, \xi)) - f(x_s)] \\
&+ \frac{b_s}{2}\mathbb{E}_\xi[\Big(\left\|x_s - \gamma\nabla f(x_s, \xi) - (z_s - \gamma'\nabla f(x_s, \xi)\right\|^2 - \left\|x_s - z_s\right\|^2\Big)] \\
&= a_s\mathbb{E}_\xi[f(x_s - \gamma\nabla f(x_s, \xi)) - f(x_s)] + b_s\frac{(\gamma' - \gamma)^2}{2}\mathbb{E}_\xi[\left\|\nabla f(x_s, \xi)\right\|^2] \\
&+ b_s(\gamma' - \gamma)\left\langle\mathbb{E}_\xi[\nabla f(x_s, \xi)], x_s - z_s\right\rangle \\
&\leq \left(b_s\rho(\gamma' - \gamma)^2 - a_s\gamma(2 - L\gamma\rho)\right)\frac{1}{2}\left\|\nabla f(x_s)\right\|^2 + b_s(\gamma' - \gamma)\left\langle\nabla f(x_s), x_s - z_s\right\rangle.
\end{aligned} \tag{4.12}$$

The last inequality uses Assumption ($SGC_\rho$) and Lemma 4.2.1. We combine (4.11) and (4.12)

$$\begin{aligned}
&\langle\nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)] \\
&\leq \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|x_s - z_s\right\|^2 + (a_s\eta - b_s(\gamma' - \gamma))\left\langle\nabla f(x_s), z_s - x_s\right\rangle \\
&- b_s(\eta + \eta')\left\|x_s - z_s\right\|^2 + \left(b_s\rho(\gamma' - \gamma)^2 - a_s\gamma(2 - L\gamma\rho)\right)\frac{1}{2}\left\|\nabla f(x_s)\right\|^2.
\end{aligned} \tag{4.13}$$

Now, using Assumption ($\mathrm{PL}_\mu$) in (4.13) and rearranging, we deduce

$$\begin{aligned}
&\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)] \\
&\leq \left(\frac{da_s}{ds} + \mu\left(b_s\rho(\gamma'-\gamma)^2 - a_s\gamma(2-L\gamma\rho)\right)\right)(f(x_s) - f^*) \\
&+ \left(\frac{1}{2}\frac{db_s}{ds} - b_s(\eta+\eta')\right)\|x_s - z_s\|^2 + (a_s\eta - b_s(\gamma'-\gamma))\langle \nabla f(x_s), z_s - x_s\rangle .
\end{aligned} \tag{4.14}$$

**Parameter Tuning** As it is standard in many convergence proofs, we now choose the parameters driving equation (CNE) such that we get to the desired result. We set $a_s = b_s = e^{\theta s}$, for some $\theta > 0$. Fixing $\gamma' = \gamma + \sqrt{\frac{\gamma}{2\rho}}$ ensures

$$\rho(\gamma'-\gamma)^2 - \gamma(2 - L\gamma\rho) \leq \frac{\gamma}{2} - \gamma = -\frac{\gamma}{2}.$$

The latter holds because $\gamma \leq \frac{1}{\rho L} \Rightarrow -\gamma(2 - L\rho\gamma) \leq -\gamma$. Setting $\eta = \gamma' - \gamma = \sqrt{\frac{\gamma}{2\rho}}$ cancels the scalar product in (4.14). We thus have

$$\begin{aligned}
&\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \mathbb{E}_\xi\left[\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)\right] \\
&\leq e^{\theta s}\left(\frac{\theta}{2} - (\eta+\eta')\right)\|x_s - z_s\|^2 + e^{\theta s}\left(\theta - \mu\frac{\gamma}{2}\right)(f(x_s) - f^*).
\end{aligned} \tag{4.15}$$

Fixing $\theta = \frac{\mu\gamma}{2}$ and $\eta' = \frac{\mu\gamma}{2} - \eta = \frac{\mu\gamma}{2} - \sqrt{\frac{\gamma}{2\rho}}$ cancels the right hand side of (4.15).

**Conclusion** From (4.10) and the previous computations, we get

$$\varphi(\overline{x}_t) \leq \varphi(\overline{x}_0) + M_t. \tag{4.16}$$

Using the definition of $\varphi$, $x_0 = z_0$ and ignoring the $e^{\frac{\mu\gamma}{2}t}\|x_t - z_t\|^2$ term, we get

$$e^{\frac{\mu\gamma}{2}t}(f(x_t) - f^*) \leq (f(x_0) - f^*) + M_t. \tag{4.17}$$

$M_t$ is a martingale with $\mathbb{E}[M_0] = 0$, such that $\mathbb{E}[M_t] = \mathbb{E}[M_0] = 0$, see Proposition 4.3.1. So, together with the fact that $x_0$ is assumed to be deterministic, taking expectation in (4.17), we obtain

$$\mathbb{E}[f(x_t) - f^*] \leq e^{-\frac{\mu\gamma}{2}t}(f(x_0) - f^*).$$

□

The only technical tool we use is a stochastic descent Lemma which controls the decrease of the function when performing a gradient step, see Lemma 4.2.1. Note that such descent lemmas typically arise in discrete analyses, rather than in the study of continuous processes. It is used in our proof to handle a term induced by the discrete component of the process. Before commenting on the result of Proposition 4.3.3, we leverage the tools introduced in Section 4.3.2 to transfer its result, applied to $(x_t)_{t\geq 0}$ in (CNE), to a result on $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ in (CNA).

**Theorem 4.3.2.** *Assume $T_1, \ldots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption (L-smooth), ($PL_\mu$) and ($SGC_\rho$), iterations of (CNA) with $\gamma \leq \frac{1}{\rho L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2\rho}}$, $\eta = \sqrt{\frac{\gamma}{2\rho}}$ and $\eta' = \frac{\mu\gamma}{2} - \sqrt{\frac{\gamma}{2\rho}}$ ensure*

$$\mathbb{E}\left[e^{\frac{\mu\gamma}{2}T_k}(f(x_{T_k}) - f^*)\right] \leq f(x_0) - f^*.$$

*Proof.* To prove Proposition 4.3.3, we established that for $(x_t, z_t)_{t\geq 0}$ solution of (CNE) with $\gamma \leq \frac{1}{\rho L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2\rho}}$, $\eta = \sqrt{\frac{\gamma}{2\rho}}$ and $\eta' = \frac{\mu\gamma}{2} - \sqrt{\frac{\gamma}{2\rho}}$, we have (4.17), namely

$$e^{\frac{\mu\gamma}{2}t}(f(x_t) - f^*) \leq f(x_0) - f^* + M_t.$$

Applying Theorem 4.3.1 with $\varphi_t := e^{\frac{\mu\gamma}{2}t}(f(x_t) - f^*)$, $K_0 := f(x_0) - f^*$ and with the almost surely finite stopping time $T_k$ yields

$$\mathbb{E}\left[e^{\frac{\mu\gamma}{2}T_k}(f(x_{T_k}) - f^*)\right] \leq f(x_0) - f^*.$$

We conclude using Proposition 4.3.2, which states that if $(x_t, z_t)_{t\geq 0}$ is generated with the stopping times $\{T_k\}_{k\in\mathbb{N}^*}$, then $\tilde{x}_k = x_{T_k}$, for $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ generated by (CNA) with choice of parameters as stated. □

By definition of the random times $\{T_k\}_{k\in\mathbb{N}^*}$, $T_k$ follows a Gamma law with shape parameter $k$ and rate 1, such that $\mathbb{E}[T_k] = k$. In order to make comparisons with other algorithms, this motivates us to interpret the bound of Theorem 4.3.2 as an $\mathcal{O}\left(\exp(-\frac{\mu\gamma}{2}k)\right)$ bound in expectation. Convergence results trajectory-wise will however make the comparison more grounded, see Section 4.4.2. Even interpreting the rate of Theorem 4.3.2 as $\mathcal{O}\left(\exp(-\frac{\mu\gamma}{2}k)\right) \approx O\left((1-\frac{\mu\gamma}{2})^k\right)$ does not provide an improved convergence rate, compared to the existing results for gradient descent (Bolte et al., 2017; Karimi et al., 2016), which are already optimal up to a constant factor (Yue et al., 2023). In the deterministic case, Polyak's momentum was known to achieve a similar speed as gradient descent (Danilova et al., 2018) when reducing momentum. However to our knowledge, there exists no convergence result of the Nesterov's momentum under the assumptions of Theorem 4.3.2.

## 4.4 Improved Factor in Convergence Rate for Strongly quasar convex Functions and Better Trajectory-Wise Guarantees

We recall existing convergence guarantees for functions satisfying Assumptions (L-smooth) and (SQC$_{\tau,\mu}$), and use the notation $\kappa := \mu/L$. Gradient descent with stepsize $1/L$ satisfies $f(x_k) - f^* = \mathcal{O}((1-\tau\kappa)^k)$, see Chapter 3. In the same setting, the classical Nesterov scheme (4.1), with a subroutine computing $\alpha_k$, achieves $f(x_k) - f^* = \mathcal{O}((1-\tau\sqrt{\kappa/2})^k)$ (Hinder et al., 2020), an improvement since typically $\kappa \ll 1$, at the cost of an additional logarithmic factor. We removed this subroutine in Chapter 3, recovering essentially the same rate $\mathcal{O}((1-\tau\sqrt{\kappa})^k)$, but requiring an extra assumption controlling the lower curvature, namely $f(x) + \langle\nabla f(x), y - x\rangle + \frac{a}{2}\|x-y\|^2 \leq f(y)$, for $a \leq 0$. This restricts how non-convex the function can be. Interestingly, we also observed in Chapter 3 that this additional assumption is unnecessary in the continuous-time analysis, revealing a gap between continuous and discrete guarantees, likely due to discretization effects. By nature, the continuized framework is less sensitive to discretization issues.

**Theorem 4.4.1** ((Wang and Wibisono, 2023, Theorem 2))**.** *Assume $T_1, \ldots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumptions (L-smooth) and (SQC$_{\tau,\mu}$), iterations of* (CNA) *with $\eta' = \tau\sqrt{\frac{\mu}{L}}$, $\eta = \sqrt{\frac{\mu}{L}}$, $\gamma' = \sqrt{\frac{1}{\mu L}}$ and $\gamma = \frac{1}{L}$ ensure*

$$\mathbb{E}\left[e^{\tau\sqrt{\frac{\mu}{L}}T_k}(f(\tilde{x}_k) - f^*)\right] \leq (f(x_0) - f^* + \mu\|x_0 - x^*\|^2).$$

Up to identifying $T_k$ to $k$ (recall $\mathbb{E}[T_k] = k$), the above result recovers the rate $\mathcal{O}((1-\tau\sqrt{\kappa})^k)$, as it is approximately $\mathcal{O}(\exp(-\tau\sqrt{\kappa}k))$ for $\tau\sqrt{\kappa}$ small. Importantly, it does not assume extra lower-curvature control. This is a first example of a potential benefit of the continuized framework. In the subsequent sections, we show it has still more to offer.

### 4.4.1 Improved Factor in Convergence Rate

Under Assumption (SQC$_{\tau,\mu}$) with $\tau = 1$, (Aujol et al., 2022) show that the solution of the Heavy Ball equation (HB) is such that $f(x_t) - f^* = \mathcal{O}(e^{-\sqrt{2\mu}t})$. This result was improving over the previously known bound $f(x_t) - f^* = \mathcal{O}(e^{-\sqrt{\mu}t})$ (Siegel, 2019). This improvement of the convergence rate by a factor $\sqrt{2}$ was due to a refined choice of Lyapunov function. To transfer this better rate in the discrete setting, authors however also assume convexity. Inspired from their Lyapunov function, we consider the following one, adapted to the continuized framework

$$\varphi(t, x_t, z_t) := a_t(f(x_t) - f^*) + \frac{b_t}{2}\|z_t - x^*\|^2 + \frac{c_t}{2}\|x_t - x^*\|^2. \tag{4.18}$$

The $\sqrt{2}$ factor improvement is due to taking $c_t < 0$, instead of $c_t = 0$, $\forall t \in \mathbb{R}_+$. We use this Lyapunov function to derive a new result in the discrete setting without assuming convexity. We begin with the continuous Lyapunov analysis, using Proposition 4.3.1.

**Theorem 4.4.2.** *Under Assumption (L-smooth)-(SQC$_{\tau,\mu}$)-(SGC$_\rho$), let $(x_t, z_t)$ follow* (CNE) *with $\gamma = \frac{1}{\rho L}$, $\gamma' = \frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}}$, $\eta = \frac{1}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}}$ and $\eta' = \frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}}$ where $\varepsilon \in (0,1)$. Then as long as $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$, we have*

$$\mathbb{E}\left[f(x_t) - f^*\right] \leq \frac{1}{\varepsilon} e^{-\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}(f(x_0) - f^* + \mu\left\|x_0 - x^*\right\|^2). \tag{4.19}$$

Detailed proof of Theorem 4.4.2 is deferred in Section 4.6.1, in which we prove in particular

$$e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}(f(x_t) - f^*) \leq \frac{1}{\varepsilon}\left(f(x_0) - f^* + \mu\left\|x_0 - x^*\right\|^2 + M_t\right), \tag{4.20}$$

for some martingale $(M_t)_{t\geq 0}$ such that $\mathbb{E}\left[M_t\right] = 0$, $\forall t \in \mathbb{R}_+$. Evaluating (4.20) at $t = T_k$ and taking expectation, by Theorem 4.3.1 and Proposition 4.3.2 we directly deduce the following result.

**Theorem 4.4.3.** *Assume $T_1, \ldots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption (L-smooth)-(SQC$_{\tau,\mu}$)-(SGC$_\rho$), iterations of (CNA) with $\eta' = \frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}}$, $\eta = \frac{1}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}}$, $\gamma' = \frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}}$ and $\gamma = \frac{1}{\rho L}$ where $\varepsilon \in (0,1)$ such that $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$ ensure*

$$\mathbb{E}\left[e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}T_k}(f(\tilde{x}_k) - f^*)\right] \leq \frac{1}{\varepsilon}(f(x_0) - f^* + \mu\left\|x_0 - x^*\right\|^2). \tag{4.21}$$

The condition $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$ is not an important restriction for many practical applications, as $\frac{\mu}{L}$ is typically very small. Note that a small $\varepsilon$ improves the rate in the exponential, leading to faster asymptotic convergence, but the $\varepsilon^{-1}$ factor makes the bound less interesting for a small number of iterations.

**Comments** As already mentioned, the continuized framework has been applied to strongly quasar convex functions in the specific case of deterministic gradients, *i.e.* $\rho = 1$ in Assumption (SGC$_\rho$). The result of Theorem 4.4.2 should be compared with (Wang and Wibisono, 2023)

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \left(f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|^2\right)\exp\left(-\tau\sqrt{\frac{\mu}{L}}t\right),$$

while Theorem 4.4.3 should be compared with

$$\mathbb{E}\left[\exp\left(\tau\sqrt{\frac{\mu}{L}}T_k\right)(f(\tilde{x}_k) - f(x^*))\right] \leq f(\tilde{x}_0) - f(x^*) + \frac{\mu}{2}\|\tilde{x}_0 - x^*\|^2,$$

see Theorem 4.4.1. In the case $\tau = 1$, we improve the convergence rate by a factor $\sqrt{2}(1-\varepsilon)$. The improvement decreases as $\tau$ goes to zero. More than this improved factor in the convergence rate, it is worth noting that it is not clear if this improvement can be obtained using more classical methods than the continuized method.

### 4.4.2 Convergence Results with High Probability and Almost Sure

Theorem 4.4.3 states a result that holds in expectation. In applications, algorithms are typically executed only once, hence the importance of trajectory-wise convergence guarantees.

**Convergence with High Probability** We focus on finite time bounds, holding with some probability. We first recall the existing result of this type.

**Corollary 4.4.1** ((Wang and Wibisono, 2023, Corollary 2))**.** *Under the same setting as Theorem 4.4.1, with probability $1 - \frac{1}{c_0} - \frac{1}{(1-c_1)^2 k}$, for some $c_0 > 1$, $c_1 \in (0,1)$ we have*

$$f(\tilde{x}_k) - f^* \leq c_0 e^{-(1-c_1)\tau\sqrt{\frac{\mu}{L}}k}(f(x_0) - f^* + \mu\left\|x_0 - x^*\right\|^2).$$

The above result ensures a convergence rate on $\{f(\tilde{x}_k) - f^*\}_{k\in\mathbb{N}}$ with probability $1 - \frac{1}{c_0} - \frac{1}{(1-c_1)^2 k}$. The convergence rate is close to the one of Theorem 4.4.1, where a constant $c_0$ acts as a degrading factor, and $c_1$ degrades the convergence rate. The proof is based on the Markov inequality and the Chebyshev inequality. The latter is used in the proof to ensure $\mathbb{P}(T_k \leq c_1 k) \leq \frac{1}{(1-c_1)^2 k}$, $c_1 \in (0,1)$. Importantly, the Chebyshev inequality is very general. As $T_k \sim \Gamma(k,1)$, it is in our case highly suboptimal.

**Lemma 4.4.1** (Chernoff inequality)**.** *Let $T_k \sim \Gamma(k,1)$, $0 < c_1 \leq 1$. Then*

$$\mathbb{P}(T_k \leq c_1 k) \leq e^{-(c_1 - 1 - \log(c_1))k}. \tag{4.21}$$

Lemma 4.4.1 bounds the probability that $T_k/k$ is below $c_1 \leq 1$, which decreases exponentially fast for $c_1 < 1$. Intuitively, it allows to replace $T_k$ by $k$ in our convergence bounds, with high probability. Using this lemma, we deduce a tighter result than Corollary 4.4.1, that holds for general exponentially fast convergence rates holding in expectation.

**Proposition 4.4.1.** *Assume $T_k \sim \Gamma(k,1)$, $k \in \mathbb{N}^*$. If we have constants $\beta$, $K_0 > 0$ and a non-negative random variable $Y$ that verifies*

$$\mathbb{E}\left[e^{\beta T_k} Y\right] \leq K_0, \tag{4.22}$$

*then, with probability $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k}$, for some $c_0 > 1$ and $c_1 \in (0,1)$ we have*

$$Y \leq c_0 K_0 e^{-\beta c_1 k}.$$

See the proof of Proposition 4.4.1 in Appendix 4.B.1. A direct application of Proposition 4.4.1 gives the following result.

**Corollary 4.4.2.** *Under the same setting as Theorem 4.4.3, for some $k \in \mathbb{N}^*$, with probability $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k}$, for some $c_0 > 1$, $c_1 \in (0,1)$, we have*

$$f(\tilde{x}_k) - f^* \leq \frac{c_0}{\varepsilon} e^{-\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}} c_1 T_k} (f(x_0) - f^* + \mu \left\|x_0 - x^*\right\|^2).$$

The constant $c_0$ is a degrading factor, while $c_1$ degrades the convergence rates. The closer to 1 are $c_0$ and $c_1$, the closer the bounds are to those that hold in expectation, at the cost of the result holding with lower probability. Using the Chernoff inequality instead of the Chebyshev inequality allows us to significantly refine the results of (Wang and Wibisono, 2023), whose results hold with probability $1 - \frac{1}{c_0} - \frac{1}{(1-c_1)^2 k}$. One can check that $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k} \geq 1 - \frac{1}{c_0} - \frac{1}{(1-c_1)^2 k}$ for any $k \in \mathbb{N}$. Apart from the clear asymptotic improvement, it also appears in finite time, as for $c_1$ close to 1, $\frac{1}{(1-c_1)^2 k}$ is larger than 1 for a significant amount of iterations, making the bound trivial. On the contrary, we have $e^{-(c_1 - 1 - \log(c_1))k} \leq 1$, $\forall k \in \mathbb{N}$, $c_1 \in (0,1)$.

**Almost Sure Convergence** Focusing on asymptotic convergence results, we can obtain an almost sure convergence rate.

**Proposition 4.4.2.** *Assume $T_1, \ldots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. If we have constants $\beta$, $K_0 > 0$ and a sequence of non-negative random variables $\{Y_k\}_{k\in\mathbb{N}^*}$ that verify*

$$\forall k \in \mathbb{N}^*,\ \mathbb{E}\left[e^{\beta T_k} Y_k\right] \leq K_0,$$

*then for any $c \in (0,1)$ we have*

$$Y_k \overset{a.s.}{=} o\left(e^{-\beta c k}\right).$$

See the proof of Proposition 4.4.2 in Appendix 4.B.2. It ensures we can transfer any exponentially fast results that hold in expectation to a result that holds almost surely, with a rate arbitrarily close to $e^{-\beta k}$. The convergence rate stated as $o(\cdot)$ instead of $\mathcal{O}(\cdot)$ typically occurs when studying asymptotic rates (Attouch and Peypouquet, 2016; Sebbouh et al., 2021; Hermant et al., 2025b). The core tool when proving Corollary 4.4.3 is the Borel-Cantelli Lemma, which can be applied because the Chernoff inequality (Lemma 4.4.1) ensures $\sum_{k\geq 0} \mathbb{P}(T_k \leq c_1 k) < +\infty$. In particular, the bound $\mathbb{P}(T_k \leq c_1 k) \leq \frac{1}{(1-c_1)^2 k}$, $c_1 \in (0,1)$ used in (Wang and Wibisono, 2023) would be insufficient to deduce our result. The following result is a direct application of Proposition 4.4.2.

**Corollary 4.4.3.** *Under the same setting as Theorem 4.4.3, for any $c \in (0,1)$, one has*

$$f(\tilde{x}_k) - f^* \overset{a.s.}{=} o\left(e^{-\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}ck}\right). \tag{4.23}$$

## 4.5 Non-Smooth Lyapunov Functions for Relaxing the Uniqueness Assumption

Because of their definition, strongly quasar convex functions have a unique minimizer. A way to relax this limitation is to use the orthogonal projection operator onto $\operatorname{argmin} f$. It has been successfully used in the case of convex functions (Aujol et al., 2024c,b), achieving near-optimal bounds for Nesterov momentum in a setting where former works assumed uniqueness of the minimizer.

**Assumption 1** (SQC'$_\mu$)**.** *$f$ verifies $f^* \geq f(x) + \langle \nabla f(x), \pi(x) - x\rangle + \frac{\mu}{2}\|x - \pi(x)\|^2$ for all $x \in \mathbb{R}^d$, some $\mu \in \mathbb{R}_+^*$, and with $\pi(\cdot)$ being the orthogonal projection onto* $\operatorname{argmin} f$*.*

Assumption (SQC'$_\mu$) is a variant of Assumption (SQC$_{\tau,\mu}$) in the case $\tau = 1$, where an explicit minimizer $x^*$ is replaced by $\pi(x)$, the orthogonal projection of $x$ onto $\operatorname{argmin} f$. This assumption is used in (Gupta and Wojtowytsch, 2025) under the name *$\mu$-strong aiming condition*. If they can obtain a convergence result for the Heavy Ball dynamical system (HB), further obtaining a convergence result on the Nesterov momentum algorithm (4.1) requires that the projection operator is affine-linear. This forces the set of minimizers to be affine, and indicates a discrepancy between continuous and discrete results. We further relax this assumption.

**Assumption 4.5.1.** $\operatorname{argmin} f$ *is a convex set with $C^2$ boundary, i.e. its boundary is locally a $C^2$ sub-manifold.*

The $C^2$ boundary assumption can be generalized to second order regularity (Shapiro, 2016; Bonnans et al., 1998), which we leave aside for clarity of exposition. Considering non-convex objects, ensuring $\pi(\cdot)$ is well-defined is technically challenging. A feasible approach is to consider local definitions, namely defining $\pi(\cdot)$ for elements that are close enough to minimizers (Poliquin et al., 2000; Gupta and Wojtowytsch, 2025). In order to define it globally, it is however needed that $\operatorname{argmin} f$ is closed and convex (Jessen, 1940; Busemann, 1947; Phelps, 1957), such that it is the most general assumption we can make to ensure it is globally defined.

### 4.5.1 A Technical Challenge: $\pi(\cdot)$ is not $C^1$

A major building block of the convergence analysis using the continuized framework is Proposition 4.3.1, which assumes the Lyapunov function belongs to $C^1$. Under Assumption (SQC'$_\mu$), we will construct our Lyapunov function so that it involves the projection $\pi(\cdot)$. However, $\operatorname{argmin} f$ being a closed convex set with $C^2$ boundary does not ensure $\pi$ to be $C^1$, as it is not differentiable on $\partial\operatorname{argmin} f$ (Holmes, 1973; Fitzpatrick and Phelps, 1982), where $\partial\operatorname{argmin} f$ is the boundary of $\operatorname{argmin} f$. Proposition 4.3.1 thus cannot be applied straightforwardly. We generalize it to include a broader set of Lyapunov functions.

**Proposition 4.5.1** (Non-Smooth Itô Formula)**.** *Let $x_t \in \mathbb{R}^d$ be a solution of*

$$dx_t = \zeta(x_t)dt + \int_\Xi G(x_{t^-}, \xi)N(dt, d\xi),$$

*where $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ is a locally Lipschitz function and $G : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ is measurable. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. Then, denoting $\phi(t) = \varphi(x_t)$, there exists $\mathring{A} \subset \mathbb{R}_+$ of zero measure such that the derivative of $\phi$, denoted by $\dot{\phi}$, exists on $\mathbb{R}_+\backslash\mathring{A}$. Moreover, if*

$$\forall s \in [0,t]\backslash\mathring{A}, \quad \dot{\phi}(s) + \mathbb{E}_\xi[(\varphi(x_s + G(x_s,\xi)) - \varphi(x_s))] \leq 0,$$

*then we have*

$$\phi(t) \leq \phi(0) + M_t,$$

*where $(M_t)_{t\geq 0}$ is a martingale such that $\mathbb{E}[M_t] = 0$ for all $t \in \mathbb{R}_+$.*

*Proof.* Note first that $T_n \to +\infty$ almost surely. This can be seen as $T_n = \sum_{i=1}^n E_i$, $E_i \overset{i.i.d}{\sim} \mathcal{E}(1)$, and by the law of large numbers $T_n/n \to \mathbb{E}[1] = 1$, almost surely. This permits us to write

$$\begin{aligned}\phi(t) - \phi(0) &= \sum_{k=0}^{+\infty} \phi(t \wedge T_{k+1}) - \phi(t \wedge T_k) \\ &= \underbrace{\sum_{k=0}^{+\infty} \left(\phi(t \wedge T_{k+1}) - \phi(t \wedge T_{k+1}^-)\right)}_{\text{(I)}} + \underbrace{\sum_{k=0}^{+\infty} \left(\phi(t \wedge T_{k+1}^-) - \phi(t \wedge T_k)\right)}_{\text{(II)}}.\end{aligned} \tag{4.24}$$

We treat (I) and (II) separately.

**Sum (I)** (I) describes the jumps. We have

$$\begin{aligned}\sum_{k=0}^{\infty} \phi(t \wedge T_{k+1}) - \phi(t \wedge T_{k+1}^-) &= \sum_{k \geq 0:\, T_{k+1} \leq t} \phi(T_{k+1}) - \phi(T_{k+1}^-) \\ &= \sum_{k \geq 0:\, T_{k+1} \leq t} \varphi(x_{T_{k+1}^-} + G(x_{T_{k+1}^-}, \xi_{k+1})) - \varphi(x_{T_{k+1}^-}) \\ &= \int_0^t \int_\Xi \varphi(x_{s-} + G(x_{s-}, \xi)) - \varphi(x_{s-})\, N(ds, d\xi).\end{aligned} \tag{4.25}$$

**Sum (II)** The sum (II) is composed of $\phi(t \wedge T_{k+1}^-) - \phi(t \wedge T_k)$ terms. For $s \in (t \wedge T_k, t \wedge T_{k+1})$, we have

$$\dot{x}_t = \zeta(x_t), \tag{4.26}$$

such that $x_s$ is $C^1$ on $(t \wedge T_k, t \wedge T_{k+1})$. Together with the fact that $\varphi$ is locally Lipschitz, it induces that $\phi(t) = \varphi(x_t)$ is locally Lipschitz on $(t \wedge T_k, t \wedge T_{k+1})$. Then, $\phi$ is locally absolutely continuous on $(t \wedge T_k, t \wedge T_{k+1})$. This means that its derivative $\dot{\phi}$ exists almost everywhere on $(t \wedge T_k, t \wedge T_{k+1})$, and that we have

$$\phi(t \wedge T_{k+1}^-) - \phi(t \wedge T_k) = \int_{t \wedge T_k}^{t \wedge T_{k+1}} \dot{\phi}(s) ds,$$

where we used that $\phi$ is right-continuous at $T_k$ because $(x_t)_{t \geq 0}$ is càdlàg. Applying the same reasoning on each interval $(t \wedge T_k, t \wedge T_{k+1})$, $k \in \mathbb{N}$, we have that the derivative of $\phi$ exists almost everywhere, and that Sum (II) rewrites

$$\sum_{k=0}^{+\infty} \left(\phi_{t \wedge T_{k+1}^-} - \phi_{t \wedge T_k}\right) = \sum_{k=0}^{+\infty} \int_{t \wedge T_k}^{t \wedge T_{k+1}} \dot{\phi}(s) ds = \int_0^t \dot{\phi}(s) ds \tag{4.27}$$

**Conclusion** Combining (4.25) and (4.27), we obtain

$$\begin{aligned}\phi(t) &= \phi(0) + \int_0^t \dot{\phi}(s) ds + \int_{[0,t]} (\varphi(x_{s^-} + G(x_{s^-}, \xi)) - \varphi(x_{s^-})) N(ds, d\xi) \\ &= \phi(0) + \int_0^t \dot{\phi}(s) ds + \int_{[0,t]} \mathbb{E}_\xi[(\varphi(x_s + G(x_s, \xi)) - \varphi(x_s))] ds \\ &+ \int_{[0,t]} \int_\Xi (\varphi(x_{s^-} + G(x_{s^-}, \xi)) - \varphi(x_{s^-}))(N(ds, d\xi) - ds \mathcal{P}(d\xi)) \\ &= \phi(0) + \int_0^t \dot{\phi}(s) + \mathbb{E}_\xi[(\varphi(x_s + G(x_s, \xi)) - \varphi(x_s))] ds + M_t,\end{aligned}$$

where $(M_t)_{t \geq 0}$ is a martingale such that $\mathbb{E}[M_t] = 0$ for all $t \in \mathbb{R}_+$. Noting $\mathring{A}$ the set of null measure on which $\dot{\phi}$ may not be defined, by non-negativity of the Lebesgue integral, we have that if

$$\forall s \in [0, t] \backslash \mathring{A}, \quad \dot{\phi}(s) + \mathbb{E}_\xi[(\varphi(x_s + G(x_s, \xi)) - \varphi(x_s))] \leq 0,$$

then

$$\phi(t) \leq \phi(0) + M_t.$$

□

Compared to Proposition 4.3.1, Proposition 4.5.1 provides a Lyapunov analysis framework in the case of non-$C^1$ Lyapunov functions. Precisely, it relaxes the $C^1$ assumption to allow locally Lipschitz functions. Intuitively, it indicates that even with non-differentiable Lyapunov function, we still have a rule to ensure $\mathbb{E}[\phi(t)] \leq \phi(0)$. In this case, having to show $\dot{\phi}(s) + \mathbb{E}_\xi[(\varphi(x_s + G(x_s, \xi)) - \varphi(x_s))] \leq 0$ plays the same role as establishing $\langle \nabla\varphi(x_s), \zeta(x_s)\rangle + \mathbb{E}_\xi\varphi(x_s + G(x_s)) - \varphi(x_s) \leq 0$ in the differentiable case, see *e.g.* the proof of Proposition 4.3.3.

**Remark 4.5.1.** *As the function $f$ we minimize is typically involved in Lyapunov functions, Proposition 4.5.1 requires $f$ to be locally Lipschitz. This is a natural assumption in non-smooth optimization, as it ensures that the slopes are locally bounded. However, it implies $f$ is continuous. It might be necessary to carry the continuized analysis, as we want to write $\phi$ as an integral in the proof of Proposition 4.5.1.*

### 4.5.2 Convergence Result

We consider the following Lyapunov function

$$\varphi(t, x_t, z_t) := a_t(f(x_t) - f^*) + \frac{b_t}{2}\left\|z_t - \pi(x_t)\right\|^2. \tag{4.28}$$

This Lyapunov function is not $C^1$ because of the projection, and we cannot apply Proposition 4.3.1. However $\pi(\cdot)$ is Lipschitz, so in particular locally Lipschitz, such that Proposition 4.5.1 applies.

**Theorem 4.5.1.** *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumptions (L-smooth)-(SQC'$_\mu$)-(SGC$_\rho$)-4.5.1, iterations of (CNA) with $\gamma = \frac{1}{\rho L}$, $\gamma' = \frac{1}{\rho\sqrt{\mu L}}$, $\eta = \frac{1 - \frac{1+\sqrt{1+4\rho}}{2}\left(\frac{\mu}{L}\right)^{1/4}}{\rho}\sqrt{\frac{\mu}{L}}$ and $\eta' = \frac{1}{\rho}\sqrt{\frac{\mu}{L}}$ ensure the following bound*

$$\mathbb{E}\left[e^{\eta T_k}(f(\tilde{x}_k) - f^*)\right] \leq f(x_0) - f^* + \mu\left(1 - \frac{1+\sqrt{1+4\rho}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right)\left\|x_0 - \pi(x_0)\right\|^2. \tag{4.29}$$

The proof of Theorem 4.5.1 is in Section 4.6.2. Its bound is nontrivial as long as $\frac{\mu}{L} < \left((1+\sqrt{1+4\rho})/2\right)^{-4}$. In the deterministic gradient case, *i.e.* $\rho = 1$, this upper bound becomes $\left((1+\sqrt{1+4\rho})/2\right)^{-4} = \left((1+\sqrt{5})/2\right)^{-4} \approx 0.15$, which is a reasonable bound as typically in most applications we expect that $\frac{\mu}{L} \ll 1$. Theorem 4.5.1 ensures an accelerated convergence rate relaxing the assumptions holding in (Wang and Wibisono, 2023) (unique minimizer, deterministic gradient) and (Gupta and Wojtowytsch, 2025) (lower curvature restriction, argmin $f$ is affine-linear).

**Remark 4.5.2.** *Note that the Lyapunov function* (4.28) *used to prove Theorem 4.5.1 does not incorporate the negative term used in the Lyapunov function* (4.18) *to prove Theorem 4.4.3, in the case of uniqueness of the minimizer. Investigating whether this modification could yield a slightly better rate is a potential direction for future work.*

**Remark 4.5.3.** *The assumption that* argmin $f$ *is convex helps to ensure $\pi(\cdot)$ is a well-defined function. It is however important to note that in the computation, this assumption is only used to have that the projection is Lipschitz with constant* 1*. It indicates that if the trajectory is such that along its path, $\pi(\cdot)$ is well-defined with existing partial derivatives, then this assumption could be relaxed by considering a higher Lipschitz constant for the projection.*

**Result Trajectory-Wise with High Probability and Almost Surely.** Following the discussion of Section 4.4.2, we extend the result of Theorem 4.5.1 to results that hold trajectory-wise.

**Corollary 4.5.1.** *Under the same setting as for Theorem 4.5.1, one has that:*

1. *With probability* $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k}$, $c_0 > 1$, $c_1 \in (0,1)$,

$$f(\tilde{x}_k) - f^* \leq c_0 \left( f(x_0) - f^* + \mu \left( 1 - \frac{1 + \sqrt{1+4\rho}}{2} \left( \frac{\mu}{L} \right)^{1/4} \right) \|x_0 - \pi(x_0)\|^2 \right) e^{-\eta c_1 k}.$$

2. *For any* $c \in (0,1)$,

$$f(\tilde{x}_k) - f^* \overset{a.s.}{=} o\left(e^{-\eta c k}\right).$$

Corollary 4.5.1 is a direct application of Proposition 4.4.1 and Proposition 4.4.2.

## 4.6 Convergence Proofs

In this section we prove our main convergence results.

### 4.6.1 Proof of Theorem 4.4.2

We start by showing a sufficient decrease of the Lyapunov function.

**Lemma 4.6.1.** *Let* $(x_t, z_t)$ *defined by* (CNE). *Let Assumptions (L-smooth)-(SQC*$_{\tau,\mu}$*)-(SGC*$_\rho$*) hold. We define the following Lyapunov function*

$$\phi(t) = a_t(f(x_t) - f^*) + \frac{b_t}{2} \|z_t - x^*\|^2 + \frac{c_t}{2} \|x_t - x^*\|^2 . \tag{4.30}$$

*We fix* $a_t = e^{2\eta' t}$, $b_t = \frac{2\mu}{2-\tau} e^{2\eta' t}$, $c_t = -(1-\varepsilon)\mu \frac{\tau}{2-\tau} e^{2\eta' t}$, $\eta' = \frac{(1-\varepsilon)\tau}{\rho} \sqrt{\frac{1}{2(2-\tau)}} \sqrt{\frac{\mu}{L}}$, *and* $\eta = \frac{1}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}}$ , $\gamma' = \frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}}$ *and* $\gamma = \frac{1}{\rho L}$ *where* $\varepsilon \in (0,1)$. *Under the assumption* $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$, *we have*

$$\phi(t) \leq \phi(0) + M_t, \tag{4.31}$$

*where by definition*

$$\phi(t) = e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t} \left( f(x_t) - f^* + \frac{\mu}{2-\tau} \|z_t - x^*\|^2 + \frac{-(1-\varepsilon)\mu}{2} \frac{\tau}{2-\tau} \|x_t - x^*\|^2 \right). \tag{4.32}$$

*Proof.* Let $\overline{x}_t = (t, x_t, z_t)$. It satisfies $d\overline{x}_t = \zeta(\overline{x}_t)dt + G(\overline{x}_t, \xi)N(dt, d\xi)$, where

$$\zeta(\overline{x}_t) = \begin{pmatrix} 1 \\ \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\overline{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma \nabla f(x_t, \xi) \\ -\gamma' \nabla f(x_t, \xi). \end{pmatrix} \tag{4.33}$$

We apply Proposition 4.3.1 to $\phi_t = \varphi(\overline{x}_t)$, where

$$\varphi(t, x, z) = a_t(f(x) - f^*) + \frac{b_t}{2} \|z - x^*\|^2 + \frac{c_t}{2} \|x - x^*\|^2 . \tag{4.34}$$

so

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \langle \nabla \varphi(\overline{x}_s), \zeta(\overline{x}_s) \rangle \, ds + \int_{[0,t]} \mathbb{E}_\xi(\varphi(\overline{x}_s + G(\overline{x}_s, \xi)) - \varphi(\overline{x}_s))ds + M_t, \tag{4.35}$$

where $M_t$ is a martingale. We have

$$\frac{\partial \varphi}{\partial s} = \frac{da_s}{ds}(f(x) - f^*) + \frac{1}{2}\frac{db_s}{ds} \|z - x^*\|^2 + \frac{1}{2}\frac{dc_s}{ds} \|x - x^*\|^2 , \tag{4.36}$$

$$\frac{\partial \varphi}{\partial x} = a_s \nabla f(x) + c_s(x - x^*), \tag{4.37}$$

$$\frac{\partial \varphi}{\partial z} = b_s(z - x^*). \tag{4.38}$$

So,

$$\begin{aligned}
\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle &= \frac{da_s}{ds}(f(x_s)-f^*) + \frac{1}{2}\frac{db_s}{ds}\|z_s-x^*\|^2 + \frac{1}{2}\frac{dc_s}{ds}\|x_s-x^*\|^2 \\
&\quad + \langle a_s\nabla f(x_s) + c_s(x_s-x^*), \eta(z_s-x_s)\rangle + \langle b_s(z_s-x^*), \eta'(x_s-z_s)\rangle \\
&= \frac{da_s}{ds}(f(x_s)-f^*) + \frac{1}{2}\frac{db_s}{ds}\|z_s-x^*\|^2 + \frac{1}{2}\frac{dc_s}{ds}\|x_s-x^*\|^2 && (4.39)\\
&\quad + a_s\eta\langle \nabla f(x_s), z_s-x_s\rangle && (4.40)\\
&\quad + c_s\eta\langle x_s-x^*, z_s-x_s\rangle + b_s\eta'\langle z_s-x^*, x_s-z_s\rangle && (4.41)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}_\xi[\varphi(\overline{x}_s+G(\overline{x}_s)) - \varphi(\overline{x}_s)] &= \mathbb{E}_\xi[a_s(f(x_s-\gamma\nabla f(x_s,\xi)) - f(x_s))] \\
&\quad + \frac{b_s}{2}\mathbb{E}_\xi[\|z_s-\gamma'\nabla f(x_s,\xi)-x^*\|^2 - \|z_s-x^*\|^2] \\
&\quad + \frac{c_s}{2}\mathbb{E}_\xi[\|x_s-\gamma\nabla f(x_s,\xi)-x^*\|^2 - \|x_s-x^*\|^2] \\
&= a_s\mathbb{E}_\xi[f(x_s-\gamma\nabla f(x_s,\xi)) - f(x_s)] && (4.42)\\
&\quad + \frac{b_s\gamma'^2 + c_s\gamma^2}{2}\mathbb{E}_\xi[\|\nabla f(x_s,\xi)\|^2] && (4.43)\\
&\quad + b_s\gamma'\langle \nabla f(x_s), x^*-z_s\rangle + c_s\gamma\langle \nabla f(x_s), x^*-x_s\rangle, && (4.44)
\end{aligned}$$

where we used $\mathbb{E}_\xi[\nabla f(x_s,\xi)] = \nabla f(x_s)$. Using Assumptions (L-smooth)-(SGC$_\rho$) + Lemma 4.2.1, we have

$$(4.42) + (4.43) \le \left(\rho b_s\gamma'^2 + \rho c_s\gamma^2 + a_s\gamma(\rho\gamma L - 2)\right)\frac{1}{2}\|\nabla f(x_s)\|^2. \tag{4.45}$$

Also,

$$(4.44) = (b_s\gamma' + c_s\gamma)\langle \nabla f(x_s), x^*-x_s\rangle + b_s\gamma'\langle \nabla f(x_s), x_s-z_s\rangle. \tag{4.46}$$

Using $\tau,\mu$-strong quasar convexity, we get

$$(4.40) + (4.46) \le -(b_s\gamma' + c_s\gamma)\tau\left(f(x_s) - f^* + \frac{\mu}{2}\|x_s-x^*\|\right) + (a_s\eta - b_s\gamma')\langle \nabla f(x_s), z_s-x_s\rangle. \tag{4.47}$$

Also,

$$\begin{aligned}
(4.41) &= -b_s\eta'\|z_s-x^*\|^2 + c_s\eta\langle x_s-x^*, z_s-x_s\rangle + b_s\eta'\langle z_s-x^*, x_s-x^*\rangle && (4.48)\\
&= -b_s\eta'\|z_s-x^*\|^2 + b_s\eta'\|x_s-x^*\|^2, && (4.49)
\end{aligned}$$

where we assumed $c_s\eta = -b_s\eta' \Rightarrow c_s = -\frac{\eta'}{\eta}b_s$. We now put everything together

$$\begin{aligned}
\langle \nabla\varphi(x_s), \zeta(x_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s+G(\overline{x}_s)) - \varphi(\overline{x}_s)] &\le \left(\frac{da_s}{ds} - \tau(b_s\gamma' + c_s\gamma)\right)(f(x_s)-f^*) && (4.50)\\
&\quad + \left(\frac{db_s}{ds} - 2b_s\eta'\right)\frac{1}{2}\|z_s-x^*\|^2 && (4.51)\\
&\quad + \left(\frac{dc_s}{ds} - \mu\tau(b_s\gamma' + c_s\gamma) + 2b_s\eta'\right)\frac{1}{2}\|x_s-x^*\|^2 && (4.52)\\
&\quad + (a_s\eta - b_s\gamma')\langle \nabla f(x_s), z_s-x_s\rangle && (4.53)\\
&\quad + (\rho b_s\gamma'^2 + \rho c_s\gamma^2 + a_s\gamma(\rho\gamma L-2))\frac{1}{2}\|\nabla f(x_s)\|^2. && (4.54)
\end{aligned}$$

**Parameter Choices** Inspired from computations on deterministic damped system, we fix $a_s = e^{2\eta' s}$, $b_s = \frac{2\mu}{2-\tau} e^{2\eta' s}$, $c_s = -\theta\mu\frac{\tau}{2-\tau}e^{2\eta' s}$, with $\theta > 0$ to be defined later. Also, $\eta' = \frac{\theta\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}}$, and $\eta = \frac{1}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}} \Rightarrow \frac{\eta'}{\eta} = \frac{\tau\theta}{2}$. Also, we fix $\gamma' = \frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}}$ and $\gamma = \frac{1}{\rho L}$.

The exponential rate ensures the term of (4.51) is non-positive.

Also, the term of (4.53) is zero as long as $a_s\eta - b_s\gamma' = 0 \Leftrightarrow \eta = \frac{2\mu}{2-\tau}\frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}} = \frac{1}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}}$.

The term of (4.50) is

$$a_s^{-1}e^{-2\eta' s}\left(\frac{da_s}{ds} - \tau(b_s\gamma' + c_s\gamma)\right) = 2\eta' - 2\tau\frac{\mu}{2-\tau}\gamma' + \theta\mu\frac{\tau^2}{2-\tau}\gamma \tag{4.55}$$

$$= 2\frac{\theta\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}} - 2\frac{\mu\tau}{2-\tau}\frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}} + \theta\mu\frac{\tau^2}{2-\tau}\frac{1}{L\rho} \tag{4.56}$$

$$= \frac{\tau}{\rho}\sqrt{\frac{2}{2-\tau}}\sqrt{\frac{\mu}{L}}\left(\theta - 1 + \frac{\theta\tau}{\sqrt{2(2-\tau)}}\sqrt{\frac{\mu}{L}}\right). \tag{4.57}$$

Omitting exponential factors, the term of (4.54) is

$$\rho b_s\gamma'^2 + \rho c_s\gamma^2 + a_s\gamma(\rho\gamma L - 2)) = \rho\frac{2\mu}{2-\tau}\frac{1}{\rho^2}\left(\sqrt{\frac{2-\tau}{2\mu L}}\right)^2 - \rho\theta\mu\frac{\tau}{2-\tau}\frac{1}{\rho^2 L^2} - \frac{1}{\rho L}$$

$$= -\theta\frac{\tau}{2-\tau}\frac{\mu}{\rho L^2}. \tag{4.58}$$

The term of (4.52) is

$$\frac{dc_s}{ds} - \mu\tau(b_s\gamma' + c_s\gamma) + 2b_s\eta' =$$

$$-2\frac{\theta\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}}\theta\mu\frac{\tau}{2-\tau} - \mu\tau\frac{2\mu}{2-\tau}\frac{1}{\rho}\sqrt{\frac{2-\tau}{2\mu L}} + \mu\tau\theta\mu\frac{\tau}{2-\tau}\frac{1}{\rho L} + 2\frac{2\mu}{2-\tau}\frac{\theta\tau}{\rho}\sqrt{\frac{1}{2(2-\tau)}}\sqrt{\frac{\mu}{L}}$$

$$= \mu\sqrt{\frac{\mu}{L}}\frac{\tau}{\rho}\left(-\theta^2\sqrt{\frac{2}{2-\tau}}\frac{\tau}{2-\tau} - \sqrt{\frac{2}{2-\tau}} + \frac{\tau\theta}{2-\tau}\sqrt{\frac{\mu}{L}} + 2\theta\sqrt{\frac{2}{2-\tau}}\frac{1}{2-\tau}\right)$$

$$= \sqrt{\frac{2}{2-\tau}}\frac{1}{2-\tau}\mu\sqrt{\frac{\mu}{L}}\frac{\tau}{\rho}\left(-\tau\theta^2 - 2 + \tau + \tau\theta\sqrt{\frac{2-\tau}{2}}\sqrt{\frac{\mu}{L}} + 2\theta\right). \tag{4.59}$$

(4.59) is not straightforward to cancel. Symbolic computations in Python show that the maximal solution is

$$\theta_- = \frac{1}{\tau} + \frac{1}{4}\sqrt{2-\tau}\sqrt{\frac{2\mu}{L}} - \frac{1}{2\sqrt{2}\tau}\sqrt{4\sqrt{2}\tau\sqrt{2-\tau}\sqrt{\frac{\mu}{L}} - \tau^2(\tau-2)\frac{\mu}{L} + 8(1-\tau)^2},$$

but the expression is barely readable. Otherwise, getting rid of the factor $\sqrt{\frac{2}{2-\tau}}\frac{1}{2-\tau}\mu\sqrt{\frac{\mu}{L}}\frac{\tau}{\rho}$, fixing $\theta = 1-\alpha$, (4.59) becomes

$$-\tau(1-\alpha)^2 - 2 + \tau + (1-\alpha)\tau\sqrt{\frac{2-\tau}{2}}\sqrt{\frac{\mu}{L}} + 2 - 2\alpha$$

$$\leq -\tau\alpha^2 + 2(\tau-1)\alpha + \tau\sqrt{\frac{2-\tau}{2}}\sqrt{\frac{\mu}{L}}$$

$$\leq -\tau\alpha^2 + \tau\sqrt{\frac{2-\tau}{2}}\sqrt{\frac{\mu}{L}}.$$

The last inequality uses $\tau - 1 \leq 0$. Now, fixing $\alpha \geq c\tau^\alpha\left(\frac{\mu}{L}\right)^\beta$, for some non-negative constants $c$, $\alpha$ and $\beta$, observe that $c = 1$, $\alpha = 0$ and $\beta = 1/4$ is sufficient to have (4.59) upper-bounded by 0. Plugging this choice

in (4.57), we have that (4.50) is non-positive as

$$\theta - 1 + \frac{\theta\tau}{\sqrt{2(2-\tau)}}\sqrt{\frac{\mu}{L}} \leq -\left(\frac{\mu}{L}\right)^{\frac{1}{4}} + \frac{(1-\left(\frac{\mu}{L}\right)^{\frac{1}{4}})\tau}{\sqrt{2(2-\tau)}}\sqrt{\frac{\mu}{L}} \tag{4.60}$$

$$\leq -\left(\frac{\mu}{L}\right)^{\frac{1}{4}} + \frac{\tau}{\sqrt{2(2-\tau)}}\sqrt{\frac{\mu}{L}} \tag{4.61}$$

$$= -\left(\frac{\mu}{L}\right)^{\frac{1}{4}}\left(1 - \frac{\sqrt{\tau}}{2^{\frac{1}{4}}\sqrt{2-\tau}}\left(\frac{\mu}{L}\right)^{\frac{1}{4}}\right) \tag{4.62}$$

$$\leq 0, \tag{4.63}$$

the last inequality holding as $\tau \leq 1$ and $\frac{\mu}{L} \leq 1$.

In particular, fixing $\theta = 1 - \varepsilon$ with $\varepsilon \in (0,1)$ such that $\varepsilon \geq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}$, we have

$$\phi(t) \leq \phi(0) + M_t, \tag{4.64}$$

where

$$\phi(t) = e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}\left(f(x_t) - f^* + \frac{\mu}{2-\tau}\left\|z_t - x^*\right\|^2 + \frac{-(1-\varepsilon)\mu}{2}\frac{\tau}{2-\tau}\left\|x_t - x^*\right\|^2\right). \tag{4.65}$$

□

**Proof of Theorem 4.4.2 .** We use the following result, ensuring a quadratic growth property.

**Lemma 4.6.2.** *(Hinder et al., 2020, Corollary 1) If $f$ verifies Assumption ($SQC_{\tau,\mu}$), we have*

$$f(x) - f^* \geq \frac{\mu\tau}{2(2-\tau)}\left\|x - x^*\right\|^2, \quad \forall x \in \mathbb{R}^d.$$

Recalling (4.65), using Lemma 4.6.2, we have

$$\phi(t) \geq \varepsilon e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}(f(x_t) - f^*). \tag{4.66}$$

So, by Lemma 4.6.1, we have

$$e^{\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}(f(x_t) - f^*) \leq \frac{1}{\varepsilon}\left(\phi(0) + M_t\right). \tag{4.67}$$

Taking expectation, we have

$$\mathbb{E}\left[f(x_t) - f^*\right] \leq \frac{1}{\varepsilon}e^{-\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}\phi(0). \tag{4.68}$$

Using $x_0 = z_0$, and $\tau \leq 1$, by the expression of $\phi(0)$ we have

$$\mathbb{E}\left[f(x_t) - f^*\right] \leq \frac{1}{\varepsilon}e^{-\frac{(1-\varepsilon)\tau}{\rho}\sqrt{\frac{2}{(2-\tau)}}\sqrt{\frac{\mu}{L}}t}(f(x_0) - f^* + \mu\left\|x_0 - x^*\right\|^2). \tag{4.69}$$

### 4.6.2 Proof of Theorem 4.5.1

Let $\overline{x}_t = (t, x_t, z_t)$. It satisfies $d\overline{x}_t = \zeta(\overline{x}_t)dt + G(\overline{x}_t, \xi)N(dt, d\xi)$, where

$$\zeta(\overline{x}_t) = \begin{pmatrix} 1 \\ \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\overline{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma\nabla f(x_t, \xi) \\ -\gamma'\nabla f(x_t, \xi). \end{pmatrix} \tag{4.70}$$

We define

$$\varphi(t, x, z) = a_t(f(x) - f^*) + \frac{b_t}{2}\left\|\pi(x) - z\right\|^2, \tag{4.71}$$

with $\pi(\cdot)$ being the orthogonal projection onto $\operatorname{argmin} f$, namely

$$\forall x \in \mathbb{R}^d,\ \pi(x) = \arg\min_{y \in \operatorname{argmin} f} \|y - x\|^2 . \tag{4.72}$$

The property of the orthogonal projection onto a convex set has been studied extensively (Zarantonello, 1971). $\operatorname{argmin} f$ being a closed convex set with $C^2$ boundary ensures $\pi$ is well-defined on $\mathbb{R}^d$ (Holmes, 1973; Fitzpatrick and Phelps, 1982), and in particular, it is 1-Lipschitz. As $f$ is $C^1$, we have that $\varphi$ is locally Lipschitz. Noting $\phi(t) := \varphi(\overline{x}_t)$, we apply Proposition 4.5.1, such that $\dot{\phi}$ the derivative of $\phi$ exists on $\mathbb{R}_+ \backslash \mathbb{A}$, where $\mathbb{A}$ is a set of null measure. In particular on $\mathbb{R}_+ \backslash \mathbb{A}$, $(x_t, z_t)_{t \geq 0}$ are $C^1$ and verify

$$\begin{cases} \dot{x}_s &= \eta(z_s - x_s) \\ \dot{z}_s &= \eta'(x_s - z_s) \end{cases} \tag{4.73}$$

So, on $\mathbb{R}_+ \backslash \mathbb{A}$, one has

$$\phi(s) = \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\|\pi(x_s) - z_s\|^2 + a_s \langle \nabla f(x_s), \dot{x}_s \rangle + b_s \langle \pi(x_s) - z_s, \partial\pi(x_s, \dot{x}_s) - \dot{z}_s \rangle ,$$

where $\partial\pi(x_s, \dot{x}_s)$ denotes the directional derivative of $\pi(\cdot)$ at point $x_s$ in the direction $\dot{x}_s$. Then, with some computations

$$\begin{aligned}
\dot{\phi}(s) &= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\|\pi(x_s) - z_s\|^2 + a_s \langle \nabla f(x_s), \dot{x}_s \rangle + b_s \langle \pi(x_s) - z_s, \partial\pi(x_s, \dot{x}_s) - \dot{z}_s \rangle \\
&= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\|\pi(x_s) - z_s\|^2 + \eta a_s \langle \nabla f(x_s), z_s - x_s \rangle \\
&+ b_s \langle \pi(x_s) - z_s, \partial\pi(x_s, \dot{x}_s) - \eta'(x_s - z_s) \rangle \\
&= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\|\pi(x_s) - z_s\|^2 + \eta a_s \langle \nabla f(x_s), z_s - x_s \rangle + \eta' b_s \langle z_s - \pi(x_s), x_s - z_s \rangle \\
&+ b_s \langle \pi(x_s) - x_s, \partial\pi(x_s, \dot{x}_s) \rangle + b_s \langle x_s - z_s, \partial\pi(x_s, \dot{x}_s) \rangle \\
&= \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\|\pi(x_s) - z_s\|^2 + \eta a_s \langle \nabla f(x_s), z_s - x_s \rangle + \eta' b_s \langle z_s - \pi(x_s), x_s - z_s \rangle \\
&+ b_s \langle \pi(x_s) - x_s, \partial\pi(x_s, \dot{x}_s) \rangle - \frac{b_s}{\eta} \langle \dot{x}_s, \partial\pi(x_s, \dot{x}_s) \rangle .
\end{aligned} \tag{4.74}$$

In the last equality, we use $\langle x_s - z_s, \partial\pi(x_s, \dot{x}_s) \rangle = -\frac{1}{\eta} \langle \dot{x}_s, \partial\pi(x_s, \dot{x}_s) \rangle$, induced by (4.73). The two last terms of (4.74) would automatically be zero if $\operatorname{argmin} f$ was reduced to a singleton, as the projection would be identically equal to $x^*$. We show that in our case we can upper-bound these terms by zero.

**Lemma 4.6.3.** *Let $U$ be a subset of $\mathbb{R}^d$ and $K \subset U$ such that the orthogonal projection $P : U \to K$ is well-defined. Then,*

1. $\forall x \in U, y \in K, \quad \langle x - y, y - P(x) \rangle \leq 0.$
2. $\forall x, d \in U, \quad \langle P(x + d) - P(x), d \rangle \geq 0.$

These are classic properties, but are often proved under the underlying assumption that $K$ is convex. We provide a proof in Appendix 4.C.2. Applying Lemma 4.6.3-1. with $P = \pi$ to $x = x_s + hd$, $d = \dot{x}_s$ and $y = \pi(x_s)$ gives

$$\langle x_s + hd - \pi(x_s), \pi(x_s) - \pi(x_s + hd) \rangle \leq 0.$$

We split this scalar product between two terms

$$\langle x_s - \pi(x_s), \pi(x_s) - \pi(x_s + hd) \rangle + h \langle d, \pi(x_s) - \pi(x_s + hd) \rangle \leq 0.$$

Dividing by $h > 0$, we obtain

$$\left\langle x_s - \pi(x_s), \frac{\pi(x_s) - \pi(x_s + hd)}{h} \right\rangle + \langle d, \pi(x_s) - \pi(x_s + hd) \rangle \leq 0.$$

By continuity of $\pi(\cdot)$, the right term goes to zero as $h$ goes to zero. So, taking the limit with $h$ going to zero yields

$$\left\langle x_s - \pi(x_s), \lim_{h\to 0} \frac{\pi(x_s) - \pi(x_s + hd)}{h} \right\rangle = -\left\langle x_s - \pi(x_s), \partial\pi(x_s, \dot{x}_s)\right\rangle \leq 0. \tag{4.75}$$

Applying again Lemma 4.6.3-1. with $P = \pi$ to $x = x_s$, $d = \dot{x}_s$ and $y = \pi(x_s + h\dot{x}_s)$ gives

$$\langle x_s - \pi(x_s + h\dot{x}_s), \pi(x_s + h\dot{x}_s) - \pi(x_s)\rangle \leq 0.$$

Dividing by $h$ which we take to 0, it becomes

$$\langle x_s - \pi(x_s), \partial\pi(x_s, \dot{x}_s)\rangle \leq 0. \tag{4.76}$$

Combining (4.75) and (4.76), we get

$$\langle x_s - \pi(x_s), \partial\pi(x_s, \dot{x}_s)\rangle = 0. \tag{4.77}$$

Applying Lemma 4.6.3-2. to $x = x_s$ and $d = h\dot{x}_s$, $h > 0$, we get

$$\langle \pi(x_s + h\dot{x}_s) - \pi(x_s), h\dot{x}_s\rangle \geq 0.$$

Dividing each side by $h^2$, we obtain

$$\left\langle \frac{\pi(x_s + h\dot{x}_s) - \pi(x_s)}{h}, \dot{x}_s \right\rangle \geq 0.$$

Taking $h \to 0$, we deduce

$$\langle \partial\pi(x_s, \dot{x}_s), \dot{x}_s\rangle \geq 0. \tag{4.78}$$

Plugging (4.77) and (4.78) in (4.74), we deduce

$$\begin{aligned}\dot{\phi}(s) = \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|\pi(x_s) - z_s\right\|^2 \\ + \eta a_s \left\langle \nabla f(x_s), z_s - x_s\right\rangle + \eta' b_s \left\langle z_s - \pi(x_s), x_s - z_s\right\rangle.\end{aligned} \tag{4.79}$$

Also, using Cauchy–Schwarz we have

$$\begin{aligned}\left\langle z_s - \pi(x_s), x_s - z_s\right\rangle &= \left\langle z_s - \pi(x_s), x_s - \pi(x_s) + \pi(x_s) - z_s\right\rangle \\ &= \left\langle z_s - \pi(x_s), x_s - \pi(x_s)\right\rangle - \left\|\pi(x_s) - z_s\right\|^2 \\ &\leq \frac{1}{2}\left\|x_s - \pi(x_s)\right\|^2 - \frac{1}{2}\left\|\pi(x_s) - z_s\right\|^2.\end{aligned}$$

Injected in (4.79), it becomes

$$\begin{aligned}\dot{\phi}(s) \leq \frac{da_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{db_s}{ds}\left\|\pi(x_s) - z_s\right\|^2 \\ + a_s\eta\left\langle \nabla f(x_s), z_s - x_s\right\rangle + \frac{b_s\eta'}{2}\left\|x_s - \pi(x_s)\right\|^2 - \frac{b_s\eta'}{2}\left\|\pi(x_s) - z_s\right\|^2.\end{aligned} \tag{4.80}$$

We now address the term $\mathbb{E}_\xi\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)$.

$$
\begin{aligned}
&\mathbb{E}_\xi \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) \\
&= \mathbb{E}_\xi[a_s(f(x_s - \gamma \nabla f(x_s,\xi)) - f(x_s))] \\
&+ \frac{b_s}{2}\mathbb{E}_\xi[\|z_s - \gamma' \nabla f(x_s,\xi) - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2 - \|z_s - \pi(x_s)\|^2] \\
&= a_s \mathbb{E}_\xi[f(x_s - \gamma \nabla f(x_s,\xi)) - f(x_s)] + \frac{b_s \gamma^{'2}}{2}\mathbb{E}_\xi[\|\nabla f(x_s,\xi)\|^2] \\
&+ \frac{b_s}{2}\mathbb{E}_\xi\left[\|z_s - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2 - \|z_s - \pi(x_s)\|^2\right] \\
&+ b_s \gamma' \mathbb{E}\left[\langle \nabla f(x_s,\xi), \pi(x_s - \gamma \nabla f(x_s,\xi)) - z_s\rangle\right] \\
&\leq a_s \mathbb{E}_\xi[f(x_s - \gamma \nabla f(x_s,\xi)) - f(x_s)] + \frac{b_s \gamma^{'2}}{2}\mathbb{E}_\xi[\|\nabla f(x_s,\xi)\|^2] + b_s\gamma' \langle \nabla f(x_s), \pi(x_s) - x_s\rangle \\
&+ b_s \gamma' \langle \nabla f(x_s), x_s - z_s\rangle + b_s\gamma' \mathbb{E}\left[\langle \nabla f(x_s,\xi), \pi(x_s - \gamma \nabla f(x_s,\xi)) - \pi(x_s)\rangle\right] \\
&+ \frac{b_s}{2}\mathbb{E}\left[\|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2\right] + b_s \mathbb{E}\left[\langle z_s - \pi(x_s), \pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\rangle\right] \\
&\leq a_s \gamma(\gamma\rho L - 2)\frac{1}{2}\|\nabla f(x_s)\|^2 + \frac{b_s \gamma^{'2}}{2}\mathbb{E}_\xi[\|\nabla f(x_s,\xi)\|^2] + b_s\gamma' \langle \nabla f(x_s), x_s - z_s\rangle \\
&- b_s\gamma'\left(f(x_s) - f^* + \frac{\mu}{2}\|x_s - \pi(x_s)\|^2\right) + b_s\gamma' \mathbb{E}\left[\langle \nabla f(x_s,\xi), \pi(x_s - \gamma \nabla f(x_s,\xi)) - \pi(x_s)\rangle\right] \\
&+ \frac{b_s}{2}\mathbb{E}\left[\|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2\right] + b_s \mathbb{E}\left[\langle z_s - \pi(x_s), \pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\rangle\right].
\end{aligned} \tag{4.81}
$$

where we used $\mathbb{E}_\xi[\nabla f(x_s,\xi)] = \nabla f(x_s)$, Lemma 4.2.1 and Assumption (SQC'$_\mu$). As $\arg\min f$ is convex, the projection $\pi(\cdot)$ is 1-Lipschitz, so

$$
\|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\| \leq \gamma \|\nabla f(x_s,\xi)\|. \tag{4.82}
$$

Using this 1-Lipschitz property of $\pi(\cdot)$ and Cauchy–Schwarz inequalities, we have

$$
\begin{aligned}
&b_s\gamma' \langle \nabla f(x_s,\xi), \pi(x_s - \gamma \nabla f(x_s,\xi)) - \pi(x_s)\rangle + \frac{b_s}{2}\|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2 \\
&+ b_s \langle z_s - \pi(x_s), \pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\rangle \\
&\leq b_s\gamma' \|\nabla f(x_s,\xi)\| \|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\| + \frac{b_s}{2}\|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\|^2 \\
&+ b_s \|z_s - \pi(x_s)\| \|\pi(x_s) - \pi(x_s - \gamma \nabla f(x_s,\xi))\| \\
&\leq b_s\gamma'\gamma \|\nabla f(x_s,\xi)\|^2 + \frac{b_s\gamma^2}{2}\|\nabla f(x_s,\xi)\|^2 + \frac{b_s\lambda}{2}\|z_s - \pi(x_s)\|^2 + \frac{b_s\gamma^2}{2\lambda}\|\nabla f(x_s,\xi)\|^2 \\
&= b_s\gamma\left(\gamma' + \frac{\gamma}{2} + \frac{\gamma}{2\lambda}\right)\|\nabla f(x_s,\xi)\|^2 + \frac{b_s\lambda}{2}\|z_s - \pi(x_s)\|^2
\end{aligned} \tag{4.83}
$$

for a parameter $\lambda > 0$, to be fixed later. Combining (4.80)-(4.81)-(4.83) and using Assumption (SGC$_\rho$), we get

$$
\begin{aligned}
&\dot{\phi}(s) + \mathbb{E}_\xi \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) \\
&\leq \left(\frac{da_s}{ds} - b_s\gamma'\right)(f(x_s) - f^*) && (4.84) \\
&+ \left(\frac{db_s}{ds} + b_s\lambda - b_s\eta'\right)\frac{1}{2}\|z_s - \pi(x_s)\|^2 && (4.85) \\
&+ (b_s\eta' - \mu b_s\gamma')\frac{1}{2}\|x_s - \pi(x_s)\|^2 && (4.86) \\
&+ (a_s\eta - b_s\gamma')\langle \nabla f(x_s), z_s - x_s\rangle && (4.87) \\
&+ (\rho b_s \gamma^{'2} + 2\rho b_s\gamma\left(\gamma' + \frac{\gamma}{2} + \frac{\gamma}{2\lambda}\right) + a_s\gamma(\gamma\rho L - 2))\frac{1}{2}\|\nabla f(x_s)\|^2. && (4.88)
\end{aligned}
$$

We choose $\eta$ of the form

$$\eta = \frac{\alpha}{\rho}\sqrt{\frac{\mu}{L}}.$$

We then cancel (4.84), (4.86) and (4.87)

$$\frac{da_s}{ds} - b_s\gamma' = 0, \quad b_s\eta' - \mu b_s\gamma' = 0, \quad a_s\eta - b_s\gamma' = 0,$$

which is ensured by choosing

$$a_s = e^{\eta s} = e^{\frac{\alpha}{\rho}\sqrt{\frac{\mu}{L}}s}, \quad \eta' = \mu\gamma', \quad b_s = \frac{\eta}{\gamma'}e^{\eta s} = \frac{\eta}{\gamma'}e^{\frac{\alpha}{\rho}\sqrt{\frac{\mu}{L}}s}.$$

We now want to cancel (4.85)

$$\frac{db_s}{ds} + b_s\lambda - b_s\eta' = 0 \Leftrightarrow \eta + \lambda - \eta' = 0 \Leftrightarrow \lambda = \eta' - \eta.$$

Fix now $\eta' = \frac{1}{\rho}\sqrt{\frac{\mu}{L}}$ and $\gamma = \frac{1}{\rho L}$, which implies

$$\gamma' = \frac{1}{\rho\sqrt{\mu L}}, \quad C_B = \alpha\mu, \quad \lambda = \frac{1-\alpha}{\rho}\sqrt{\frac{\mu}{L}}.$$

It remains to cancel (4.88), which we recall to be

$$\begin{aligned}
&\rho b_s {\gamma'}^2 + 2\rho b_s\gamma\left(\gamma' + \frac{\gamma}{2} + \frac{\gamma}{2\lambda}\right) + a_s\gamma(\rho\gamma L - 2)\\
&= a_s\left(\rho C_B{\gamma'}^2 + 2\rho C_B\gamma\left(\gamma' + \frac{\gamma}{2} + \frac{\gamma}{2\lambda}\right) + \gamma(\rho\gamma L - 2)\right)\\
&= a_s\left(\frac{\alpha}{\rho L} + \alpha\frac{\mu}{L}\left(\frac{2}{\rho\sqrt{\mu L}} + \frac{1}{\rho L} + \frac{1}{(1-\alpha)L}\sqrt{\frac{L}{\mu}}\right) - \frac{1}{\rho L}\right)\\
&= \frac{a_s}{\rho L}\left(\alpha - 1 + \rho\frac{\alpha}{(1-\alpha)}\sqrt{\frac{\mu}{L}} + 2\alpha\sqrt{\frac{\mu}{L}} + \alpha\frac{\mu}{L}\right).
\end{aligned} \tag{4.89}$$

The term $\frac{\alpha}{(1-\alpha)}\sqrt{\frac{\mu}{L}}$ requires a special care, as it grows with $\alpha$ getting closer to 1. Trying to make $\alpha - 1 + \frac{\alpha}{(1-\alpha)}\sqrt{\frac{\mu}{L}}$ non-positive, we see that a good choice of $\alpha$ is of the form $1 - c\left(\frac{\mu}{L}\right)^{\frac{1}{4}}$. Plugging this choice in (4.89)$\times\frac{\rho L}{a_s}$, we have

$$-c\left(\frac{\mu}{L}\right)^{\frac{1}{4}} + \rho\left(1 - c\left(\frac{\mu}{L}\right)^{\frac{1}{4}}\right)\frac{1}{c}\left(\frac{\mu}{L}\right)^{\frac{1}{4}} + 2\left(1 - c\left(\frac{\mu}{L}\right)^{\frac{1}{4}}\right)\sqrt{\frac{\mu}{L}} + \left(1 - c\left(\frac{\mu}{L}\right)^{\frac{1}{4}}\right)\frac{\mu}{L} \tag{4.90}$$

$$\leq \left(\frac{\mu}{L}\right)^{\frac{1}{4}}\left(-c + \frac{\rho}{c} - \rho\left(\frac{\mu}{L}\right)^{\frac{1}{4}} + 2\left(\frac{\mu}{L}\right)^{\frac{1}{4}} - 2c\sqrt{\frac{\mu}{L}} + \left(\frac{\mu}{L}\right)^{\frac{3}{4}} - c\frac{\mu}{L}\right). \tag{4.91}$$

As long as $c \geq \frac{1}{2}$, one has $-2c\sqrt{\frac{\mu}{L}} + \left(\frac{\mu}{L}\right)^{\frac{3}{4}} - c\frac{\mu}{L} \leq 0$ because $\frac{\mu}{L} \leq 1$. So, it suffices to ensure

$$-c + \frac{\rho}{c} + (2-\rho)\left(\frac{\mu}{L}\right)^{\frac{1}{4}} \leq 0. \tag{4.92}$$

One has $-c + \frac{\rho}{c} + (2-\rho)\left(\frac{\mu}{L}\right)^{\frac{1}{4}} \leq -c + \frac{\rho}{c} + 1$, such that the choice $c = \frac{1+\sqrt{1+4\rho}}{2}$ suffices to verify (4.92). So, we fix $\alpha = 1 - \left(\frac{1+\sqrt{1+4\rho}}{2}\right)\left(\frac{\mu}{L}\right)^{\frac{1}{4}}$, which is non-negative as long as

$$\frac{\mu}{L} \leq \left(\frac{1+\sqrt{1+4\rho}}{2}\right)^{-4}$$

In the deterministic gradient case, *i.e.* $\rho = 1$, this condition becomes $\frac{\mu}{L} \leq \left(\frac{1+\sqrt{1+4\rho}}{2}\right)^{-4} = \left(\frac{1+\sqrt{5}}{2}\right)^{-4} \approx 0.15$.
In conclusion, if $\gamma = \frac{1}{\rho L}$, $\gamma' = \frac{1}{\rho\sqrt{\mu L}}$, $\eta = \frac{1-\frac{1+\sqrt{1+4\rho}}{2}\left(\frac{\mu}{L}\right)^{1/4}}{\rho}\sqrt{\frac{\mu}{L}}$ $\eta' = \frac{1}{\rho}\sqrt{\frac{\mu}{L}}$ and $C_B = \left(1 - \frac{1+\sqrt{1+4\rho}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right)\mu$ we have

$$\forall s \in [0,t]\backslash\mathring{A}, \quad \dot{\phi}(s) + \mathbb{E}_\xi[(\varphi(x_s + G(x_s,\xi)) - \varphi(x_s))] \leq 0.$$

According to Proposition 4.5.1, we deduce

$$\phi(t) \leq \phi(0) + M_t,$$

where $(M_t)_{t\geq 0}$ is a martingale such that $\mathbb{E}[M_t] = 0$ for all $t \in \mathbb{R}_+$. This in turn implies

$$e^{\eta t}(f(x_t) - f^*) \leq \phi(t) \leq \phi(0) + M_t. \tag{4.93}$$

Evaluating the latter at $t = T_k$ and taking expectation, thanks to the stopping Theorem 4.3.1 we have

$$\begin{aligned}\mathbb{E}\left[e^{\eta T_k}(f(\tilde{x}_k) - f^*)\right] &= \mathbb{E}\left[e^{\left(1-\frac{1+\sqrt{1+4\rho}}{2})\left(\frac{\mu}{L}\right)^{1/4}\right)\frac{1}{\rho}\sqrt{\frac{\mu}{L}}T_k}(f(\tilde{x}_k) - f^*)\right] \\ &\leq f(x_0) - f^* + \mu\left(1 - \frac{1+\sqrt{1+4\rho}}{2}\left(\frac{\mu}{L}\right)^{1/4}\right)\|x_0 - \pi(x_0)\|^2.\end{aligned}$$

## 4.A Proof of Proposition 4.3.2.

The proof generalizes (Even et al., 2021), which addressed specific choices of parameters.

**Step 1** If we have

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt \\ dz_t &= \eta'(x_t - z_t)dt \end{cases} \tag{4.94}$$

integrating between $t_0$ and $t \geq t_0$, assuming $\eta' > -\eta$, we have

$$\begin{aligned} x(t) &= \frac{\frac{\eta'}{\eta}x(t_0) + z(t_0)}{\frac{\eta+\eta'}{\eta}} + \frac{x(t_0) - z(t_0)}{\frac{\eta+\eta'}{\eta}}e^{-(\eta+\eta')(t-t_0)} \\ &= x(t_0) + \frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(t-t_0)}\right)(z(t_0) - x(t_0)) \end{aligned}$$

$$\begin{aligned} z(t) &= \frac{x(t_0) + \frac{\eta}{\eta'}z(t_0)}{\frac{\eta+\eta'}{\eta'}} + \frac{z(t_0) - x(t_0)}{\frac{\eta+\eta'}{\eta'}}e^{-(\eta+\eta')(t-t_0)} \\ &= z(t_0) + \frac{\eta'}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(t-t_0)}\right)(x(t_0) - z(t_0)) \end{aligned}$$

**Step 2** Noting $t_0 = T_k$ and $t = T_{k+1^-}$, we have

$$\begin{aligned} \tilde{y}_k := x(T_{k+1^-}) &= (z(T_k) - x(T_k))\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + x(T_k) \\ &:= (\tilde{z}_k - \tilde{x}_k)\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + \tilde{x}_k \end{aligned}$$

Now, because of the gradient step at time $T_{k+1}$,

$$\begin{aligned} \tilde{x}_{k+1} &:= x(T_{k+1}) \\ &= (z(T_k) - x(T_k))\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + x(T_k) - \gamma\nabla f(x(T_{k+1}^-), \xi_{k+1}) \\ &= \tilde{y}_k - \gamma\nabla f(\tilde{y}_k, \xi_{k+1}) \end{aligned}$$

Also, as $\tilde{y}_k = (\tilde{z}_k - \tilde{x}_k)\alpha + \tilde{x}_k \Rightarrow \tilde{x}_k = \frac{\tilde{y}_k - \tilde{z}_k\alpha}{1-\alpha} = \tilde{z}_k + \frac{\tilde{y}_k - \tilde{z}_k}{1-\alpha}$, in our case $1 - \alpha = 1 - \frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) = \frac{1}{\eta+\eta'}\left(\eta e^{-(\eta+\eta')(T_{k+1}-T_k)} + \eta'\right)$, we deduce

$$\begin{aligned} z(T_{k+1^-}) &= z(T_k) + \frac{\eta'}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)(x(T_k) - z(T_k)) \\ &= z(T_k) + \frac{\eta'}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)\frac{(x(T_{k+1}^-) - z(T_k))}{\frac{1}{\eta+\eta'}\left(\eta e^{-(\eta+\eta')(T_{k+1}-T_k)} + \eta'\right)} \\ &= z(T_k) + \eta'\frac{\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta' + \eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(x(T_{k+1}^-) - z(T_k)). \end{aligned}$$

Finally,

$$\begin{aligned} \tilde{z}_{k+1} := z(T_{k+1}) &= z(T_k) + \eta'\frac{\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta' + \eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(x(T_{k+1}^-) - z(T_k)) - \gamma'\nabla f(x(T_{k+1}^-), \xi_{k+1}) \\ &= \tilde{z}_k + \eta'\frac{\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta' + \eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(\tilde{y}_k - \tilde{z}_k) - \gamma'\nabla f(\tilde{y}_k, \xi_{k+1}). \end{aligned}$$

Finally, the algorithm is

$$\left\{\begin{array}{ll}\tilde{y}_k & = (\tilde{z}_k - \tilde{x}_k)\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + \tilde{x}_k \\ \tilde{x}_{k+1} & = \tilde{y}_k - \gamma\nabla f(\tilde{y}_k, \xi_{k+1}) \\ \tilde{z}_{k+1} & = \tilde{z}_k + \eta'\frac{\left(1-e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta'+\eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(\tilde{y}_k - \tilde{z}_k) - \gamma'\nabla f(\tilde{y}_k, \xi_{k+1})\end{array}\right.$$

# 4.B Convergence Results Trajectory-Wise

## 4.B.1 Proof of Proposition 4.4.1

Before proving Proposition 4.4.1, we prove the Chernoff inequality for Gamma laws (Lemma 4.4.1). This is rather a classical result, but we provide a proof for the sake of completeness.

**Proof of Lemma 4.4.1** The Chernoff inequality is deduced by applying the Markov inequality to the right transformation of the random variable. Indeed for a non-negative random variable $X$, for $t < 0$ and $a > 0$ we have

$$\mathbb{P}(X \leq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \mathbb{E}\left[e^{tX}\right] e^{-ta} \leq \inf_{t<0} \mathbb{E}\left[e^{tX}\right] e^{-ta}. \tag{4.95}$$

The moment-generating function of a Gamma law with parameters $(n, 1)$ is well known, namely $\mathbb{E}\left[e^{tT_n}\right] = (1-t)^{-n}$, for $t < 1$. Applying (4.95) to $X = T_n$, $a = c\mathbb{E}[T_n] = cn$ with $c \leq 1$, we deduce

$$\begin{aligned}\mathbb{P}(T_n \leq cn) &\leq \inf_t (1-t)^{-n} e^{-cnt} \\ &= \inf_t e^{-n(ct+\log(1-t))}.\end{aligned}$$

The infimum is reached at $t = 1 - \frac{1}{c}$, such that

$$\mathbb{P}(T_n \leq cn) \leq e^{-n(c-1-\log(c))}.$$

**Proof of Proposition 4.4.1** By Markov's inequality and because $\mathbb{E}\left[e^{\beta T_k} Y\right] \leq K_0$, for some $C_0 > 0$ we have

$$\mathbb{P}\left(e^{\beta T_k} Y > C_0\right) \leq \frac{\mathbb{E}\left[e^{\beta T_k} Y\right]}{C_0} \leq \frac{K_0}{C_0}.$$

Set $C_0 := c_0 K_0$, $c_0 \geq 1$. With probability $1 - \frac{1}{c_0}$, we have

$$e^{\beta T_k} Y \leq c_0 K_0. \tag{4.96}$$

Using that $T_k \sim \Gamma(k, 1)$, we apply Lemma 4.4.1 (Chernoff inequality) with $c = c_1 \in (0, 1)$, such that

$$\mathbb{P}(T_k \leq c_1 k) \leq e^{-(c_1 - 1 - \log(c_1))k}.$$

We thus have

$$T_k \geq c_1 k \tag{4.97}$$

with probability at least $1 - e^{-(c_1-1-\log(c_1))k}$. Combining (4.96) and (4.97), and using that for two events $A$, $B$ we have $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, we deduce that with probability $1 - \frac{1}{c_0} - e^{-(c_1-1-\log(c_1))k}$

$$Y \leq c_0 K_0 e^{-\beta c_1 k}.$$

## 4.B.2 Proof of Proposition 4.4.2

Fix any $c \in (0, 1)$. Choose $\bar{c} \in (c, 1)$ and $\tilde{\varepsilon}' > 0$ such that $\tilde{\varepsilon}' < \beta(\bar{c} - c)$. For $k \geq 1$, define the events

$$A_k := \{T_k \leq \bar{c}k\}, \qquad B_k := \{e^{\beta T_k} Y_k \geq K_0 e^{\tilde{\varepsilon}' k}\}.$$

By assumption $\mathbb{E}\left[e^{\beta T_k} Y_k\right] \leq K_0$ and $Y_k \geq 0$. Markov's inequality and the Chernoff inequality (Lemma 4.4.1) imply, with $\Delta_{\bar{c}} := \bar{c} - 1 - \log(\bar{c}) > 0$, that

$$\mathbb{P}(A_k \cup B_k) \leq \mathbb{P}(A_k) + \mathbb{P}(B_k) \tag{4.98}$$

$$\leq e^{-k\Delta_{\bar{c}}} + e^{-\tilde{\varepsilon}' k}. \tag{4.99}$$

Summing (4.98) over $k \geq 1$ gives

$$\sum_{k\geq 1} \mathbb{P}(A_k \cup B_k) \leq \sum_{k\geq 1} \left(e^{-k\Delta_{\bar{c}}} + e^{-\tilde{\varepsilon}' k}\right) < +\infty. \tag{4.100}$$

By (4.100) and the Borel–Cantelli lemma, $\mathbb{P}\left(\liminf_{k\to\infty}(A_k^c \cap B_k^c)\right) = 1$. Thus, almost surely, there exists $K(\omega)$ such that for all $k \geq K(\omega)$, $T_k > \bar{c}k$ and $e^{\beta T_k} Y_k < K_0 e^{\tilde{\varepsilon}' k}$. Hence

$$0 \leq Y_k e^{\beta c k} < K_0 e^{-\left[\beta(\bar{c}-c) - \tilde{\varepsilon}'\right]k} \longrightarrow 0.$$

Since $c \in (0,1)$ was arbitrary, the claimed rate follows:

$$Y_k \overset{a.s.}{=} o\left(e^{-\beta c k}\right).$$

# 4.C Technical Lemmas

## 4.C.1 Proof of Lemma 4.2.1

By Assumption (L-smooth)

$$\begin{aligned} f(x - \gamma \nabla f(x,\xi)) &\leq f(x) + \langle \nabla f(x), x - \gamma \nabla f(x,\xi) - x\rangle + \frac{L}{2}\left\|x - \gamma \nabla f(x,\xi) - x\right\|^2 \\ &= f(x) - \gamma \left\langle \nabla f(x), \nabla f(x,\xi)\right\rangle + \frac{\gamma^2 L}{2}\left\|\nabla f(x,\xi)\right\|^2. \end{aligned} \tag{4.101}$$

We take expectation on both sides of (4.101)

$$\mathbb{E}\left[f(x - \gamma \nabla f(x,\xi)) - f(x)\right] \leq -\gamma \left\langle \nabla f(x), \mathbb{E}\left[\nabla f(x,\xi)\right]\right\rangle + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\nabla f(x,\xi)\right\|^2\right]. \tag{4.102}$$

By Assumption ($\mathrm{SGC}_\rho$), we have

$$\left\langle \nabla f(x), \mathbb{E}\left[\nabla f(x,\xi)\right]\right\rangle = \left\|\nabla f(x)\right\|^2, \quad \mathbb{E}\left[\left\|\nabla f(x,\xi)\right\|^2\right] \leq \rho \left\|\nabla f(x)\right\|^2. \tag{4.103}$$

Injecting (4.103) in (4.102), we have

$$\mathbb{E}\left[f(x - \gamma \nabla f(x,\xi)) - f(x)\right] \leq \gamma\left(\frac{L\rho}{2}\gamma - 1\right)\left\|\nabla f(x)\right\|^2.$$

## 4.C.2 Proof of Lemma 4.6.3

**Statement 1.** By definition of the orthogonal projection, for $x \in U$ and $y \in K$

$$\left\|x - P(x)\right\|^2 = \left\|x - y + y - P(x)\right\|^2 \leq \left\|x - y\right\|^2.$$

Developing the squared norm, we obtain

$$\left\|y - P(x)\right\|^2 + 2\left\langle x - y, y - P(x)\right\rangle \leq 0,$$

which gives the result.

**Statement 2.** For $x, d \in U$, we have

$$\begin{aligned}
&\langle P(x+d) - P(x), d\rangle \\
&= \langle P(x+d) - x, d\rangle - \langle P(x) - x, d\rangle \\
&= \frac{1}{2}\left[\|P(x+d) - x\|^2 + \|d\|^2 - \|P(x+d) - (x+d)\|^2\right. \\
&\left. - \|P(x) - x\|^2 - \|d\|^2 + \|P(x) - (x+d)\|^2\right] \\
&= \frac{1}{2}\left[\|P(x+d) - x\|^2 - \|P(x+d) - (x+d)\|^2 - \|P(x) - x\|^2 + \|P(x) - (x+d)\|^2\right].
\end{aligned}$$

Note that

$$\|P(x+d) - x\|^2 \geq \|P(x) - x\|^2, \quad \|P(x) - (x+d)\|^2 \geq \|P(x+d) - (x+d)\|^2,$$

such that

$$\langle P(x+d) - P(x), d\rangle \geq 0.$$

## 4.D Background

In this section, we give additional details about different formulations of momentum ODEs.

The Heavy Ball equation is defined by the following equation

$$\ddot{X}_t + \alpha(t)\dot{X}_t + \nabla f(X_t) = 0, \tag{HB}$$

where $\alpha(t) > 0$. Intuitively, it models a rolling object on the surface defined by the function, submitted to a friction which is parameterized by $\alpha(t)$. Another popular equation is the following (Alvarez et al., 2002)

$$\ddot{X}_t + \alpha(t)\dot{X}_t + b\nabla^2 f(X_t) + \nabla f(X_t) = 0. \tag{HB-Hess}$$

Importantly, (HB-Hess) describes more accurately the behavior of the Nesterov accelerated gradient algorithm than (HB), see (Shi et al., 2022). In section 4.3, we introduced the following equation

$$\begin{cases} \dot{x}_t &= \eta_t(z_t - x_t) - \gamma_t \nabla f(x_t) \\ \dot{z}_t &= \eta_t'(x_t - z_t) - \gamma_t' \nabla f(x_t) \end{cases} \tag{4.104}$$

with $\eta_t > 0$ for all $t \in \mathbb{R}_+$. This formulation is more general than (HB-Hess), as we show next.

**Proposition 4.D.1.** *Let $(x_t, z_t)_{t\geq 0}$ verify (4.104). Then, it verifies*

$$\ddot{x}_t + \left(\eta_t - \frac{\dot{\eta}_t}{\eta_t} + \eta_t'\right)\dot{x}_t + \gamma_t \nabla^2 f(x_t)\dot{x}_t + \left(\eta_t'\gamma_t + \eta_t\gamma_t' - \frac{\dot{\eta}_t}{\eta_t}\right)\nabla f(x_t) = 0. \tag{4.105}$$

*Proof.* We have

$$z_t = x_t + \frac{1}{\eta_t}\left(\dot{x}_t + \gamma_t \nabla f(x_t)\right) \tag{4.106}$$

$$\dot{z}_t = \dot{x}_t - \frac{\dot{\eta}_t}{\eta_t^2}\left(\dot{x}_t + \gamma_t \nabla f(x_t)\right) + \frac{1}{\eta_t}\left(\ddot{x}_t + \gamma_t \nabla^2 f(x_t)\dot{x}_t\right) \tag{4.107}$$

From this and the second line of (4.104)

$$\dot{x}_t - \frac{\dot{\eta}_t}{\eta_t^2}\left(\dot{x}_t + \gamma_t \nabla f(x_t)\right) + \frac{1}{\eta_t}\left(\ddot{x}_t + \gamma_t \nabla^2 f(x_t)\dot{x}_t\right) = \eta_t'(x_t - z_t) - \gamma_t' \nabla f(x_t) \tag{4.108}$$

$$= -\frac{\eta_t'}{\eta_t}\dot{x}_t - \frac{\eta_t'\gamma_t}{\eta_t}\nabla f(x_t) - \gamma_t' \nabla f(x_t). \tag{4.109}$$

We multiply by $\eta_t$ and rearrange, to get

$$\ddot{x}_t + \left(\eta_t - \frac{\dot{\eta}_t}{\eta_t} + \eta_t'\right)\dot{x}_t + \gamma_t \nabla^2 f(x_t)\dot{x}_t + \left(\eta_t'\gamma_t + \eta_t\gamma_t' - \frac{\dot{\eta}_t}{\eta_t}\right)\nabla f(x_t) = 0. \tag{4.110}$$

□

# Chapter 5

# Reconsideration of Acceleration for Strongly-Quasar Convex functions, Existing Flaws and Alternative

We recall that the content of this chapter is based on the following paper.

J. Hermant.
**Acceleration for Polyak-Łojasiewicz Functions with a Gradient Aiming Condition**.
*arXiv preprint*, 2026

## 5.1 Introduction

In previous chapters, we provided new results for the optimization of strongly quasar convex functions with Nesterov's Momentum. Those were presented as *accelerated* results. Indeed, in the $L$-smooth setting, minimizing these functions using momentum is often referred to as *accelerated optimization* (Hinder et al., 2020; Fu et al., 2023), suggesting better convergence bounds than gradient descent. This expectation partly stems from the fact that its complexity bounds take a form similar to those obtained in the smooth, strongly convex case. As Polyak-Łojasiewicz functions do not allow to provide acceleration with momentum (Yue et al., 2023), strong quasar convexity seems to be the right intermediate class, relaxing strong convexity while retaining its acceleration property. In this chapter, we are seriously questioning this claim.

### 5.1.1 Contributions

We are interested in what additional structure beyond the PL condition is required to obtain improved theoretical convergence bounds over gradient descent when using momentum. We summarize our contributions as follows:
**(1.)** We exhibit in Section 5.3 flaws of strong quasar convexity, and show that in contrast with existing interpretations, minimizing these functions with momentum does not necessarily yield better theoretical convergence bounds than gradient descent. We demonstrate that an acceleration phenomenon similar to strong convexity does not occur in general for this class.
**(2.)** We propose in Section 5.4 a different viewpoint by considering an *aiming condition*

$$\langle \nabla f(x), x - x^* \rangle \geq a \, \|\nabla f(x)\| \, \|x - x^*\| \,, \tag{5.1}$$

for $1 \geq a > 0$, which quantifies how well the descent direction aims toward a minimizer $x^*$. We show how (5.1) relates to strong-quasar convexity, and show that it helps to define a specific reparameterization of this latter class. In Section 5.5, we study the convergence of gradient descent, gradient flow, the Nesterov momentum algorithm and the Nesterov ODE for functions that are both PL and satisfy the aiming condition.

Using a stochastic parameterization of Nesterov momentum's algorithm, we show that we obtain accelerated bounds when the alignment constant $a$ is sufficiently large (Section 5.5.2).
**(3.)** We relax the aiming condition by requiring it to hold only on average along the optimization trajectory, recovering similar convergence rates (Section 5.6). It shows that acceleration can be achieved as long as (5.1) holds over a sufficiently large fraction of the optimization path, better reflecting the behavior of practical optimization algorithms.
**(4.)** We provide numerical experiments to support our theoretical insights. We exhibit a two-dimensional PL function with a unique minimizer that fails to satisfy (5.1), resulting in gradient descent to outperform momentum methods during the early stages of optimization (Section 5.7.1). We then provide an heuristic validation of our theoretical convergence bounds on the hard function for PL functions from (Yue et al., 2023), see Section 5.7.2.

### 5.1.2 Related Works on Aiming Conditions

Studying SGD, (Liu et al., 2023; Gupta and Wojtowytsch, 2025; Alimisis et al., 2025) use conditions close to quasar convexity they name *aiming condition*. While they consider the interpretation of pointing toward a minimizer, this condition actually carries more structure. Studying gradient descent, (Guille-Escuret et al., 2024; Singh et al., 2025) consider (5.1) from distinct perspectives: the first studies it empirically along the optimization path, the second uses it to prove a stability result, in the idea of better understanding the Edge of Stability phenomenon (Cohen et al., 2021).

## 5.2 Background

This chapter uses several algorithms/dynamical systems and functions classes. We restate them in this section.

### 5.2.1 Algorithms and Equations

We introduce the algorithms and dynamical systems that we study throughout this work.

**Gradient Descent and Gradient Flow** In this chapter, we consider gradient descent (GD)

$$\tilde{x}_{k+1} = \tilde{x}_k - \gamma \nabla f(\tilde{x}_k). \tag{GD}$$

We recall it is the building block of many important algorithms used in modern applications (Hinton, 2012; Ghadimi and Lan, 2013; Kingma and Ba, 2015; Romano et al., 2017), and is still an important object of study (Malitsky and Mishchenko, 2020; Cohen et al., 2021). Writing gradient descent as $\frac{x_{k+1}-x_k}{\gamma} = -\nabla f(x_k)$, setting the *Ansatz* $\tilde{x}_k \approx x_{k\gamma}$ for some smooth curve $(x_t)_{t\geq 0}$, we consider the ODE limit of (GD) as the stepsize $\gamma$ vanishes to zero, named gradient flow (GF)

$$\dot{x}_t = -\nabla f(x_t). \tag{GF}$$

In words, (GF) is a continuous version of (GD). This dynamical system viewpoint is longstanding for gaining insights into the corresponding algorithms (Ambrosio et al., 2005; Benaïm, 2006). See our discussion in Section 1.2.

**Nesterov's Momentum: Ode and Algorithm** We recall the formulation of Nesterov momentum (NM) we conder

$$\begin{cases} \tilde{y}_k &= \tilde{x}_k + \alpha_k(\tilde{z}_k - \tilde{x}_k), \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma \nabla f(\tilde{y}_k), \\ \tilde{z}_{k+1} &= \tilde{z}_k + \beta_k(\tilde{y}_k - \tilde{z}_k) - \gamma' \nabla f(\tilde{y}_k). \end{cases} \tag{NM}$$

Although its formulation is slightly more involved than Polyak's Heavy Ball, the latter exhibits worse convergence guarantees in some settings (Ghadimi et al., 2015; Goujaud et al., 2025). As for gradient descent,

Table 5.1: Comparison of convergence rates up to a numerical constant, for (GF), (NMO), (GD) and (NM) properly parameterized, for functions belonging to $\mathrm{SC}^\mu$, $\mathrm{SQC}^\mu_\tau$ and $\mathrm{PL}^\mu$, intersected with $\mathrm{LS}^L$ for algorithms. Most can be found in the literature, see Appendix 5.A.1.

| Function class \ Equation/Algorithm | GF | NMO | GD | NM |
|---|---|---|---|---|
| $\mu$-strongly convex ($\mathrm{SC}^\mu$) | $\mu$ | $\sqrt{\mu}$ | $\mu/L$ | $\sqrt{\mu/L}$ |
| $(\tau,\mu)$-strongly quasar convex ($\mathrm{SQC}^\mu_\tau$) | $\tau\mu$ | $\tau\sqrt{\mu}$ | $\tau\mu/L$ | $\tau\sqrt{\mu/L}$ |
| $\mu$-Polyak-Łojasiewicz ($\mathrm{PL}^\mu$) | $\mu$ | $(-)$ | $\mu/L$ | $\mu/L$ |

taking the limit of (NM) as stepsize $\gamma$ vanishes yields a limit Nesterov momentum ODE (NMO), (Kim and Yang, 2023, Appendix A.1)

$$\begin{cases} \dot{x}_t &= \eta_t(z_t - x_t) - \gamma\nabla f(x_t), \\ \dot{z}_t &= \eta'_t(x_t - z_t) - \gamma'\nabla f(x_t). \end{cases} \tag{NMO}$$

Since the strong connection between these ODEs and the corresponding algorithm has been established (Su et al., 2016), their study has gained a rising interest both in the convex (Siegel, 2019; Shi et al., 2022; Attouch et al., 2022; Aujol et al., 2023a; Li et al., 2024; Guo et al., 2024) and non-convex case (Okamura et al., 2024; Hermant et al., 2026; Gupta and Wojtowytsch, 2025; Renaud et al., 2026).

**Continuized Parameterization** We use the *continuized* method (Even et al., 2021), that we described in Chapter 4. As a reminder, it can be seen as a stochastic parameterization of (NM). For some $\eta, \eta' \in \mathbb{R}$ such that $\eta + \eta' > 0$, and some $\{T_k\}_{k\in\mathbb{N}}$ such that $T_{k+1} - T_k \overset{i.i.d}{\sim} \mathcal{E}(1)$, we fix in (NM):

$$\alpha_k = \frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right),$$
$$\beta_k = \eta'\frac{\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta' + \eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}.$$

With this parameterization, the algorithm is inherently stochastic, yielding convergence guarantees holding with high probability. In the case of minimizing smooth strongly-quasar convex functions, the best known bounds are achieved using this parameterization (Wang and Wibisono, 2023). In contrast, more classical deterministic parameterization achieves it only if assuming more structure (Chapter 3), or when using search-procedures to compute parameters (Hinder et al., 2020; Nesterov et al., 2021; Guminov et al., 2023), see details in Appendix 5.A.1.

### 5.2.2 Classical setting for acceleration

In this work we consider several different classes of functions. To help the presentation, we will recall each class and attribute them a specific notation. First, we recall a geometrical setting in which momentum-based acceleration is long established.

**Definition 5.2.1.** *For $L > 0$, we call $LS^L$ the set of $L$-smooth functions, namely $f$ such that $\forall x, y \in \mathbb{R}^d$, $f(x) - f(y) - \langle\nabla f(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2$.*

**Definition 5.2.2.** *For $\mu > 0$, we call $SC^\mu$ the set of $\mu$-strongly convex functions, namely $f$ such that $\forall x, y \in \mathbb{R}^d$, $f(x) + \langle\nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2 \leq f(y)$.*

Intuitively, $L$-smoothness and $\mu$-strong convexity provide global upper and lower bounds on the curvature of $f$. For a $C^2$ function, $f \in \mathrm{SC}^\mu \cap \mathrm{LS}^L$ implies that any eigenvalue $\lambda(x)$ of $\nabla^2 f(x)$ is such that $\mu \leq \lambda(x) \leq L$, for any $x \in \mathbb{R}^d$. The ratio $L/\mu \geq 1$ is called *condition number*, and characterizes how ill-conditioned the minimization problem is—a higher condition number typically reflecting a more difficult problem. $\mathrm{SC}^\mu \cap \mathrm{LS}^L$ is a favorable setting for minimization: there exists a unique critical point $x^*$, satisfying $f(x^*) = f^*$, and linear convergence of function values is ensured by first-order methods, namely

$$\log(f(x_t) - f^*) \leq -\Lambda t + \log(K(x_0)),$$

with $\Lambda > 0$ being the **convergence rate**, $K(x_0) > 0$ depends on the initialization. Linear convergence also holds in the discrete case, *i.e.*, replacing $(x_t)_{t\geq 0}$ by $\{x_n\}_{n\geq 0}$ and $t$ by $k$. In this work, we specifically focus on the convergence rate value $\Lambda$. In the $\mathrm{SC}^\mu \cap \mathrm{LS}^L$ regime, the improvement of (NM) over (GD) is well established (Nesterov, 2013), see the convergence rates in Table 5.1. Precisely, as typically $L/\mu \gg 1$, improving from $\mu/L$ to $\sqrt{\mu/L}$ yields a substantial gain, often called *acceleration*. However, many problems of interest do not fall within $\mathrm{SC}^\mu \cap \mathrm{LS}^L$. This motivates to identify more general settings where momentum-based acceleration remains possible.

### 5.2.3 Relaxed setting

A first relaxation of $\mathrm{SC}^\mu$ and $\mathrm{LS}^L$ consists in lower and upper global quadratic bounds of the function.

**Definition 5.2.3.** *For $\mu_0, L_0 > 0$, we call: $QG_-^{\mu_0}$ the set of functions $f$ that verify $f(x) - f^* \geq \frac{\mu_0}{2}\|x - x^*\|^2$, $QG_+^{L_0}$ the set of functions $f$ that verify $f(x) - f^* \leq \frac{L_0}{2}\|x - x^*\|^2$.*

$\mathrm{QG}_-^{\mu_0} \cap \mathrm{QG}_+^{L_0}$ includes functions with pathological behavior, *e.g.* having an infinite amount of critical points in any vicinity of $x^*$, see Chapter 3. This may discard the possibility of targeting a global minimum $f^*$. A key relaxation of strong convexity that does not preclude this objective is the class of Polyak-Łojasiewicz (PL) functions (Polyak, 1963; Łojasiewicz, 1963).

**Definition 5.2.4.** *For $\mu > 0$, we call $PL^\mu$ the set of functions $f$ that satisfies $\forall x \in \mathbb{R}^d$, $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$.*

It immediately follows that for $f \in \mathrm{PL}^\mu$, any critical point is a global minimizer of $f$. From an optimization perspective, two key results hold for the class $\mathrm{PL}^\mu \cap \mathrm{LS}^L$: (i) gradient descent achieves a linear convergence rate (Polyak, 1963), and (ii) this rate is optimal among first-order algorithms (Yue et al., 2023). As a consequence, (NM) cannot yield any significant theoretical improvement over this class, see Table 5.1. This motivates studying more structured non-convex relaxations of $\mathrm{SC}^\mu \cap \mathrm{LS}^L$. The class of strongly quasar convex functions (SQC) is expected to fulfill the demand.

**Definition 5.2.5.** *For $(\tau, \mu) \in (0, 1] \times \mathbb{R}_+^*$, we call $SQC_\tau^\mu$ the set of functions $f$ that verify $\forall x \in \mathbb{R}^d$, $f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2$.*

$\mathrm{SQC}_\tau^\mu$ is a relaxation of $\mathrm{SC}^\mu$, while being more structured than $\mathrm{PL}^\mu$, see (5.2). One possible application of these functions is the training of generalized linear models (Wang and Wibisono, 2023, Section 3). Compared to $\mathrm{PL}^\mu$, an important supplementary structure of $\mathrm{SQC}_\tau^\mu$ is to keep the property of $\mathrm{SC}^\mu$ to having a unique minimizer, although it can be relaxed, see (Gupta and Wojtowytsch, 2025) or Chapter 4. These functions can exhibit can exhibit strong oscillations, see Figure 5.6. Unlike PL functions, this class admits distinct convergence rates for (GD) and (NM) when intersected with $\mathrm{LS}^L$, see Table 5.1.

**Relation between classes** The following inclusions hold, see (Karimi et al., 2016) and Chapter 3),

$$\mathrm{SC}^\mu \subset \mathrm{SQC}_1^\mu \subset \mathrm{SQC}_\tau^\mu \subset \mathrm{PL}^{\mu\tau^2} \subset \mathrm{QG}_-^{\mu\tau^2}, \quad \mathrm{LS}^L \subset \mathrm{QG}_+^L. \tag{5.2}$$

As observed by (Guille-Escuret et al., 2021), the constants inherited from these inclusions may be severely suboptimal. For instance, oscillations of the functions outside a vicinity of minimizers may affect significantly the $\mathrm{LS}^L$ constant but not the $\mathrm{QG}_+^{L_0}$ constant, such that we could have that $f \in \mathrm{LS}^L \cap \mathrm{QG}_+^{L_0}$ with $L_0 \ll L$. Similarly, a sudden drop of the gradient norm may affect the $\mathrm{PL}^\mu$ constant while letting the $\mathrm{QG}_-^{\mu_0}$ almost untouched, such that $f \in \mathrm{PL}^\mu \cap \mathrm{QG}_-^{\mu_0}$ while $\mu_0 \gg \mu$. This induces that comparing convergences rates across different classes must be handled with care when using same constants for different classes, $\mu$ or $L$ classically.

## 5.3 Acceleration under Strong Quasar Convexity: Hidden Pitfalls

Using momentum to optimize functions in $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$ has been described as accelerated optimization in the literature (Hinder et al., 2020; Fu et al., 2023), and by ourselves in previous chapter. If so, this would place this class as an interesting intermediate class between $\mathrm{SC}^\mu \cap \mathrm{LS}^L$ and $\mathrm{PL}^\mu \cap \mathrm{LS}^L$: retaining the accelerated convergence bounds of the first, while allowing non-convex behavior as the second does. The interpretation of retaining accelerated bounds of $\mathrm{SC}^\mu \cap \mathrm{LS}^L$ is motivated by an analogy between complexity bounds. We show that performing such analogy is misleading.

**A hard PL instance is strongly quasar convex** (Yue et al., 2023) introduces a *hard function* belonging to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$ such that that the number of needed gradient calls needed to obtain a point $\tilde{x}$ ensuring $f(\tilde{x}) - f^* \leq \varepsilon$ scales at best as $\mathcal{O}(\frac{L}{\mu}\log(\frac{1}{\varepsilon}))$ for any first-order algorithms. This bound corresponds to a $\mu/L$ convergence rate, inducing that, up to numerical constants, the convergence rate of (GD) is optimal, see Table 5.1. Perhaps surprisingly, this hard function satisfies some strong regularity, for instance it has a unique global minimizer. More interesting in our case, it is actually strongly-quasar convex.

**Proposition 5.3.1.** *The hard function belonging to $PL^\mu \cap LS^L$ from (Yue et al., 2023), for which the complexity bound of* (GD) *can be improved at most by a numerical factor, also belongs to $SQC_\tau^{\mu'} \cap LS^L$ for some parameters $(\tau, \mu') \in (0,1] \times \mathbb{R}_+^*$.*

The proof is essentially a collection of known result, see Appendix 5.B.3. This proposition definitely shows that an acceleration of the same nature than for $\mathrm{SC}^\mu \cap \mathrm{LS}^L$ is not possible in general when considering $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$. Indeed, in the case of $\mathrm{SC}^\mu \cap \mathrm{LS}^L$, the ratio of the convergence rates between (NM) and (GD) is $\sqrt{\mu/L}/(\mu/L) = \sqrt{L/\mu}$, which goes to infinity with the conditioning $L/\mu$ going to infinity. In particular, if we fix $L$, this shows that for this class, when taking $\mu \to 0$ the bound of (NM) is arbitrarily better compared with (GD). Such behavior is not possible for the function from (Yue et al., 2023).

**Weaker Assumption, Sharper Rate** To compare the convergence rates of (GD) and (NM) when minimizing functions in $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$, it is natural to rely on the rates established in the literature for this class, namely $\tau\mu/L$ for (GD) and $\tau\sqrt{\mu/L}$ for (NM), see Table 5.1. However, any function belonging to $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$ also belongs to $\mathrm{PL}^{\mu'} \cap \mathrm{LS}^L$, for some parameter $\mu'$. As a result, one may alternatively consider the convergence rate $\mu'/L$ of gradient descent derived for functions in $\mathrm{PL}^{\mu'} \cap \mathrm{LS}^L$. Strikingly, in some cases, this latter rate **can be sharper** than those given for the class $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$. In Figure 5.1, we plot two examples of functions that belong to $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$, and so also belong to $\mathrm{PL}^{\mu'} \cap \mathrm{LS}^L$, for some parameters. For a given pair of algorithm and function class, we numerically compute the parameters of this class such that the associated convergence rate is maximized. For the first example, we find that the rate of (NM) characterized by $\mathrm{SQC}_\tau^\mu$ is $\approx 10^2$ times larger than the one of (GD) characterized by $\mathrm{PL}^{\mu'}$. However, for the second example, we find that the rate of (GD) characterized by $\mathrm{PL}^{\mu'}$ is $\approx 10^2$ times better than the rate of (NM) characterized by $\mathrm{SQC}_\tau^\mu$. Thus, the worst-case bounds of (GD) and (NM) under $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$ may be improved if considering the worst-case bound of (GD) under $\mathrm{PL}^\mu \cap \mathrm{LS}^L$.

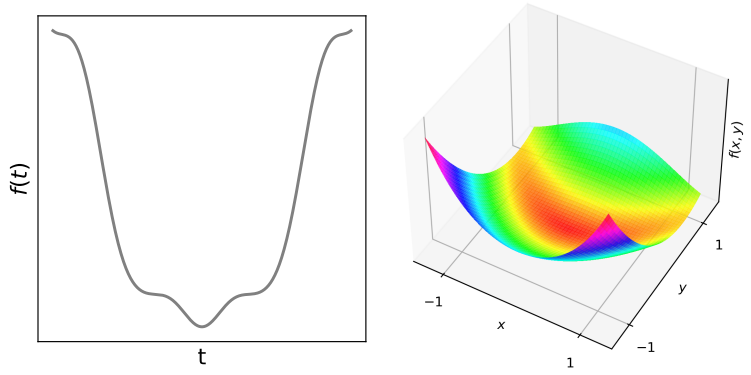
These findings indicate that our interpretation of strong quasar convexity yielding accelerated convergence bound using (NM) is lacking. In fact, this class suffers from some problems, as we discuss in the next section.

### 5.3.1 Problems of Strong Quasar Convexity

We provide below some arguments on why we should treat with special care convergence bounds involving $\mathrm{SQC}_\tau^\mu$.

**Parameter Interpretation** The constant $\mu$ in $\mathrm{SC}^\mu$ parametrizes a global lower bound on the curvature. Reusing the notation $\mu$ in a similar way in $\mathrm{SQC}_\tau^\mu$ may therefore suggest a similar interpretation. In particular, for $f \in \mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$, one may be tempted to view $L/\mu$ as a conditioning measurement, analogously to the setting $\mathrm{SC}^\mu \cap \mathrm{LS}^L$. Under this interpretation, one would expect $L/\mu \geq 1$, see Section 5.2.2. However, since $f \in \mathrm{SQC}_\tau^\mu \Rightarrow f \in \mathrm{SQC}_{\theta\tau}^{\mu/\theta}$, for any $\theta \in (0,1]$ (Hinder et al., 2020, Observation 5), it is possible to have $f \in \mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$ for some parameters satisfying $\mu > L$. This observation challenges both the condition-number interpretation of $L/\mu$ and calls into question the interpretation of the convergence rates in this setting.

**The Optimal Choice of Parameters** For many function classes, there exists a natural notion of optimal parameters. For instance, if $f \in \mathrm{PL}^\mu$ for a range of constants $\mu > 0$, the tighter class to describe $f$ is $\mathrm{PL}^{\hat{\mu}}$ with $\hat{\mu}$ the largest admissible value. Accordingly, the rates reported in Table 5.1 are maximized by choosing $\mu = \hat{\mu}$. In contrast, because of the double parameterization, there is no canonical way to choose among all admissible pairs $(\tau, \mu)$ an optimal one for a function in $\mathrm{SQC}_\tau^\mu$. Choosing the largest admissible $\tau$ and then the largest admissible $\mu$ leads to a different parameter pair than optimizing in the reverse order, see




| Algorithm (class) | $\Lambda$ (left plot) | $\Lambda$ (right plot) |
|---|---|---|
| GD ($\mathrm{PL}^{\mu}$) | $3.2 \cdot 10^{-4}$ | $1.6 \cdot 10^{-2}$ |
| GD ($\mathrm{SQC}_{\tau}^{\mu}$) | $1.3 \cdot 10^{-2}$ | $7.7 \cdot 10^{-5}$ |
| NM ($\mathrm{SQC}_{\tau}^{\mu}$) | $1.8 \cdot 10^{-2}$ | $1.8 \cdot 10^{-4}$ |

Figure 5.1: Left: graph of $f(t) = 5(t + 0.19\sin(5t))^2$, minimizer is 0. Right: graph of $f(x, y) = 0.5(0.5x^2 - y)^2 + 0.05x^2$, minimizer is $(0, 0)$. Bottom: table giving for each function the numerical value of the theoretical convergence rate for (GD) or (NM), precising what class of function is used to characterize the bound (see Table 5.1), where $\Lambda$ is the convergence rates. See implementation details in Appendix 5.B.1. It shows that depending on the functions, $\mathrm{PL}^{\mu}$-based bound may be sharper than $\mathrm{SQC}_{\tau}^{\mu}$-based bounds, and conversely.

Appendix 5.A.2. More critically, for a given function $f$, among the admissible pair $(\tau, \mu)$ ensuring $f \in \mathrm{SQC}_{\tau}^{\mu}$, the one that maximizes a given convergence rate depends on the algorithm or dynamic under consideration, as illustrated in Figure 5.2. As a consequence, statements of the form *"the rates of* (GF) *and* (NMO) *scale as* $\tau\mu$ *and* $\tau\sqrt{\mu}$*"* without clarifying how the parameters are chosen may result in misleading comparisons.

### 5.3.2 Discussion

We note that the problems raised in Section 5.3.1 can be solve if splitting $\mathrm{SQC}_{\tau}^{\mu}$ into two assumptions, *i.e.* assuming $f \in \mathrm{QG}_{-}^{\mu_0}$ and that it is quasar convex, namely $\exists \tau \in (0, 1]$ such that $f^* \geq f(x) + \frac{1}{\tau}\langle \nabla f(x), x^* - x \rangle$. However, the other points remain. The problem stems in the fact that convergence rates deduced under $\mathrm{PL}^{\mu}$ are not directly comparable to those deduce for classes such as $\mathrm{SQC}_{\tau}^{\mu}$, as the underlying geometrical constants encompasses different information. Our idea in the remaining of the paper is then to keep working with the $\mathrm{PL}^{\mu}$ class, and pairing it with an additional and intuitive geometrical information.

## 5.4 Aiming condition as an Alternative Characterization of Convergence

We introduce a class of function that we will further intersect with $\mathrm{PL}^{\mu}$. We assume that there exists a unique global minimizer $x^*$. We initially assume the existence of a unique minimizer $x^*$. We note that as discussed in Section 5.3, functions that serve as lower bound for $\mathrm{PL}^{\mu} \cap \mathrm{LS}^{L}$ are designed in this setting, such that it remains an interesting setting to discuss acceleration. Anyway this assumption can be relaxed, as discussed in Section 5.6

**Definition 5.4.1.** *For* $a \in (0, 1]$*, we call* $AC^a$ *the set of functions satisfying the aiming condition, i.e.* $\forall x \in \mathbb{R}^d,\ \langle \nabla f(x), x - x^* \rangle \geq a \|\nabla f(x)\| \|x - x^*\|$.

The parameter $a$ measures the alignment between the descent direction $-\nabla f$ and the direction toward the minimizer. Values of $a$ close to 1 indicate a strong alignment. In one dimension, any function in $\mathrm{AC}^a$ has perfect alignment, yielding $a = 1$. Such strong alignment is particularly favorable to momentum methods, as it ensures that the accumulated momentum reinforces progress toward $x^*$. In contrast, if $a$ is small, momentum may amplify steps taken in poorly aligned directions. For example, the rightmost function in

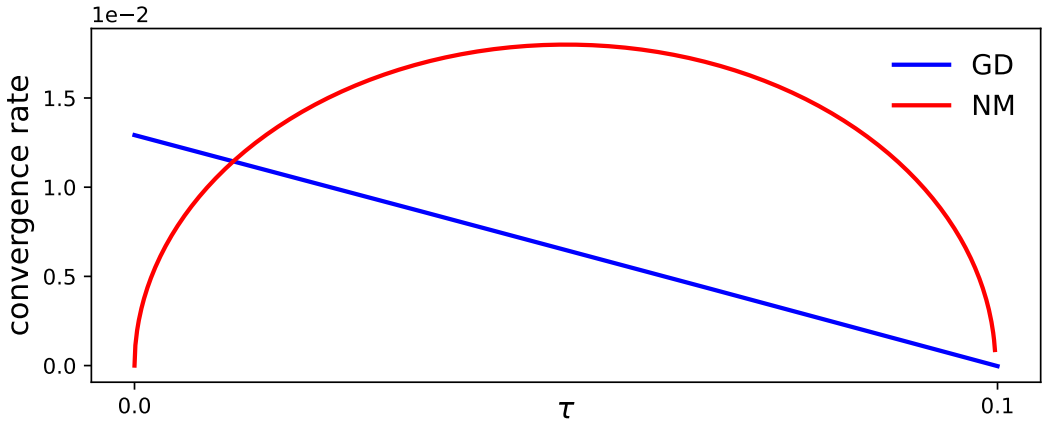



Figure 5.2: For a range of parameters $\tau$, we numerically compute the highest admissible $\mu$ ensuring $f \in \mathrm{SQC}_\tau^\mu$, and plot the numerical values of the associated theoretical convergence rates for (GD) and (NM) under $\mathrm{SQC}_\tau^\mu \cap \mathrm{LS}^L$, namely $\tau\mu/L$ and $\tau\sqrt{\mu/L}$, with $f(t) = 5(t + 0.19\sin(5t))^2$. See implementation details in Appendix 5.B.1. Note that the pair $(\tau, \mu)$ that maximizes the convergence rate of (GD) differs from the pair that maximizes the one of (NM).

Figure 5.1 exhibits a small aiming condition value, $a \approx 3 \cdot 10^{-4}$, which helps to explain why the convergence rate of (GD) is sharper than that of (NM) in this case. We discuss the important connection between $\mathrm{AC}^a$ and $\mathrm{SQC}_\tau^\mu$ in Appendix 5.4.1. Before, we provide a toy construction that belongs to $\mathrm{AC}^a$.

**A toy example** Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $f(0) = 0$ and $f'(t) > 0$ for all $t > 0$. Let $g : \mathbb{R}^d \to \mathbb{R}$, such that $g \geq 1$. Define $h : \mathbb{R}^d \to \mathbb{R}$

$$h(x) = f(\|x\|)g\left(\frac{x}{\|x\|}\right). \tag{5.3}$$

$f$ describes the radial movement, while $g$ describes the angular oscillations. Such constructions were introduced as quasar convex examples (Lee and Valiant, 2016). In particular, $f \in \mathrm{SQC}_\tau^\mu$ implies $h \in \mathrm{SQC}_\tau^\mu$, see Chapter 3. The more $g$ oscillates, the lower is $a$. This is formalized with the following result.

**Proposition 5.4.1.** *Let $h$ defined as* (5.3)*. Then, $h \in AC^a$ with*

$$a = \inf_{r>0, u \in \mathbb{S}^{d-1}} \frac{1}{\sqrt{1 + \left(\frac{f(r)}{rf'(r)}\right)^2 \left(\frac{\|\nabla_{\mathbb{S}} g(u)\|}{g(u)}\right)^2}},$$

*where $\nabla_{\mathbb{S}} g(u)$ is the spherical gradient of $g$ restricted to $\mathbb{S}^{d-1}$.*

The spherical gradient $\nabla_{\mathbb{S}} g(u)$ encodes the angular variation. If $g$ is constant or has no angular variation, *e.g.* $g(x) = \|x\|^2$, then $\nabla_{\mathbb{S}} g(u) = 0$. In this case $h$ is a radial function and it follows that $h \in \mathrm{AC}^1$. Supplementary structure on $f$–such as belonging to $\mathrm{PL}^\mu$ or $\mathrm{QG}_-^{\mu_0}$–will transfer to $h$. However it appears clearly in our result that it will also impact the value of $a$ through the ratio $f(r)/(rf'(r))$.

### 5.4.1 Link between Aiming Condition and Strong Quasar Convexity

In this section, we investigate the link between $\mathrm{AC}^a \cap \mathrm{PL}^\mu$ and relaxations of strong convexity.

**Lemma 5.4.1.** *Let $f \in PL^\mu \cap AC^a \cap QG_+^{L_0}$. Let $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$. Then, for all $x \in \mathbb{R}^d$,*

$$f^* \geq f(x) + \frac{1}{a}\sqrt{\frac{L_0}{\mu}}\langle \nabla f(x), x^* - x\rangle + \frac{\sqrt{\mu\mu_0}}{2}\|x - x^*\|^2.$$

*That is, $f \in SQC_{a\sqrt{\mu/L_0}}^{\sqrt{\mu\mu_0}}$.*

*Proof.* If $x = x^*$, the inequality holds trivially. Assume $x \neq x^*$.

**Step 1 (AC + PL).** Since $f \in \mathrm{AC}^a$, we have

$$\langle \nabla f(x), x - x^* \rangle \geq a \|\nabla f(x)\| \|x - x^*\|, \quad \text{hence} \quad \langle \nabla f(x), x^* - x \rangle \leq -a \|\nabla f(x)\| \|x - x^*\|.$$

Since $f \in \mathrm{PL}^\mu$, we also have $\|\nabla f(x)\| \geq \sqrt{2\mu}\sqrt{f(x) - f^*}$, so

$$\langle \nabla f(x), x^* - x \rangle \leq -a\sqrt{2\mu}\sqrt{f(x) - f^*}\, \|x - x^*\|. \tag{5.4}$$

**Step 2 (use $\mathrm{QG}_+^{L_0}$).** Since $f \in \mathrm{QG}_+^{L_0}$,

$$f(x) - f^* \leq \frac{L_0}{2}\|x - x^*\|^2 \quad \Rightarrow \quad \|x - x^*\| \geq \sqrt{\frac{2}{L_0}}\sqrt{f(x) - f^*}.$$

Plugging this into (5.4) yields

$$\langle \nabla f(x), x^* - x \rangle \leq -a\sqrt{2\mu}\sqrt{\frac{2}{L_0}}\,(f(x) - f^*) = -2a\sqrt{\frac{\mu}{L_0}}\,(f(x) - f^*). \tag{5.5}$$

**Step 3 (use $\mathrm{QG}_-^{\mu_0}$).** By definition of $\mu_0$, we have

$$f(x) - f^* \geq \frac{\mu_0}{2}\|x - x^*\|^2 \quad \Rightarrow \quad \sqrt{f(x) - f^*} \geq \sqrt{\frac{\mu_0}{2}}\|x - x^*\|.$$

Plugging this into (5.4) yields

$$\langle \nabla f(x), x^* - x \rangle \leq -a\sqrt{2\mu}\sqrt{\frac{\mu_0}{2}}\|x - x^*\|^2 = -a\sqrt{\mu\mu_0}\,\|x - x^*\|^2. \tag{5.6}$$

**Step 4 (combine the two bounds).** From (5.5) and (5.6), we have

$$\langle \nabla f(x), x^* - x \rangle \leq -\max\Big\{2a\sqrt{\frac{\mu}{L_0}}\,(f(x) - f^*),\ a\sqrt{\mu\mu_0}\,\|x - x^*\|^2\Big\}.$$

Since $\max\{u, v\} \geq \frac{u+v}{2}$ for $u, v \geq 0$, we obtain

$$\langle \nabla f(x), x^* - x \rangle \leq -\frac{1}{2}\Big(2a\sqrt{\frac{\mu}{L_0}}\,(f(x) - f^*) + a\sqrt{\mu\mu_0}\,\|x - x^*\|^2\Big),$$

which is the claim. □

This lemma shows that $\mathrm{PL}^\mu \cap \mathrm{AC}^a \cap \mathrm{QG}_+^{L_0}$ induces a specific parameterization of strong-quasar convexity. We note that this parameterization does not suffer from the problem raised in Section 5.3.1. Importantly, it also follows that similar convergence analysis ideas can be used for $\mathrm{PL}^\mu \cap \mathrm{AC}^a \cap \mathrm{QG}_+^{L_0}$ and $\mathrm{SQC}_\tau^\mu$. Conversely, if assuming some smoothness regularity, strong-quasar convexity implies the aiming condition.

**Proposition 5.4.2.** *Let $f \in SQC_\tau^\mu$. If $f$ also satisfies*

$$\forall x \in \mathbb{R}^d,\ \|\nabla f(x)\| \leq L\|x - x^*\|, \tag{5.7}$$

*then $f \in AC^{\frac{2\mu}{(2-\tau)L}}$.*

See the proof in Appendix 5.G.1.

## 5.5 Acceleration for PL Functions with an Aiming Condition

In Section 5.5.1, we observe the impact of the aiming condition when studying standard gradient methods (GF) and (GD). We consider in Section 5.5.2 the convergence of (NMO) and (NM), and the question of acceleration.

### 5.5.1 $\mathrm{AC}^a$ Gives Weight to the Local Information at $x^*$

We first compare convergence bounds obtained for (GF), depending on $f$ belongs to $\mathrm{AC}^a$ or not.

**Theorem 5.5.1.** *Let $(x_t)_{t\geq 0}$ follow* (GF).
*(i) (Polyak and Shcherbakov, 2017, Theorem 3) If $f \in PL^\mu$, then*

$$f(x_t) - f^* \leq (f(x_0) - f^*)e^{-2\mu t}.$$

*(ii) If $f \in PL^\mu \cap AC^a \cap QG_+^{L_0}$, then*

$$f(x_t) - f^* \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(x_0) - f^*)e^{-a\sqrt{\mu\mu_0}t},$$

*where $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$.*

See the proof in Appendix 5.D.1. The assumption $f \in \mathrm{QG}_+^{L_0}$ for bound (ii) is fairly mild. When considering algorithms, we will assume $f \in \mathrm{LS}^L$, such that they will naturally belong to $\mathrm{QG}_+^{L_0}$. We note that under the aiming condition, there is a worse constant factor $1 + \sqrt{L_0/\mu_0}$, which mainly affects the bound for early iterations. In this work, we specifically focus on convergence rates.

**Comparison of Convergence Rates** In addition to the PL constant $\mu$, the convergence rates in (ii) depends on a quadratic lower bound parameter $\mu_0$ and the aiming condition parameter $a$. As a result, rate (ii) improves upon rate (i) whenever $a \geq 2\sqrt{\mu/\mu_0}$. So, for functions with bad alignment, characterizing the convergence rate solely through the PL condition may lead to sharper rates. It is interesting to observe that $\mu_0$ appears in (ii), but not in (i). In the presence of flat plateaus, we may have $\mu \ll \mu_0$. This is due to $\mu_0$ being less affected by global irregularities of the function landscape, as we discuss next.
**The Aiming Condition Exploits the Local Information at $x^*$** For functions in $\mathrm{SC}^\mu \cap \mathrm{LS}^L$, the slowest direction toward the minimizer is the one with the smallest curvature $\mu$, since it corresponds to the flattest direction (left plot of Figure 5.8 in the appendix). For functions in $\mathrm{PL}^\mu \cap \mathrm{LS}^L$, the situation is less clear: the trajectory may approach the minimizer either through low-curvature directions near $x^*$ or through plateau-like regions (right plot of Figure 5.8). Thus, it is ambiguous whether the relevant $\mu$ is determined by local curvature at $x^*$ or by the plateaus, as $\mathrm{PL}^\mu$ handles both information. Bound (i) of Theorem 5.5.2 does not separate these effects, but bound (ii) does: it reveals a trade-off between global plateaus and local curvature $\mu_0$. In the special case $a = 1$, if $\mu = \mu_0$, flatness near $x^*$ dominates and we recover the rate $\mathcal{O}(e^{-\mu t})$. If $\mu < \mu_0$, the plateaus matter, yielding a rate $e^{-\sqrt{\mu\mu_0}}$, which shows that plateaus do not fully determine the convergence.

**Convergence of** (GD) Assuming the function belongs to $\mathrm{LS}^L$, we deduce similar convergence rates for (GD), with the difference that the $L$-smoothness constant intervenes.

**Theorem 5.5.2.** *Let $f \in LS^L$, $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ follow* (GD), *$\gamma = \frac{1}{L}$.*
*(i) (Karimi et al., 2016, Theorem 1) If we also have $f \in PL^\mu$, then*

$$f(\tilde{x}_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(\tilde{x}_0) - f^*).$$

*(ii) If we also have $f \in PL^\mu \cap AC^a$, then*

$$f(\tilde{x}_k) - f^* \leq K_0 \left(1 - a\frac{\sqrt{\mu\mu_0}}{L}\right)^k (f(\tilde{x}_0) - f^*).$$

*with $K_0 := \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)$, $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$ and $L_0 := \inf\{L' \leq L : f \in QG_+^{L'}\}$.*

Broadly speaking, the convergence rates of Theorem 5.5.2 are the same as those of Theorem 5.5.1, divided by $L$, such that we can draw similar insights, see the proof in Appendix 5.D.2.

### 5.5.2 Acceleration using Momentum with Large Enough Aiming Condition Constant

We state the convergence rates of (NMO) and (NM), under the aiming condition.

**Theorem 5.5.3.** *Let $f \in PL^{\mu} \cap AC^{a} \cap LS^{L}$, $L_0 := \inf\{L' \leq L : f \in QG_{+}^{L'}\}$, $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_{-}^{\mu'}\}$.*

*(i) Let $(x_t)_{t\geq 0}$ follow* (NMO) *and constant parameters $\gamma \geq 0$, $\gamma' = (\mu_0 L_0)^{-1/4}$, $\eta = (\mu_0 L_0)^{1/4}$, $\eta' = a\,(\mu_0/L_0)^{1/4}\sqrt{\mu}$. Then,*

$$f(x_t) - f^* \leq K_0 e^{-a\left(\frac{\mu_0}{L_0}\right)^{\frac{1}{4}}\sqrt{\mu}t},$$

*where $K_0 := \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(x_0) - f^*)$*

*(ii) Let $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ verify* (NM) *with the continuized parameterization, with $\eta = (\mu_0 L_0)^{1/4}/\sqrt{L}$, $\gamma = 1/L$, $\gamma' = ((\mu_0 L_0)^{1/4}\sqrt{L})^{-1}$, and $\eta' = a\,(\mu_0/L_0)^{1/4}\sqrt{\mu/L}$. Then, with probability $1 - \frac{1}{c_0} - \exp(-(c_1 - 1 - \log(c_1))k)$, for some $c_0 > 1$ and $c_1 \in (0,1)$ we have*

$$f(\tilde{x}_k) - f^* \leq K_1 e^{-a\left(\frac{\mu_0}{L_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}(1-c_1)k},$$

*where $K_1 := c_0\left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(x_0) - f^*)$.*

See the proof in Appendix 5.D.3. The result of Theorem 5.5.3 (ii) holds under high probability, due to the use of the *continuized* version of Nesterov. In the following, we compare convergence rates despite this difference.

**Comparison of the Convergence Rates: Discrete Case** The convergence rates in Theorem 5.5.2 and Theorem 5.5.3 (ii) are

$$\frac{\mu}{L}, \qquad a\frac{\sqrt{\mu\mu_0}}{L}, \qquad a\left(\frac{\mu_0}{L_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}(1 - c_1).$$

Ignoring the factor $(1-c_1)$, the latter always dominates $a\frac{\sqrt{\mu\mu_0}}{L}$, such that we compare $\mu/L$ and $a(\mu_0/L_0)^{1/4}\sqrt{\mu/L}$. Thus, for functions in $\mathrm{PL}^{\mu} \cap \mathrm{AC}^{a} \cap \mathrm{LS}^{L}$, the rate of (NM) is better than the rate of (GD) whenever

$$a \;\geq\; \left(\frac{L_0}{\mu_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}.$$

This bound suggests that, in the presence of oscillations and flat plateaus, leading to $\mu/L \ll 1$, there is increased tolerance for poor alignment while still allowing acceleration. We test this bound numerically on the PL lower-bound function of (Yue et al., 2023), and find that the two quantities scale similarly, in a sense described in Appendix 5.7.2, providing a heuristic support for the relevance of our bound.

**Acceleration for Dynamical Systems: Flatness Matters** Comparing the rates of Theorem 5.5.1 (i) and Theorem 5.5.3 (i) yields an improvement of momentum if $a \geq (L_0/\mu_0)^{1/4}\sqrt{\mu}$, namely a similar bound as a in the discrete case, up to the $\sqrt{L}^{-1}$ factor. Contrasting with the discrete case, the rate of Theorem 5.5.3 (i) does not necessarily improve upon Theorem 5.5.1 (ii). Improvement occurs whenever

$$\mu_0 L_0 < 1.$$

Since $L_0 \geq \mu_0$, this implies that if $\mu_0 > 1$ (the function is sharp near the minimizer), then (GF) is preferable, whereas if $L_0 < 1$ (flatness near minimizer), then (NMO) performs better. In the intermediate regime $\mu_0 < 1 < L_0$, the outcome depends on the balance between the smallest and largest local curvatures: if $\mu_0$ is much smaller than 1 while $L_0$ is only moderately above 1, flatness dominates and (NMO) wins; conversely, in the opposite case. The advantage of (NMO) on flat landscapes is consistent with previous observations for the Heavy Ball dynamics when $f \in \mathrm{PL}^{\mu}$ (Apidopoulos et al., 2022; Gess and Kassing, 2023), whose dynamics are close to (NMO).

### 5.5.3 Decoupling information: more interpretability, but may create sub-optimality.

From Lemma 5.4.1, it follows that we do not study in this Section 5.5 a class of function that is fundamentally different from $\mathrm{SQC}^{\mu}_{\tau}$. We rather introduce a new characterization of these kind of functions in the idea of having better interpretability. This is done by decoupling the information: angular information with $\mathrm{AC}^{a}$, and curvature with $\mathrm{PL}^{\mu}$. In contrast, all these information are blended with $\mathrm{SQC}^{\mu}_{\tau}$, and according to Section 5.3, in a fairly hidden way.

This separation of information comes at a cost: this may create sub-optimality. A clear example is that for a quadratic function in $\mathrm{SC}^{\mu} \cap \mathrm{LS}^{L}$, one may have $a = 2\sqrt{\mu L}/(\mu + L)$. Plugged into our Table bounds, this creates significantly poorer bounds compared with the known bounds, see Section 5.1. This sub-optimality follows from the definition of the constants parameterizing the class, that follow from *worst-case points*. Define $a(x) = \langle \nabla f(x), x - x^* \rangle / (\|\nabla f(x)\| \, \|x - x^*\|)$, and assume the worst-case point $\hat{x} = \mathrm{arginf}_{x \in \mathbb{R}^d} a(x)$ is such that $a(\hat{x}) > 0$. Then, we have that $f \in \mathrm{AC}^{a}$ for $a = a(\hat{x})$. For the PL condition, the $\mu$ constant is also determined by such a worst-case point. However, the worst case points might not be the same for the two conditions; this is exactly what happens for a strongly convex quadratic function, where the worst aiming condition point is reached where the curvature is high. When an assumption blends the two information in a unique inequality, the associated worst case point balance between each condition. For such rigid functions such as quadratic, the different information (angular, curvature) are strongly coupled, such that a unique inequality can encompasses efficiently all the information. In a non-convex case these information may be decoupled. This may be seen with our toy example in Section 5.4, where we can fix $a = 1$ while playing with the curvature through $f$. This indicates that such a decoupling can bring interesting insight for non-convex functions, compared to the more traditional $\mathrm{SQC}^{\mu}_{\tau}$ based on a unique inequality.

Anyway, we note that the aforementioned limitation is rather artificial. A first reason is that it can be corrected if allowing a more involved definition of the classes. Precisely, the aiming condition will be used in our convergence proofs through Lemma 5.4.1. Its proof remains valid for $a$ defined as

$$a := \frac{\langle \nabla f(\hat{x}), \hat{x} - x^* \rangle}{\|\hat{x} - x^*\| \, \|\nabla f(\hat{x})\|},$$

where

$$\hat{x} := \mathrm{argmin}_{x \in \mathbb{R}^d} \frac{\langle \nabla f(x), x - x^* \rangle}{\|x - x^*\| \, \|\nabla f(x)\|} \cdot \frac{\|\nabla f(x)\|}{\sqrt{f(x) - f^*}}.$$

Namely, the aiming condition constant may be considered at a point that creates a balance with the PL condition. This allows to keep the information of the aiming condition, while not inducing an artificial sub-optimality. A second justification is provided in the next section.

## 5.6 Relaxed Setting: Aiming Condition on Average

There is a gap between the convergence observed in practice and the theoretical bounds established for globally defined classes such as PL functions. The reason is intuitive: if the global landscape of non-convex functions can be highly irregular, nevertheless, the optimization path may remain in a favorable region. Importantly, the performance of first-order algorithms such as (GD) and (NM) only depends on the geometry surrounding their optimization path.

**Naive Relaxation** A first direct idea to take this intuition into account is to consider an algorithm $\{x_n\}_{n \geq 0}$ and a minimizer $\hat{x}$ such that $\tilde{x}_k \to_{k \to +\infty} \hat{x}$, and assuming

$$\forall k \in \mathbb{N}, \langle \nabla f(\tilde{x}_k), \tilde{x}_k - \hat{x} \rangle \geq a \, \|\nabla f(\tilde{x}_k)\| \, \|\tilde{x}_k - \hat{x}\| \, . \tag{5.8}$$

With words, $f$ is assumed to satisfy the $a$-aiming condition along the trajectory $\{x_n\}_{n \geq 0}$, with respect to its convergence point $\hat{x}$. For a function in $\mathrm{PL}^{\mu} \cap \mathrm{LS}^{L}$, if $\{x_n\}_{n \geq 0} \sim$ (GD) or (NM) verify (5.8), we would deduce for this sequence the same result as for Section 5.5. The relaxation is twofold: (i) we get rid of the uniqueness minimizer assumption, and (ii) the convergence rate is not affected by potential pathological regions of $f$, as long as the sequence does not cross it. However, it suffices for (5.8) to not be verified for a single iteration to make this reasoning fail. We thus derive a relaxation that, in the case where pathological regions remain a negligible section of the optimization path, still allows to keep accelerated results.

**Averaging Relaxation** We relax the viewpoint (5.8). We address the continuous case; the discrete case follows similarly, see Appendix 5.E.

**Assumption 5.6.1.** *For some minimizer $x^* \in \mathbb{R}^d$, $\theta > 0$, $a_{avg} > 0$, and $(x_t)_{t\geq 0}$, we have*

$$\int_0^t e^{\theta s} \Delta_s ds \geq 0,$$

$\Delta_s := \langle \nabla f(x_s), x_s - x^* \rangle - a_{avg} \|\nabla f(x_s)\| \|x_s - x^*\|$.

A function $f$ satisfies Assumption 5.6.1 as long as the aiming condition holds along a specific average of the trajectory. If $f \in \mathrm{AC}^a$, Assumption 5.6.1 holds with $a_{avg} = a$ for any initialization. However, we could have $a \ll a_{avg}$, or having that the path crosses some regions where the aiming condition is not verified, while still having $a_{avg} > 0$.

**Theorem 5.6.1.** *Let $f \in PL^\mu \cap QG_+^{L_0}$, $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$ and $a_{avg} > 0$. If:*

*(i) $(x_t)_{t\geq 0}$ follows* (GF) *for a given initialization $x_0 \in \mathbb{R}^d$ is such that Assumption 5.6.1 holds with $\theta = a_{avg}\sqrt{\mu\mu_0}$. , then*

$$f(x_t) - f^* \leq K_0 e^{-a_{avg}\sqrt{\mu\mu_0}t}(f(x_0) - f^*).$$

*(ii) $(x_t)_{t\geq 0}$ follows* (NMO) *for $\eta' = a_{avg}\left(\mu_0/L_0\right)^{1/4}\sqrt{\mu}, \gamma' = (\mu_0 L_0)^{-1/4}$, $\eta = (\mu_0 L_0)^{1/4}$, $\gamma \geq 0$ and such that Assumption 5.6.1 holds with $\theta = a_{avg}\left(\mu_0/L_0\right)^{1/4}\sqrt{\mu}$, then*

$$f(x_t) - f^* \leq K_0 e^{-a_{avg}\left(\frac{\mu_0}{L_0}\right)^{\frac{1}{4}}\sqrt{\mu}}(f(x_0) - f^*).$$

*In both statements, $K_0 := \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)$.*

The proof is in Appendix 5.E.1. We obtain the same convergence rate as for Theorem 5.5.1 (ii) and Theorem 5.5.3 (i), replacing $a$ with $a_{avg}$. However, the rates of Theorem 5.6.1 could be either significantly sharper, or allow for obtaining such a rate while the algorithm crosses regions with negative aiming condition values. This reasoning can be extended to other geometrical conditions, such as $\mathrm{PL}^\mu$. This perspective might explain why, in practice, grid-search selection of parameters often leads to relatively large constants, compared to what would be allowed by geometrical constants that parameterize globally the associated landscape.

## 5.7 Numerical Experiments

We provide numerical experiments to support our theoretical findings. We first design a simple 2d example that belongs to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$, has a unique minimizer, but does not belong to $\mathrm{AC}^a$. This will result in momentum hurting the convergence speed in the early stage of optimization, see Section 5.7.1. Then, in Section 5.7.2, we provide a heuristic validation of the bound to have acceleration with (NM) provided in Section 5.5.2.

### 5.7.1 An Example with Negative Aiming Condition

In this section, we design an example that shows how negative aiming condition values can hurt the convergence of momentum algorithms, and then discuss the aiming condition in neural networks. For our experiment, we consider an alternative formulation of (NM) that we call (NM').

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \alpha(\tilde{x}_k - \tilde{x}_{k-1}) \\ \tilde{x}_{k+1} = \tilde{y}_k - \gamma \nabla f(\tilde{y}_k) \end{cases} \tag{NM'}$$

While (NM) with the continuized parameterization is convenient for theoretical analysis, (NM') is more amenable to practical implementation, as it involves only two easily tunable parameters with clear interpretations: the stepsize $s$ and the momentum parameter $\alpha$. See (Defazio, 2019) for a discussion on the different forms of Nesterov momentum.

Consider the 2-dimensional function

$$F_\varepsilon(x, y) = 0.5(y - \sin(x))^2 + 0.5\varepsilon x^2.$$

The special case $\varepsilon = 0$ is introduced in (Apidopoulos et al., 2022). For $\varepsilon > 0$, $F_\varepsilon$ has $(0, 0)$ as unique minimizer, and it belongs to $\text{PL}^\mu$ with $\mu \leq \frac{2+\varepsilon-\sqrt{\varepsilon^2+4}}{2}$, see Lemma 5.B.1. We display $F_{0.001}$ in Figure 5.3. We also provide a three-dimensional visualization in the appendix, see Figure 5.9. At certain points of the landscape, the descent direction is anti-correlated with the direction pointing toward the minimizer, thereby violating the aiming condition.

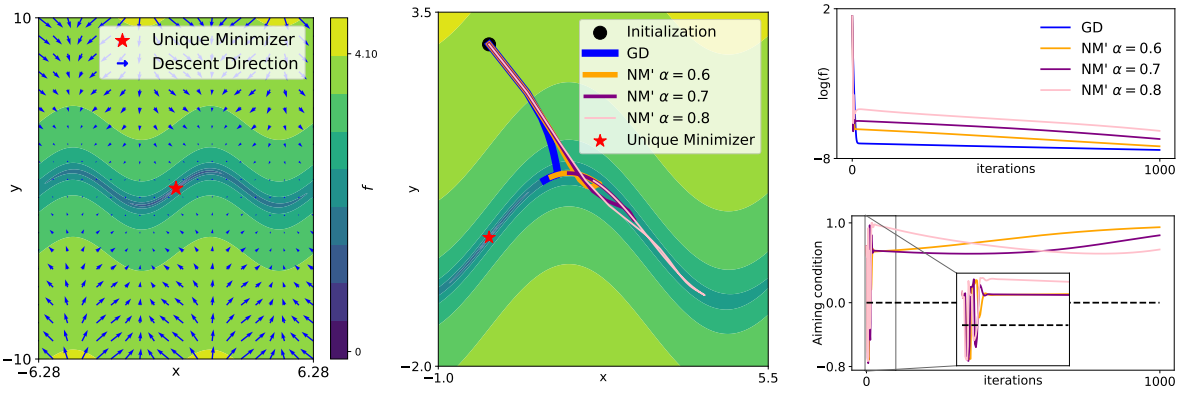



Figure 5.3: Left: Heatmap of $F_{0.001}(x, y) = 0.5(y - \sin(x))^2 + 0.001 \cdot 0.5x^2$. Blue arrows indicate the descent direction $-\nabla F_{0.001}$. Center: First 1000 iterations of the trajectories of (GD) and (NM'), both starting from the initialization point $(0, 3)$, for different values of momentum parameter $\alpha$. Top right: Corresponding decrease of $\log(f)$. Bottom right: Values of the aiming condition along the iterates, zoomed on the first 100 iterations. Early negative aiming condition values cause momentum to drive the trajectory away from the minimizer, leading (GD) to outperform momentum in early iterations, with increasing effect for larger $\alpha$.

**Negative aiming condition shoots** (NM') **away** In Figure 5.3, we compare (GD) and (NM') for different values of momentum parameter $\alpha$, see details in Appendix 5.B.2. Due to negative aiming condition values during the early phase of the optimization process, iterates of (NM') are initially driven away from the minimizer. Increasing the momentum parameter amplifies this effect, causing the trajectory to deviate even further. As a result, (GD) outperforms (NM') for up to 1000 iterations in terms of function value decrease. This example illustrates how negative aiming condition values can hinder momentum methods. More broadly, designing functions with unfavorable aiming behavior while retaining desirable optimization properties—such as satisfying the PL condition and having a unique minimizer—appears to be an interesting direction for better understanding landscapes on which momentum is ineffective.

(NM') **eventually catches up due to favorable averaged aiming condition** Figure 5.3 shows that, although the aiming condition takes negative values at early iterations, it then stabilizes around a positive value close to one. As discussed in Section 5.6, this suggests that (NM') should eventually benefit from momentum for a large enough number of iterations. We confirm this behavior in Figure 5.10 in the appendix.

**Aiming Condition and Neural Networks** Because of the curse of dimensionality and their high level of non-convexity, we could expect that large-scale models such as those involved in deep learning do not verify any aiming condition. This intuition is challenged by the fact that well-designed neural networks are known to have favorable optimization properties. It was observed in (Li et al., 2018) that around 90% of the variations of the path of SGD with large batch size lie in a dimension 2 space, despite the high dimensionality of the models. Of special interest to our work, (Guille-Escuret et al., 2024) performs an extensive empirical evaluation of the aiming condition, which they name the *cosine similarity*. They observe that for various tasks and architectures, $a$ remains between $10^{-1}$ and $10^{-3}$, despite the high dimensionality of the model. This observation was consistent across several stochastic first-order algorithms, including SGD with momentum. Although our theoretical results do not apply to general deep learning models, these

experiments tend to highlight the crucial role of the aiming condition as a key ingredient in explaining the effectiveness of momentum.

### 5.7.2 Heuristic Evaluation of the Acceleration Bound

In this section, we provide a heuristic validation of the bound on the aiming condition to provide acceleration, derived in Section 5.5.2. To do so, we consider the $\mathrm{PL}^\mu$ function introduced in (Yue et al., 2023), which serves to obtain a lower bound on the number of gradient call needed by first-order algorithms–algorithms that make use of gradient information–to obtain a point $\hat{x} \in \mathbb{R}^d$ such that $f(\hat{x}) - \min f \leq \varepsilon$. Precisely, it is shown that the number of needed gradient calls scales at best as $\mathcal{O}(\frac{L}{\mu}\log(\frac{1}{\varepsilon}))$ for any first-order algorithms. This bound corresponds to a $\mu/L$ convergence rate, inducing that, up to constants, the convergence rate of (GD) is optimal, see Table 5.1. The definition of the hard instance is rather involved see details in Appendix 5.B.3.

**Non acceleration of momentum on the hard function** We observe empirically that, as expected, momentum does not allow improvement over gradient descent when minimizing the hard function of (Yue et al., 2023). We consider (GD) and (NM) with the continuized parameterization. We chose a parameterization of the hard function such that it belongs to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$ with $\mu = 10^{-4}$, $L = 10^3$, and the dimension is 4550 (see details in Appendix 5.B.3). For both (GD) and (NM), we tune the parameters such that the decrease at iteration 1000 is maximized. We plot the decrease of function values of the hard function in Figure 5.4. We observe that tuning of (NM) only permits to catch up with the decrease of (GD), but provides no acceleration.

**Heuristic test of the acceleration bound** In Section 5.5.2, we stated that for $f \in \mathrm{PL}^\mu \cap \mathrm{AC}^a \cap \mathrm{LS}^L$, momentum provide acceleration as long as

$$a \geq \left(\frac{L_0}{\mu_0}\right)^{1/4} \sqrt{\frac{\mu}{L}}.$$

It is natural to test whether this bound is verified or not on the aforementioned hard function. Importantly, the following discussion should be regarded as a **heuristic**, rather than a rigorous derivation.

For our test, we do not consider the global parameters $\mu, a, L$ such that $f$ belongs to $\mathrm{PL}^\mu \cap \mathrm{AC}^a \cap \mathrm{LS}^L$. This is because (i) we do not know the optimal parameters $a, \mu_0$ and $L_0$, and (ii) in the lower bound provided in (Yue et al., 2023), there is a hidden constant with values $\approx 10^6$, which is not negligible. Moreover, the decrease observed in Figure 5.4 is significantly faster than what is predicted by the theoretical bound derived under the global parameters, such that it actually poorly describes the algorithm's behavior. This motivates to use a perspective that has a similar spirit to the one described in Section 5.6, such that we will consider the values of PL and aiming condition taken along the optimization path. Along the iterates of $\{x_n\}_{n\geq 0}$ generated by (GD) or (NM), we compute the point-wise PL and aiming condition values, namely

$$\mu(\tilde{x}_i) := 2\frac{\|\nabla f(\tilde{x}_i)\|^2}{f(\tilde{x}_i) - f^*}, \quad a(\tilde{x}_i) := \frac{\langle \nabla f(\tilde{x}_i), \tilde{x}_i - x^* \rangle}{\|\nabla f(\tilde{x}_i)\| \, \|\tilde{x}_i - x^*\|}.$$

Empirically, we observe that $\mu(\tilde{x}_i) \approx 1$ (see Figure 5.5), while the function is $\mathrm{PL}^\mu$ with $\mu = 10^{-4}$. This suggests that the algorithm stays in favorable regions compared to those existing on the global landscape. As a result, in the case of (GD), compared to the rate $\mu/L$, we expect that its convergence rate is more accurately described by

$$\tilde{\mu}_{gd}/L,$$

with $\tilde{\mu}_{gd} = \frac{1}{N}\sum_{i=1}^N \mu(\tilde{x}_i)$ is the averaged point-wise PL values along the $N$ iterations of the algorithm. Moreover, because we observed that (GD) achieves a better decrease with stepsize $\gamma_{gd} \geq 1/L$, running the algorithm with this stepsize suggests that an even more accurate rate is

$$\tilde{\mu}_{gd}\gamma_{gd}. \tag{11}$$

We plot the curve associated to this theoretical rate in Figure 5.4. We observe that, if not matching perfectly the decrease of (GD), it is significantly more accurate compared with the rate $\mu/L$. Following the same reasoning, we will model the rate for (NM) as (see Theorem 5.5.3 (ii))

$$\tilde{a}(\tilde{\mu}_0/\tilde{L}_0)^{1/4}\sqrt{\tilde{\mu}_{nm}\gamma_{nm}}, \tag{12}$$

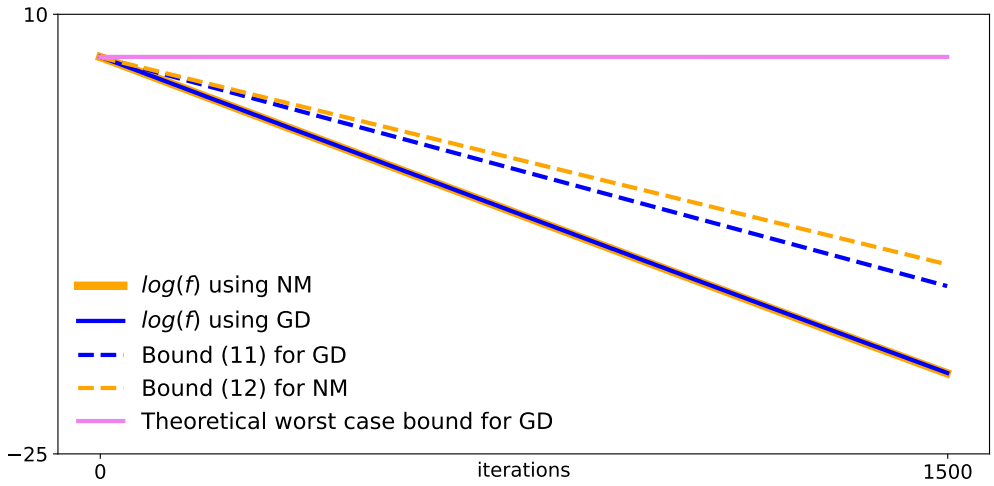



Figure 5.4: Function values decrease when running (GD) and (NM) on the hard instance of (Yue et al., 2023). We also plot the log of the theoretical convergence upper bound $\log((1-\mu\gamma)^k(f(x_0)-f^*)$ with $\mu$ the global PL constant, and the log of the theoretical convergence upper bound using parameter values across the path, see (11) and (12) in Section 5.7.2. It can be observed that (GD) and (NM) decrease at same speed, and that the bounds proposed at (11) and (12) are significantly more accurate compared with the global worst case bound.

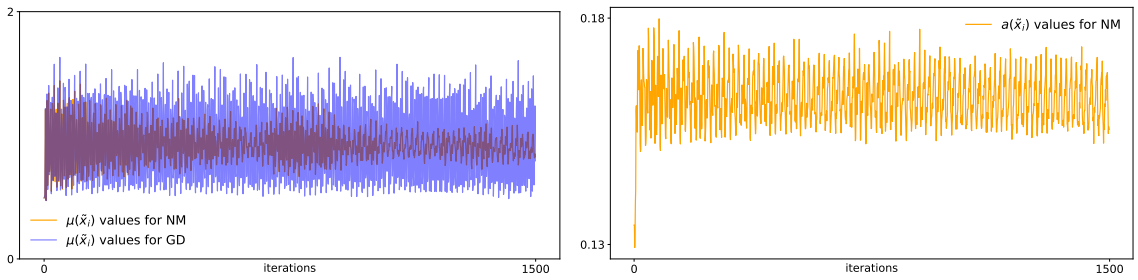



Figure 5.5: Left: Values of $\mu(\tilde{x}_i)$ along the iterates of (GD) and (NM). Values of $a(\tilde{x}_i)$ along the iterates of (NM). The underlying function is the hard instance of (Yue et al., 2023), see details in Appendix 5.B.3.

where $\tilde{a} = \frac{1}{N}\sum_{i=1}^N a(\tilde{x}_i)$ and $\tilde{\mu}_{nm} = \frac{1}{N}\sum_{i=1}^N \mu(\tilde{x}_i)$ are respectively the point-wise aiming condition value and point-wise PL values along the algorithm, and $\tilde{\mu}_0$ and $\tilde{L}_0$ are estimation of the optimal constant involved in $\mathrm{QG}_-^{\mu_0}$ and $\mathrm{QG}_+^{L_0}$, and $\gamma_{nm} \geq L$ is the used stepsize for (NM) that arises from the tuning of parameters. The parameters used in (NM) result from grid-search rather than the theoretical recommendation of Theorem 5.5.3, and we observe that (NM) has a slightly better decreasing rate compared to the bound (12). Up to associating the empirical decreasing rate of (GD) and (NM) with the bounds (11) and (12), we thus test the acceleration condition of Section 5.5.2 by comparing $\tilde{a}$ with

$$\frac{\tilde{\mu}_{gd}\gamma_{gd}}{\sqrt{\gamma_{nm}\tilde{\mu}_{nm}}}\left(\frac{\tilde{L}_0}{\tilde{\mu}_0}\right)^{\frac{1}{4}}.$$

We find numerically that $\tilde{a} \approx 0.16$, while $\frac{\tilde{\mu}_{gd}\gamma_{gd}}{\sqrt{\gamma_{nm}\tilde{\mu}_{nm}}}\left(\frac{\tilde{L}_0}{\tilde{\mu}_0}\right)^{\frac{1}{4}} \approx 0.19$ such that they are roughly the same. So, in this setting where (NM) does not accelerate over (GD), we observe that the acceleration condition of Section 5.5.2 is indeed not verified. This gives heuristic weights to the relevance of this bound. Finally, this suggests that the function that serves as lower bound for PL functions, which actually belongs to $\mathrm{PL}^\mu \cap \mathrm{AC}^a \cap \mathrm{LS}^L$, is such that the aiming condition values do not scale properly with other geometrical constants to permit acceleration using momentum.

## Reproducibility Statement

Our figures and experiments are reproducible, see our code `https://github.com/J-Hermant/Nesterov_Acceleration_Aiming_Condition/tree/main`.

## 5.A Related Works and Background Details

In this section, we first provide bibliographical details about Table 5.1 (Appendix 5.A.1).

### 5.A.1 Details for Table 5.1

In this section, we detail Table 5.1, giving in particular references for the results. We note that some of the results do not exist in the literature. Before, we state the following result that we will use in this section.

**Proposition 5.A.1.** *Let $(x_t, z_t)_{t\geq 0}$ verify* (NMO). *Then, it verifies the Heavy Ball equation with Hessian damping*

$$\ddot{x}_t + \left(\eta_t - \frac{\dot{\eta}_t}{\eta_t} + \eta_t'\right)\dot{x}_t + \gamma\nabla^2 f(x_t)\dot{x}_t + \left(\eta_t'\gamma + \eta_t\gamma' - \frac{\dot{\eta}_t}{\eta_t}\gamma\right)\nabla f(x_t) = 0. \tag{5.9}$$

Proposition 5.A.1 states that (NMO) can be written as the Heavy Ball equation with Hessian damping (Alvarez et al., 2002). It was proved in the appendices of Chapter 4. We will use results established for this equation in the case of (NMO).

**About $\mathbf{SC}^\mu$** The rate for (GF) can be deduced from the one in the $\mathrm{PL}^\mu$ case. The rate stated for (NMO) can be found in the case of heavy ball with Hessian damping in (Shi et al., 2022), from which we deduce the same rate for (NMO). The rates for (GD) and (NM) can be found in (Nesterov, 2013).

**About $\mathbf{SQC}^\mu_\tau$** See Proposition 5.A.2 for the rate of (GF). The rate of (GD) can be found in Chapter 3. The rate stated for (NMO) can be found in the case of heavy ball with Hessian damping, see also Chapter 3, from which we deduce the same rate for (NMO). The rate for (NM) requires some care. It can be found in (Hinder et al., 2020, Theorem 1), but involves a line-search procedure that requires computing evaluations of $f$ and $\nabla f$ at each iteration to compute one of the parameters. It can also be found in Chapter 3 up to a supplementary assumption of weak-convexity, namely $f + \rho/2\left\|\cdot\right\|^2$ is convex for some $\rho \in \mathbb{R}$, with $\rho \leq L$. Finally, it can be found in (Wang and Wibisono, 2023, Theorem 2) up to non-deterministic results, because of the use of the continuized parameterization.

**About $\mathbf{PL}^\mu$** The rate $\mu$ for (GF) can be found in (Polyak and Shcherbakov, 2017, Theorem 3). The rate $\mu/L$ for (GD) has been known for a long time (Polyak, 1963; Bolte et al., 2017), see (Karimi et al., 2016, Theorem 1) for a direct proof. The rate for (NM) can be found in Chapter 4, but holds with high probability because of the use of the continuized framework. We note that using the Heavy Ball algorithm, this rate can be improved to a rate close to $\sqrt{\mu/L}$ if assuming that $f$ is four times differentiable, and if the algorithm is initialized close enough to the minimizer (Kassing and Weissmann, 2024). The case of (NMO) is more complicated. For Polyak's momentum (Heavy Ball equation), a result exists assuming $f$ is also in $\mathrm{LS}^L$, yielding a convergence rate $\approx \mu/\sqrt{L}$ (Apidopoulos et al., 2022). For the same equation, the rate $\sqrt{\mu}$ holds asymptotically, provided that $f$ is four times differentiable, providing strong local regularity around minimizers (Kassing and Weissmann, 2024). In the case of Heavy Ball with Hessian damping, a rate $\sqrt{\mu}$ is deduced (Apidopoulos et al., 2025), but the gradient term in the equation is rescaled by $\mathcal{O}(1/\sqrt{\mu})$, which means we are minimizing a rescaled version of the function. In Proposition 5.A.3, we actually show without assuming $f \in \mathrm{LS}^L$ or smoother than $C^1$–and with a considerably simpler proof than the previously mentioned works–that (NMO) achieves a rate $\gamma\mu$, where $\gamma$ is a parameter of (NMO). It highlights that we can achieve whatever rate we want, as long as we take $\gamma$ big enough. However, in the discrete case, $\gamma$ corresponding intuitively to the stepsize, it appears clearly that the descent lemma induced by $f \in \mathrm{LS}^L$ will impose a bound on $\gamma$, namely $\gamma \leq 1/L$, yielding a rate $\mu/L$.

**Proof of not existing rates of Table 5.1**

**Proposition 5.A.2.** *Let $f \in SQC^\mu_\tau$, $(x_t)_{t\geq 0} \sim$* (GF). *Then,*

$$f(x_t) - f^* = \mathcal{O}(e^{-\tau\mu t}).$$

*Proof.* Let

$$E_t = f(x_t) - f^* + \frac{\mu}{2}\left\|x_t - x^*\right\|^2.$$

Then

$$\begin{aligned}\dot{E}_t &= \langle \nabla f(x_t), \dot{x}_t \rangle + \mu \langle x_t - x^*, \dot{x}_t \rangle \\ &\overset{(i)}{=} -\|\nabla f(x_t)\|^2 - \mu \langle \nabla f(x_t), x_t - x^* \rangle \\ &\leq -\mu \langle \nabla f(x_t), x_t - x^* \rangle \\ &\overset{(ii)}{\leq} -\mu\tau\Big(f(x_t) - f^* + \frac{\mu}{2}\|x_t - x^*\|^2\Big) = -\mu\tau E_t,\end{aligned}$$

where (i) follows from (GF) and (ii) from $f \in \mathrm{SQC}^\mu_\tau$. Integrating yields $E_t \leq e^{-\mu\tau t}E_0$, hence $f(x_t) - f^* \leq E_t = \mathcal{O}(e^{-\mu\tau t})$. □

**Proposition 5.A.3.** *Let $f \in \mathrm{PL}^\mu$. Let $(x_t)_{t\geq 0}$ follow*

$$\text{(NMO)} \textit{with}$$

*$\eta_t \equiv \eta$ and $\eta'_t \equiv \eta'$ constants and under the constraint $\gamma = (\eta + \eta')/\mu$. Then, we have*

$$f(x_t) - f^* = \mathcal{O}(e^{-2\mu\gamma t}).$$

*Proof.* Let

$$E_t = f(x_t) - f^* + \frac{\delta}{2}\|x_t - z_t\|^2.$$

Then

$$\begin{aligned}\dot{E}_t &= \langle \nabla f(x_t), \dot{x}_t \rangle + \delta \langle x_t - z_t, \dot{x}_t - \dot{z}_t \rangle \\ &\overset{(i)}{=} \langle \nabla f(x_t), \eta(z_t - x_t) - \gamma \nabla f(x_t)) \rangle + \delta \langle x_t - z_t, (\eta + \eta')(z_t - x_t) + (\gamma' - \gamma)\nabla f(x_t) \rangle \\ &= \big(\eta + \delta(\gamma - \gamma')\big) \langle \nabla f(x_t), z_t - x_t \rangle - \gamma \|\nabla f(x_t)\|^2 - \delta(\eta + \eta')\|z_t - x_t\|^2,\end{aligned}$$

where (i) follows from (NMO). Using $f \in \mathrm{PL}^\mu$, we have $-\gamma\|\nabla f(x_t)\|^2 \leq -2\mu\gamma(f(x_t) - f^*)$, hence

$$\dot{E}_t \leq \big(\eta + \delta(\gamma - \gamma')\big) \langle \nabla f(x_t), z_t - x_t \rangle - 2\mu\gamma(f(x_t) - f^*) - \delta(\eta + \eta')\|z_t - x_t\|^2.$$

Impose $\eta + \delta(\gamma - \gamma') = 0$ (*i.e.* $\delta = \eta/(\gamma' - \gamma)$) and $\eta + \eta' = \mu\gamma$. Then $\dot{E}_t \leq -2\mu\gamma E_t$, so $E_t \leq e^{-2\mu\gamma t}E_0$. Since $f(x_t) - f^* \leq E_t$, the claim follows. □

### 5.A.2 Choice of Parameters for $\mathrm{SQC}^\mu_\tau$.

In Section 5.3, we discuss that $\sup_\tau \sup_\mu \{(\tau, \mu) \in (0, 1] \times \mathbb{R}_+ : f \in \mathrm{SQC}^\mu_\tau\}$ and $\sup_\mu \sup_\tau \{(\tau, \mu) \in (0, 1] \times \mathbb{R}_+ : f \in \mathrm{SQC}^\mu_\tau\}$ leads to different choices of parameters. To justify this statement, we assume there exists $\tau, \mu \in (0, 1] \times \mathbb{R}_+$ such that $f \in \mathrm{SQC}^\mu_\tau$. We first recall $f \in \mathrm{SQC}^\mu_\tau \Rightarrow f \in \mathrm{SQC}^{\mu/\theta}_{\theta\tau}$, for all $\theta \in (0, 1]$ (Hinder et al., 2020, Observation 5). Letting $\theta \to 0$, it follows that

$$\arg\sup_\tau \arg\sup_\mu \{(\tau, \mu) \in (0, 1] \times \mathbb{R}_+ : f \in \mathrm{SQC}^\mu_\tau\} = (0, +\infty).$$

On the other hand, it is clear that $\arg\sup_\mu \arg\sup_\tau \{(\tau, \mu) \in (0, 1] \times \mathbb{R}_+ : f \in \mathrm{SQC}^\mu_\tau\} \neq (0, +\infty)$.

## 5.B More Details on Figures and Numerical Experiments

In this section, we provide details about our method to generate Figure 5.1, Figure 5.2 (Appendix 5.B.1) and Figure 5.3 (Appendix 5.B.2). We provide in Appendix 5.7.2 a heuristic discussion of the bound over aiming condition to have acceleration derived in Section 5.5.2. We also provide in Appendix 5.B.4 additional figures, useful for the main text but that could not be included because of space limitations. Our figures and experiments are reproducible; see the code in the supplement.

### 5.B.1 Details about Figure 5.1 and Figure 5.2

In this section, we detail how we numerically compute the convergence rate of the functions displayed in Figure 5.1 and Figure 5.2. We split our presentation between the two underlying functions.

**One-dimensional example** The function displayed on the left in Figure 5.1 is

$$f(t) = 5(t + 0.19\sin(5t))^2,$$

defined on $[-2, 2]$. It has 0 as its unique critical point and minimizer, which satisfies $f(0) = 0$. We define a grid of $10^5$ points in $[-2, 2]$, equally spaced. This grid has been refined until the subsequent numerical values are not impacted by further refinement. For the $i$-th point $x_i$ in this grid, we compute the associated point-wise PL value

$$\mu_i = \frac{f'(x_i)^2}{2f(x_i)},$$

and point-wise $L$-smooth value

$$L_i = |f''(x_i)|.$$

We define $\mu := \min_{1\leq i\leq 10^5} \mu_i$ and $L := \max_{1\leq i\leq 10^5} L_i$ the constant such that numerically, $f \in \mathrm{PL}^\mu \cap \mathrm{LS}^L$ on $[-2, 2]$. On this domain, we thus obtain a convergence rate $\mu/L$ for (GD).

Because of the double parameterization, the case of $\mathrm{SQC}^\mu_\tau$ is slightly more involved. We define a grid of 1000 potential $\tau$ parameters, equally spaced in $[10^{-5}, 0.1]$. Note that we take $10^{-5}$ as the lower bound for this interval instead of zero, to avoid dividing by zero. For a given $\tau_j$ in the grid on $[10^{-5}, 0.1]$ , and each $x_i$ belonging to the grid on $[-2, 2]$, we compute the point-wise strong quasar convexity constant $\mu'_{i,j}$

$$\mu'_{i,j} = -2\frac{\Big(f(x_i) - f'(x_i)x_i\Big)}{x_i^2}.$$

For this $\tau_j$, we define $\mu'_j = \min_{1\leq i\leq 10^5} \mu_{i,j}$ the constant such that numerically, $f \in \mathrm{SQC}^{\mu_j}_{\tau_j}$. In the case where $\mu_j \leq 0$, we consider $\tau_j$ is not admissible, which occurs for $\tau_j \geq 0.1$. The pairs $(\tau_j, \mu_j)_{1\leq j\leq 1000}$ thus generate Figure 5.2, by plotting $(\tau_j\mu_j/L)_{1\leq j\leq 1000}$ for (GD) and $(\tau_j\sqrt{\mu_j/L})_{1\leq j\leq 1000}$ for (NM). As a supplement, we plot in Figure $(\tau_j\mu_j)_{1\leq j\leq 1000}$ and $(\tau_j\sqrt{\mu_j})_{1\leq j\leq 1000}$ which are the convergence rates of (GF) and (NMO). We observe a similar behavior compared with the discrete case. Finally, we obtain the convergence rate values of Figure 5.1 by considering $\max_{1\leq j\leq 1000} \tau_j\mu_j/L$ for (GD), and $\max_{1\leq j\leq 1000} \tau_j\sqrt{\mu_j/L}$ for (NM). As illustrated in Figure 5.2, the pair $(\tau_{j_0}, \mu_{j_0})$ that maximizes the rate of (GD) differs from the pair $(\tau_{j_1}, \mu_{j_1})$ that maximizes the rate of (NM).

**Two-dimensional example** The function displayed on the left in Figure 5.1 is

$$f(x, y) = 0.5(0.5x^2 - y)^2 + 0.05x^2,$$

defined on the square $[-1.2638, 1.2638]^2$. It has a unique minimizer and critical point at $(0, 0)$. To compute the convergence rate for this function, we proceed similarly to the one-dimensional case. The only difference is that we define a two-dimensional grid on $[-1.2638, 1.2638]^2$ of $10^6$ points, equally spaced. The grid has been refined until the numerical values we compute are not impacted by further refinement.

### 5.B.2 Details about Figure 5.3

In Figure 5.3, we optimize

$$F_\varepsilon(x, y) = 0.5(y - \sin(x))^2 + 0.5\varepsilon x^2.,$$

with $\varepsilon = 10^{-3}$. It is clear that $F_\varepsilon(0, 0) = 0$ is a global minimizer, and for all $(x, y) \neq (0, 0)$, $F_\varepsilon(x, y) > 0$. This function has no other critical point, which is induced by the fact that it is PL.

**Lemma 5.B.1.** *For $x, y \in \mathbb{R}$, $\varepsilon > 0$ the function $F_\varepsilon(x, y) = \frac{1}{2}(y - \sin x)^2 + \frac{1}{2}\varepsilon x^2$ belongs to $PL^\mu$ with*

$$\mu \leq \frac{2 + \varepsilon - \sqrt{\varepsilon^2 + 4}}{2}.$$

*Proof.* Consider the function

$$F_\varepsilon(x,y) = \tfrac{1}{2}\,(y-\sin x)^2 + \tfrac{1}{2}\,\varepsilon x^2, \qquad x,y\in\mathbb{R},\ \varepsilon>0.$$

Let

$$u(x,y) := y - \sin x.$$

Then

$$F_\varepsilon = \tfrac{1}{2}\,(u^2+\varepsilon x^2).$$

The gradient is

$$\nabla F_\varepsilon = \begin{pmatrix} -u\cos x + \varepsilon x \\ u \end{pmatrix}, \qquad \|\nabla F_\varepsilon\|^2 = u^2 + (-u\cos x+\varepsilon x)^2.$$

Writing both sides as quadratic forms in $z = (u(x,y),x)^\top$, we obtain

$$2F_\varepsilon = z^\top \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} z, \qquad \|\nabla F_\varepsilon\|^2 = z^\top \begin{pmatrix} 1+\cos^2 x & -\varepsilon\cos x \\ -\varepsilon\cos x & \varepsilon^2 \end{pmatrix} z.$$

Define

$$N = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \qquad M(x) = \begin{pmatrix} 1+\cos(x)^2 & -\varepsilon\cos(x) \\ -\varepsilon\cos(x) & \varepsilon^2 \end{pmatrix}, \qquad c := \cos x \in [-1,1].$$

So, $f$ belong to $\mathrm{PL}^\mu$ with

$$\begin{aligned} \mu = \min_{\substack{z=(u(x,y),x)\neq 0,\\ (x,y)\in\mathbb{R}^2}} \frac{z^T M(x) z}{z^T N z} &\geq \min_{\substack{z=(z_1,z_2),\\ (z_1,z_2)\in\mathbb{R}^2\setminus(0,0)}} \frac{z^T M(x) z}{z^T N z} \\ &\overset{(i)}{=} \min_{\substack{z=(z_1,z_2),\\ (z_1,z_2)\in\mathbb{R}^2\setminus(0,0)}} \frac{z^T N^{-1/2} M(x) N^{-1/2} z}{z^T z} \\ &\overset{(ii)}{=} \lambda_{\min}(N^{-1/2}M(x)N^{-1/2}), \end{aligned} \tag{5.10}$$

where (i) holds with the reparameterization $z' = N^{-1/2}z$, and (ii) is a property of the Rayleigh coefficient. We consider the matrix

$$N^{-1/2}M(x)N^{-1/2} = \begin{pmatrix} 1+\cos(x)^2 & -\sqrt{\varepsilon}\,\cos(x) \\ -\sqrt{\varepsilon}\,\cos(x) & \varepsilon \end{pmatrix} := \begin{pmatrix} 1+c^2 & -\sqrt{\varepsilon}\,c \\ -\sqrt{\varepsilon}\,c & \varepsilon \end{pmatrix} =: B(c), \quad c := \cos(x) \in [-1,1].$$

So, we have

$$\lambda_{\min}(N^{-1/2}M(x)N^{-1/2}) \geq \inf_{c\in[-1,1]}\ \lambda_{\min}(B(c)).$$

The eigenvalues of $B(c)$ are

$$\lambda_\pm(c) = \frac{\mathrm{tr}(B(c)) \pm \sqrt{\mathrm{tr}(B(c))^2 - 4\det(B(c))}}{2}.$$

We have

$$\det(B(c)) = \varepsilon, \qquad \mathrm{tr}(B(c)) = 1+c^2+\varepsilon.$$

Such that we solve

$$\min_{c\in[-1,1]} \frac{1+c^2+\varepsilon - \sqrt{(1+c^2+\varepsilon)^2-4\varepsilon}}{2} = \min_{t\in[0,1]} \frac{1+t+\varepsilon - \sqrt{(1+t+\varepsilon)^2-4\varepsilon}}{2} := \phi(t).$$

We have for $t\in[0,1]$,

$$2\phi'(t) = 1 - \frac{1+t+\varepsilon}{\sqrt{(1+t+\varepsilon)^2-4\varepsilon}} \leq 0,$$

such that $\phi$ is decreasing on $[0,1]$. So, the minimum is reached for $t = 1 \Rightarrow c^2 = 1$, which leads to the eigenvalue

$$\inf_{c\in[-1,1]} \lambda_{\min}(B(c)) = \lambda_{\min}(B(1)) = \frac{2+\varepsilon-\sqrt{\varepsilon^2+4}}{2} > 0.$$

Therefore, for all $(x,y) \in \mathbb{R}^2$,

$$\|\nabla F_\varepsilon(x,y)\|^2 \;\geq\; 2\mu\, F_\varepsilon(x,y),$$

with

$$\mu \leq \frac{2+\varepsilon-\sqrt{\varepsilon^2+4}}{2}.$$

□

The Hessian matrix of $F_\varepsilon$ at $(x,y)$ is

$$\nabla^2 F_\varepsilon(x,y) = \begin{pmatrix} \cos(x)^2 + \sin(x)(y-\sin(x)+\varepsilon & -\cos(x) \\ -\cos(x) & 1 \end{pmatrix}.$$

The function is not globally $L$-smooth, because of the $y$ factor in $\nabla^2_{1,1} F_\varepsilon(x,y)$. However, for a fixed $y$, the eigenvalues are bounded for any $x \in \mathbb{R}$.

**Implementation Details** We fix $\varepsilon = 10^{-3}$, and chose $(0,3)$ as the initialization point. We compute the $L$-smooth value of $F_{0.001}$ on the domain $[-2\pi, 2\pi] \times [-3,3]$, using a grid of $10^6$ equally spaced points. We then fix $\gamma = 1/L$ for both (GD) and (NM').

### 5.B.3 Proof of Proposition 5.3.1

The definition of the hard instance is rather involved. The core object of its definition is the following function

$$g_{T,t}(x) = q_{T,t}(x) + \sum_{i=1}^{Tt} v_{y_i}(x_i),$$

where $x_i$ is the $i$-th coordinate of $x \in \mathbb{R}^{Tt} \to \mathbb{R}$, where $q_{T,t}(x)$ is defined as follows

$$q_{T,t}(x) = \frac{1}{2}\sum_{i=0}^{t-1}\left[\left(\frac{7}{8}x_{iT} - x_{iT+1}\right)^2 + \sum_{j=1}^{T-1}(x_{iT+j+1} - x_{iT+j})^2\right],$$

and for all $y \in \mathbb{R}$, $v_y(t)$ is the following one-dimensional function

$$v_y(t) = \begin{cases} \frac{1}{2}t, & t \leq \frac{31}{32}y, \\ \frac{1}{2}t^2 - 16(t-\frac{31}{32}y)^2, & \frac{31}{32}y < t \leq y, \\ \frac{1}{2}t^2 - \frac{y^2}{32} + 16(t-\frac{33}{32}y)^2, & y < t \leq \frac{33}{32}y, \\ \frac{1}{2}t^2 - \frac{y^2}{32}, & t > \frac{33}{32}y. \end{cases}$$

The function $q_{T,t}$ belongs to the class of zero-chain functions (Nesterov, 2013; Carmon et al., 2020, 2021a), which are hard to minimize with first-order algorithms. The key idea is that for these functions, if starting from 0, first-order algorithms will be able to explore only one new coordinate at each iteration for a number of iterations equal to the dimension of the problem. Here, the $v_y(\cdot)$ components introduce non-convexity. Up to a rescaling, the function $g_{T,t}$ belongs to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$ for desired constants $\mu$ and $L$, see all details in (Yue et al., 2023).

*Proof of Proposition 5.3.1.* The function $q_{T,t}$ is a strongly convex quadratic function with unique minimizer 0, such that it belongs to $\mathrm{SQC}^\mu_\tau$ for some parameters. The $v_y(\cdot)$ components introduces non-convexity. For any $y \in \mathbb{R}$, the function $v_y(\cdot)$ has zero as its unique minimizer, and has it belongs to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$ for some parameter, which implies it belongs to $\mathrm{SQC}^\mu_\tau \cap \mathrm{LS}^L$ because the function is one-dimensional. It follows that as a sum of one-dimensional functions in $\mathrm{SQC}^\mu_\tau \cap \mathrm{LS}^L$, the function $\sum_{i=1}^{Tt} v_{y_i}(x_i)$ belongs to $\mathrm{SQC}^\mu_\tau \cap \mathrm{LS}^L$ for some parameters, with unique minimizer $0 \in \mathbb{R}^d$. Then, by sum of strongly quasar convex functions with same minimizer, $g_{t,T}$ is strongly quasar convex (Hinder et al., 2020, Appendix D.3). Finally, to have the function $g_{T,t}$ belongs to $\mathrm{PL}^\mu \cap \mathrm{LS}^L$ for desired constants $\mu$ and $L$, it remains to rescale it as $\tilde{g}(x) = a_0 g_{t,T}(y - a_1 x)$ for some constants $a_0, a_1 \in \mathbb{R}$ and where $y = (y_1, \ldots, y_{Tt})$, see all details in (Yue et al., 2023). Such rescaling preserve the strong quasar convexity property, see (Hinder et al., 2020, Appendix D.3).

### 5.B.4 Supplementary Figures

In this section, we provide some figures that are not in the main text, due to available space.

**About Figure 5.6** We display in Figure 5.6 two synthetic examples of functions belonging to $\mathrm{SQC}_\tau^\mu$, in dimension one and two. The one-dimensional example is

$$f(t) = 0.5 \cdot (t + 0.15 \sin(5t))^2.$$

The two dimensional example is $h : \mathbb{R}^2 \to \mathbb{R}$, built as follows:

$$h(x) = f(\|x\|) g\left(\frac{x}{\|x\|}\right) \tag{5.11}$$

where $f(t) = t^2$ and

$$g(x_1, x_2) = \frac{1}{4N} \sum_{i=1}^{N} \left(a_i \sin(b_i x_1)^2 + c_i \cos(d_i x_2)^2\right) + 1 \tag{5.12}$$

with $N = 10$ and the $\{a_i\}_i$, $\{c_i\}_i$ are independently and uniformly distributed on $[0, 20]$, and the $\{b_i\}_i$, $\{d_i\}_i$ are independently and uniformly distributed on $[-25, 25]$.

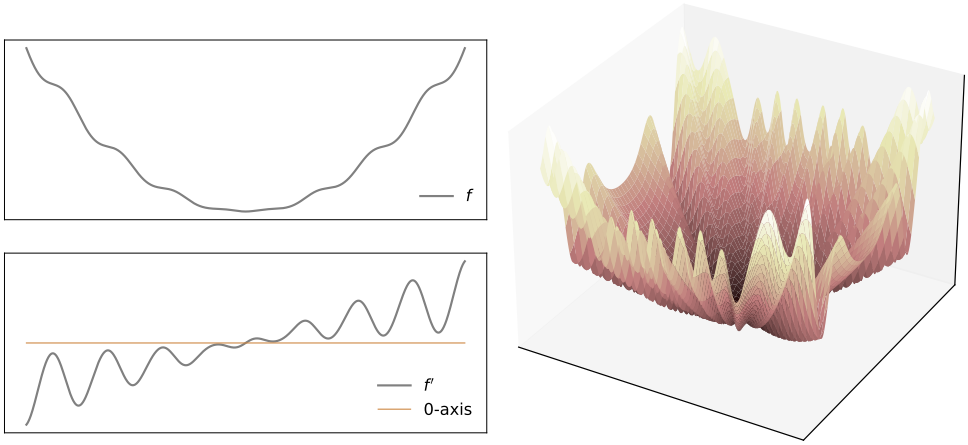



Figure 5.6: Visualization of functions belonging in $\mathrm{SQC}_\tau^\mu$, for some parameters. See details in Appendix 5.B.4.

**About Figure 5.7** Figure 5.7 displays a simple one-dimensional example where $f \in \mathrm{LS}^L \cap \mathrm{QG}_+^{L_0}$ with $L$ significantly larger than $L_0$, and similarly for $f \in \mathrm{PL}^\mu \cap \mathrm{QG}_-^{\mu_0}$. Such differences are created by local variations outside of the vicinity of the minimizer. This is because the parameterizing $\mathrm{QG}_+^{L_0}$ and $\mathrm{QG}_-^{\mu_0}$ is mainly affected by the geometry in the vicinity of the minimizer, while $\mathrm{LS}^L$ and $\mathrm{PL}^\mu$ are strongly sensitive to variations at any points.

**About Figure 5.8** Figure 5.8 illustrates that, for a function belonging to $\mathrm{PL}^\mu$, the optimal $\mu$ may depend on the local smaller curvature at a minimizer $x^*$, or by plateaus.

## 5.C Technical Lemmas

The following lemmas serve as technical tools for our convergence proofs.

**Lemma 5.C.1.** *Let $f \in PL^\mu \cap QG_+^{L_0}$. Noting $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$, we have for all $x \in \mathbb{R}^d$,*

$$\|x - x^*\| \, \|\nabla f(x)\| \geq \sqrt{\frac{\mu}{L_0}}(f(x) - f^*) + \frac{\sqrt{\mu\mu_0}}{2} \|x - x^*\|^2 .$$

*Proof.* If $x = x^*$, the statement is trivial. Assume $x \neq x^*$ and write

$$\|x - x^*\| \, \|\nabla f(x)\| = \|x - x^*\| \sqrt{f(x) - f^*} \, \frac{\|\nabla f(x)\|}{\sqrt{f(x) - f^*}}.$$

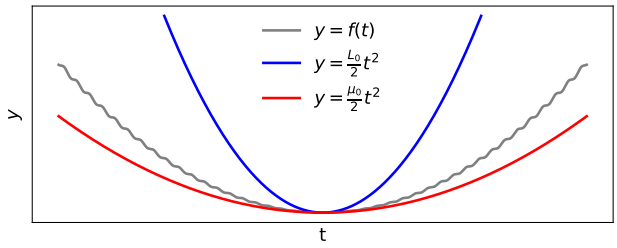



Figure 5.7: The grey curve is $f(t) = 0.5 \cdot 5(t + 0.07\sin(13t))^2$. The blue (red) curve is a quadratic upper (lower) bound. We compute the largest $\mu$ and smallest $L$ such that $f \in \mathrm{PL}^\mu \cap \mathrm{LS}^L$ on the displayed domain, numerically evaluated at $\mu \approx 4 \cdot 10^{-2}$, $L \approx 6 \cdot 10^2$. Also, the parameters $L_0$ and $\mu_0$ that parameterize the quadratic bounds are $\mu_0 \approx 3$ and $L_0 \approx 18$. On this prototype 1-dimensional example, $\frac{\mu_0}{L_0} \approx 0.2$ while $\frac{\mu}{L} \approx 7 \cdot 10^{-5}$. This highlight the possibility of a significant gap between these two ratios.

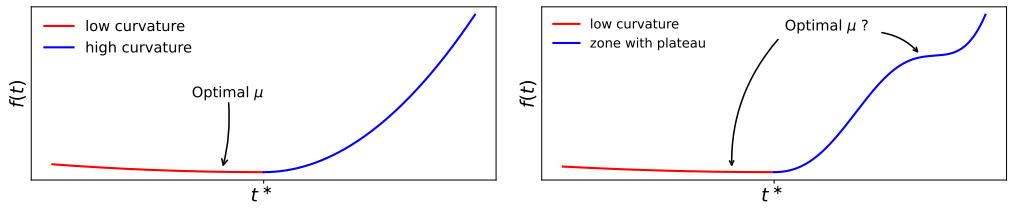



Figure 5.8: Left: Simple 1-dimensional functions in $\mathrm{SC}^\mu$, so also in $\mathrm{PL}^\mu$. The worst approaching direction toward the minimizer $t^*$ is the red zone, which determines the optimal $\mu$ constant. Right: Simple non-convex function that belongs to $\mathrm{PL}^\mu$ for some $\mu > 0$. The $\mu$ constant could be determined either by the local flatness around the minimizer $t^*$ (red zone), or by the flatness of the plateau (blue zone).

**(1) Step 1 ($\mathrm{PL}^\mu$ and $\mathrm{QG}_+^{L_0}$).** Since $f \in \mathrm{PL}^\mu$, $\|\nabla f(x)\| \geq \sqrt{2\mu}\sqrt{f(x) - f^*}$, and since $f \in \mathrm{QG}_+^{L_0}$,

$$f(x) - f^* \leq \frac{L_0}{2}\|x - x^*\|^2 \quad \Rightarrow \quad \|x - x^*\| \geq \sqrt{\frac{2}{L_0}}\sqrt{f(x) - f^*}.$$

Therefore,

$$\|x - x^*\|\,\|\nabla f(x)\| \geq \sqrt{\frac{2}{L_0}}(f(x) - f^*)\sqrt{2\mu} = 2\sqrt{\frac{\mu}{L_0}}(f(x) - f^*). \tag{5.13}$$

**Step 2 ($\mathrm{PL}^\mu$ and $\mathrm{QG}_-^{\mu_0}$).** By definition of $\mu_0$

$$f(x) - f^* \geq \frac{\mu_0}{2}\|x - x^*\|^2 \quad \Rightarrow \quad \sqrt{f(x) - f^*} \geq \sqrt{\frac{\mu_0}{2}}\|x - x^*\|.$$

Combining with $\|\nabla f(x)\| \geq \sqrt{2\mu}\sqrt{f(x) - f^*}$ gives

$$\|x - x^*\|\,\|\nabla f(x)\| \geq \sqrt{\mu\mu_0}\,\|x - x^*\|^2. \tag{5.14}$$

**Combine the bounds** From (5.13) and the previous bound,

$$\|x - x^*\|\,\|\nabla f(x)\| \geq \max\Big\{2\sqrt{\frac{\mu}{L_0}}(f(x) - f^*),\ \sqrt{\mu\mu_0}\,\|x - x^*\|^2\Big\}.$$

Using $\max\{u, v\} \geq \frac{u+v}{2}$ for $u, v \geq 0$, we obtain

$$\|x - x^*\|\,\|\nabla f(x)\| \geq \sqrt{\frac{\mu}{L_0}}(f(x) - f^*) + \frac{\sqrt{\mu\mu_0}}{2}\|x - x^*\|^2,$$

which concludes the proof. □

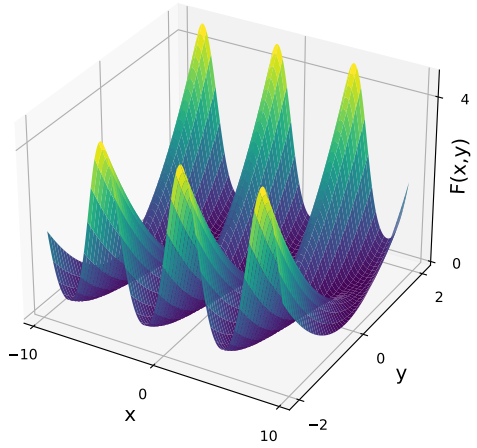



Figure 5.9: 3-dimensional visualization of $F_{0.001}(x, y) = 0.5(y - \sin(x))^2 + 0.001 \cdot 0.5x^2$, considered in Section 5.7

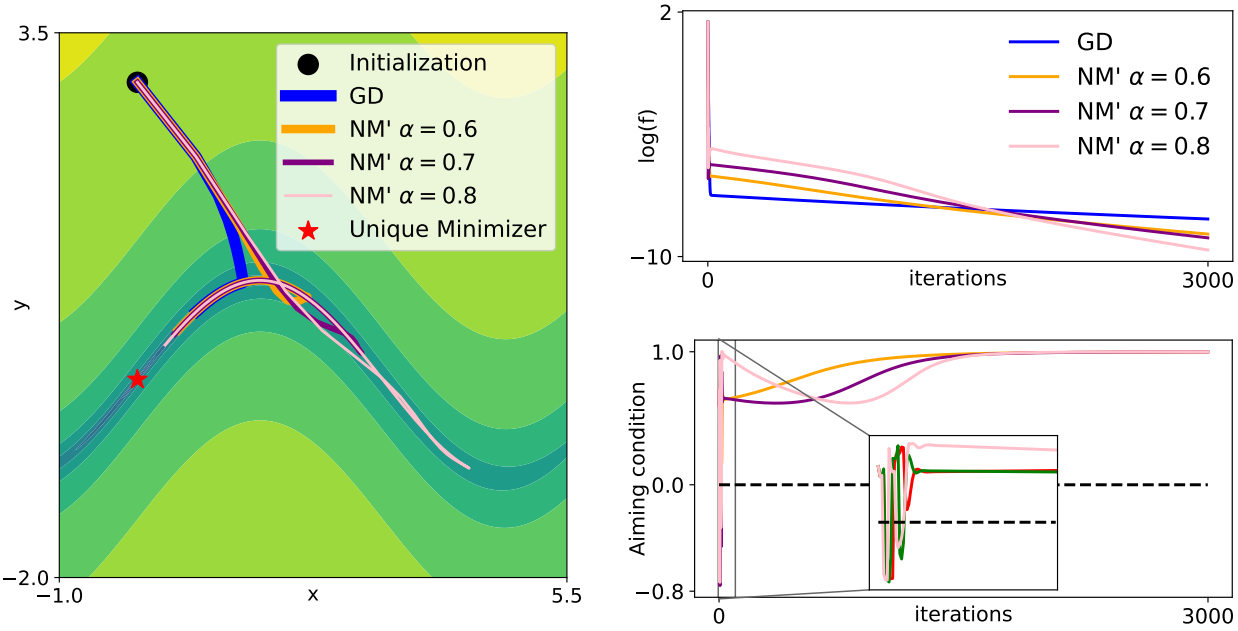



Figure 5.10: Plot of the 3000 first iterations of the trajectories of (GD) and (NM') on $f$, when minimizing $F_{0.001}(x, y) = 0.5(y - \sin(x))^2 + 0.001 \cdot 0.5x^2$. This is a version of Figure 5.3 with 3000 iterations instead of 1000. Observe that the decrease of the momentum method (NM') catches up with the one of (GD), which is coherent with the values of the aiming condition.

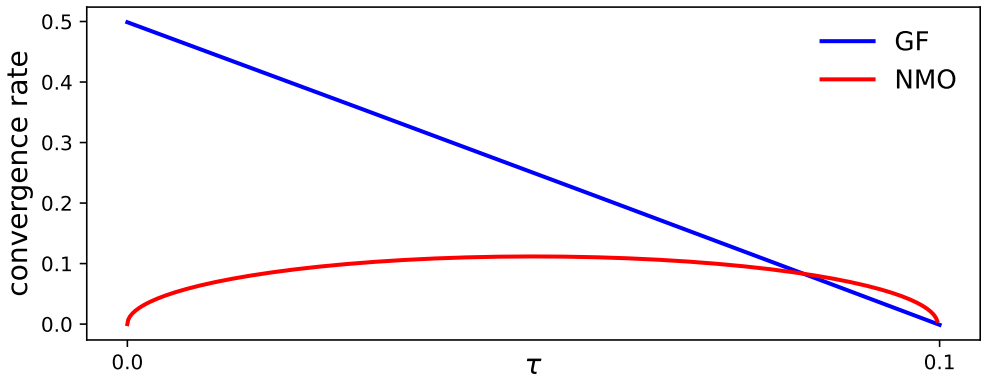



Figure 5.11: For a range of parameters $\tau$, we compute the highest admissible $\mu$ ensuring $f \in \mathrm{SQC}_\tau^\mu$, and plot the associated theoretical convergence rates of (GF) and (NMO) under $\mathrm{SQC}_\tau^\mu$, namely $\tau\mu$ and $\tau\sqrt{\mu}$, with $f(t) = 5(t + 0.19\sin(5t))^2$. See details in Appendix 5.B.1. This is the continuous counterpart of Figure 5.2. Similarly to the discrete setting, the pair $(\tau, \mu)$ that maximizes the convergence rate differ in both cases.

**Lemma 5.C.2.** *Assume $x \neq x^*$ and $f(x) \neq f^*$. Define the point-wise quantities*

$$\mu(x) = \frac{1}{2}\frac{\|\nabla f(x)\|^2}{f(x) - f^*}, \qquad a(x) = \frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\| \, \|x - x^*\|}, \qquad \mu_0(x) = 2\frac{f(x) - f^*}{\|x - x^*\|^2}.$$

*Then,*

$$\langle \nabla f(x), x^* - x \rangle = -a(x)\sqrt{\frac{\mu(x)}{\mu_0(x)}}(f(x) - f^*) - \frac{a(x)}{2}\sqrt{\mu(x)\mu_0(x)}\, \|x - x^*\|^2 .$$

*If $f \in PL^\mu$, it follows*

$$\langle \nabla f(x), x^* - x \rangle = -a(x)\sqrt{\frac{\mu}{\mu_0}}(f(x) - f^*) - \frac{a(x)}{2}\sqrt{\mu\mu_0}\, \|x - x^*\|^2 ,$$

*with $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$.*

*Proof.* By definition of $a(x)$,

$$\langle \nabla f(x), x^* - x \rangle = -a(x) \, \|\nabla f(x)\| \, \|x - x^*\| .$$

Moreover, by the definitions of $\mu(x)$ and $\mu_0(x)$,

$$\|\nabla f(x)\| = \sqrt{2\mu(x)}\sqrt{f(x) - f^*}, \qquad \|x - x^*\| = \sqrt{\frac{2}{\mu_0(x)}}\sqrt{f(x) - f^*}.$$

Hence

$$\|\nabla f(x)\| \, \|x - x^*\| = 2\sqrt{\frac{\mu(x)}{\mu_0(x)}}\, (f(x) - f^*).$$

Also, using $\|\nabla f(x)\| = \sqrt{2\mu(x)}\sqrt{f(x) - f^*} = \sqrt{2\mu(x)}\sqrt{\frac{\mu_0(x)}{2}}\, \|x - x^*\|$,

$$\|\nabla f(x)\| \, \|x - x^*\| = \sqrt{\mu(x)\mu_0(x)}\, \|x - x^*\|^2 .$$

Combining the two equal expressions and using $\xi = \frac{1}{2}\xi + \frac{1}{2}\xi$ with $\xi = \|\nabla f(x)\| \, \|x - x^*\|$ gives

$$\|\nabla f(x)\| \, \|x - x^*\| = \sqrt{\frac{\mu(x)}{\mu_0(x)}}\, (f(x) - f^*) + \sqrt{\mu(x)\mu_0(x)}\, \frac{1}{2}\, \|x - x^*\|^2 ,$$

and multiplying by $-a(x)$ yields the claim. □

## 5.D Proofs of Section 5.5

This section is dedicated to proving the results of Section 5.5. In the discrete setting, we will actually prove a more general result, replacing the gradient with a stochastic estimator, under a strong growth condition, also sometimes named multiplicative noise assumption.

**Assumption 5.D.1** (Strong growth condition)**.** *We assume we have access to stochastic approximations $\nabla f(x,\xi)$ of the real gradient $\nabla f(x)$, where $\xi \sim \mathcal{P}(\Xi)$, such that*

*(i) (Unbiased estimator)* $\forall x \in \mathbb{R}^d,\ \mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla f(x)$.

*(ii) (Strong Growth Condition)* $\exists \rho \geq 1, \forall x \in \mathbb{R}^d,\ \mathbb{E}_\xi\left[\|\nabla f(x,\xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2$.

To our knowledge, this assumption was introduced in (Polyak, 1987), and further used by (Cevher and Vu, 2019) and (Vaswani et al., 2019) as a relaxation of the maximal strong growth condition (Tseng, 1998; Solodov, 1998). We avoid this more general setting in the main text, to discard adding a layer of complexity. The deterministic case is a special instance of Assumption 5.D.1, corresponding to $\rho = 1$. For functions in $\mathrm{LS}^L$ that verify Assumption 5.D.1, we have a stochastic descent lemma, see (Gupta et al., 2024, Lemma 16).

**Lemma 5.D.1.** *If $f \in LS^L$ and Assumption 5.D.1 holds, then we have*

$$\mathbb{E}_\xi\left[f\Big(x - \frac{1}{\rho L}\nabla f(x,\xi)\Big) - f(x)\right] \leq -\frac{1}{2\rho L}\|\nabla f(x)\|^2. \tag{5.15}$$

In the subsequent sections, for an algorithm $\{x_n\}_{n\geq 0}$ we denote $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. For a sequence $\{\xi_k\}_{k\in\mathbb{N}}$ of random variables, we note $\mathcal{F}_k = \sigma(\xi_0, \ldots, \xi_k)$ the filtration generated by $\xi_0, \ldots, \xi_k$.

### 5.D.1 Proof of Theorem 5.5.1

**(i)** $f \in \mathbf{PL}^\mu$ The following proof is classical (Polyak and Shcherbakov, 2017). We provide a proof the sake of completeness. Let

$$E_t = f(x_t) - f^*.$$

By definition of (GF), we have

$$\dot{E}_t = \langle \nabla f(x_t), \dot{x}_t\rangle = -\|\nabla f(x_t)\|^2 \leq -2\mu\,(f(x_t) - f^*) = -2\mu\, E_t,$$

where we used $f \in \mathrm{PL}^\mu$. By integration,

$$f(x_t) - f^* \leq e^{-2\mu t}(f(x_0) - f^*).$$

**(ii)** $f \in \mathbf{PL}^\mu \cap \mathbf{AC}^a \cap \mathbf{QG}^{L_0}_+$ Let

$$E_t = f(x_t) - f^* + \frac{\beta}{2}\|x_t - x^*\|^2.$$

Then

$$\begin{aligned}\dot{E}_t &= \langle \nabla f(x_t), \dot{x}_t\rangle + \beta\langle x_t - x^*, \dot{x}_t\rangle \\ &= -\|\nabla f(x_t)\|^2 - \beta\langle \nabla f(x_t), x_t - x^*\rangle \\ &\leq -\beta\langle \nabla f(x_t), x_t - x^*\rangle = \beta\langle \nabla f(x_t), x^* - x_t\rangle.\end{aligned}$$

By Lemma 5.4.1, for all $t \geq 0$,

$$\langle \nabla f(x_t), x^* - x_t\rangle \leq -a\sqrt{\frac{\mu}{L_0}}\,(f(x_t) - f^*) - \frac{a}{2}\sqrt{\mu\mu_0}\,\|x_t - x^*\|^2,$$

hence

$$\dot{E}_t \leq -\beta a\sqrt{\frac{\mu}{L_0}}\,(f(x_t) - f^*) - \frac{\beta a}{2}\sqrt{\mu\mu_0}\,\|x_t - x^*\|^2.$$

Choose $\beta$ so that the two coefficients match:

$$\beta a\sqrt{\frac{\mu}{L_0}} = a\sqrt{\mu\mu_0} \quad \Longrightarrow \quad \beta = \sqrt{\mu_0 L_0}.$$

Then

$$\dot{E}_t \leq -a\sqrt{\mu\mu_0}\Big(f(x_t) - f^* + \frac{\beta}{2}\left\|x_t - x^*\right\|^2\Big) = -a\sqrt{\mu\mu_0}\, E_t,$$

and by integration,

$$E_t \leq e^{-a\sqrt{\mu\mu_0}\, t} E_0.$$

Since $f \in \mathrm{QG}_-^{\mu_0}$, we have

$$\frac{\beta}{2}\left\|x_0 - x^*\right\|^2 \leq \beta\, \frac{1}{\mu_0}\left(f(x_0) - f^*\right) = \sqrt{\frac{L_0}{\mu_0}}\left(f(x_0) - f^*\right),$$

so $E_0 \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)\left(f(x_0) - f^*\right)$. Finally, since $f(x_t) - f^* \leq E_t$, we deduce

$$f(x_t) - f^* \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right) e^{-a\sqrt{\mu\mu_0}\, t}\left(f(x_0) - f^*\right).$$

### 5.D.2 Proof of Theorem 5.5.2

We prove the following more general result.

**Theorem 5.D.1.** *Let $f \in LS^L$. Let $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ follow* (GD) *with $\nabla f(\tilde{x}_k)$ replaced by $\nabla f(\tilde{x}_k, \xi_k)$, where $\xi_k$ is independent of $\tilde{x}_k$, and assume Assumption 5.D.1. Let $\gamma = \frac{1}{\rho L}$.*

*(i) If $f \in PL^\mu$, then*

$$\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] \leq \left(1 - \frac{\mu}{\rho L}\right)^k \left(f(\tilde{x}_0) - f^*\right).$$

*(ii) If $f \in PL^\mu \cap AC^a \cap QG_+^{L_0}$, then*

$$\mathbb{E}\left[f(\tilde{x}_k) - f^*\right] \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)\left(1 - \frac{a\sqrt{\mu\mu_0}}{\rho L}\right)^k \left(f(\tilde{x}_0) - f^*\right),$$

*with $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$ and $L_0 := \inf\{L' \leq L : f \in QG_+^{L'}\}$.*

The result for Theorem 5.5.2 is deduced in the case $\rho = 1$, which removes the expectations.

*Proof of Theorem 5.D.1.* **(i)** $f \in \mathbf{PL}^\mu$ This result and proof is well known, see by instance (Karimi et al., 2016) in the deterministic setting, (Vaswani et al., 2019) for the stochastic case. We provide a proof for completeness. Define

$$E_k = f(\tilde{x}_k) - f^*.$$

By Lemma 5.D.1,

$$\mathbb{E}_n\left[E_{k+1} - E_k\right] = \mathbb{E}_n\left[f(\tilde{x}_{k+1}) - f(\tilde{x}_k)\right] \leq -\frac{1}{2\rho L}\left\|\nabla f(\tilde{x}_k)\right\|^2.$$

Since $f \in \mathrm{PL}^\mu$, $\left\|\nabla f(\tilde{x}_k)\right\|^2 \geq 2\mu\left(f(\tilde{x}_k) - f^*\right) = 2\mu E_k$, hence

$$\mathbb{E}_n\left[E_{k+1}\right] \leq \left(1 - \frac{\mu}{\rho L}\right) E_k.$$

By induction and taking expectation,

$$\mathbb{E}\left[E_k\right] \leq \left(1 - \frac{\mu}{\rho L}\right)^k E_0,$$

which is the claim.

**(ii)** $f \in \mathbf{PL}^{\mu} \cap \mathbf{AC}^{a} \cap \mathbf{QG}_{+}^{L_0}$ Set $\beta := \sqrt{\mu_0 L_0}$ and define

$$E_k = f(\tilde{x}_k) - f^* + \frac{\beta}{2} \|\tilde{x}_k - x^*\|^2 .$$

Using the update $\tilde{x}_{k+1} = \tilde{x}_k - \gamma \nabla f(\tilde{x}_k, \xi_k)$, we expand

$$\|\tilde{x}_{k+1} - x^*\|^2 - \|\tilde{x}_k - x^*\|^2 = -2\gamma \langle \tilde{x}_k - x^*, \nabla f(\tilde{x}_k, \xi_k) \rangle + \gamma^2 \|\nabla f(\tilde{x}_k, \xi_k)\|^2 .$$

Taking $\mathbb{E}_n[\cdot]$, using Assumption 5.D.1 (unbiasedness + SGC) and Lemma 5.D.1, we obtain

$$\begin{aligned}\mathbb{E}_n[E_{k+1} - E_k] &\leq -\frac{1}{2\rho L} \|\nabla f(\tilde{x}_k)\|^2 + \frac{\beta}{2}\Big(-2\gamma \langle \tilde{x}_k - x^*, \nabla f(\tilde{x}_k) \rangle + \gamma^2 \rho \|\nabla f(\tilde{x}_k)\|^2\Big) \\ &= \Big(-\frac{1}{2\rho L} + \frac{\beta}{2}\gamma^2 \rho\Big) \|\nabla f(\tilde{x}_k)\|^2 - \beta\gamma \langle \tilde{x}_k - x^*, \nabla f(\tilde{x}_k) \rangle .\end{aligned}$$

With $\gamma = \frac{1}{\rho L}$, we have $\gamma^2 \rho = \frac{1}{\rho L^2}$, hence

$$-\frac{1}{2\rho L} + \frac{\beta}{2}\gamma^2 \rho = -\frac{1}{2\rho L} + \frac{\beta}{2\rho L^2} = -\frac{1}{2\rho L}\left(1 - \frac{\beta}{L}\right) \leq 0,$$

because $\beta = \sqrt{\mu_0 L_0} \leq L_0 \leq L$. Therefore,

$$\mathbb{E}_n[E_{k+1} - E_k] \leq -\frac{\beta}{\rho L} \langle \tilde{x}_k - x^*, \nabla f(\tilde{x}_k) \rangle = \frac{\beta}{\rho L} \langle \nabla f(\tilde{x}_k), x^* - \tilde{x}_k \rangle .$$

By Lemma 5.4.1,

$$\langle \nabla f(\tilde{x}_k), x^* - \tilde{x}_k \rangle \leq -a\sqrt{\frac{\mu}{L_0}} (f(\tilde{x}_k) - f^*) - \frac{a}{2}\sqrt{\mu\mu_0} \|\tilde{x}_k - x^*\|^2 .$$

Multiplying by $\beta/(\rho L)$ and using $\beta\sqrt{\mu/L_0} = \sqrt{\mu\mu_0}$ (since $\beta = \sqrt{\mu_0 L_0}$), we get

$$\mathbb{E}_n[E_{k+1} - E_k] \leq -\frac{a\sqrt{\mu\mu_0}}{\rho L} (f(\tilde{x}_k) - f^*) - \frac{a\sqrt{\mu\mu_0}}{\rho L} \frac{\beta}{2} \|\tilde{x}_k - x^*\|^2 = -\frac{a\sqrt{\mu\mu_0}}{\rho L} E_k.$$

Hence

$$\mathbb{E}_n[E_{k+1}] \leq \left(1 - \frac{a\sqrt{\mu\mu_0}}{\rho L}\right) E_k,$$

taking the expectation and by induction,

$$\mathbb{E}[E_k] \leq \left(1 - \frac{a\sqrt{\mu\mu_0}}{\rho L}\right)^k E_0.$$

Finally, since $f(\tilde{x}_k) - f^* \leq E_k$, and using $f \in \mathrm{QG}_{-}^{\mu_0}$,

$$\frac{\beta}{2} \|\tilde{x}_0 - x^*\|^2 \leq \beta \frac{1}{\mu_0}(f(\tilde{x}_0) - f^*) = \sqrt{\frac{L_0}{\mu_0}}(f(\tilde{x}_0) - f^*),$$

so $E_0 \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(\tilde{x}_0) - f^*)$. This yields the claim. □

### 5.D.3 Proof of Theorem 5.5.3

**(i)** Assume that $\eta_t \equiv \eta$, $\eta'_t \equiv \eta'$ are constants. Let

$$E_t = e^{\theta t}\Big(f(x_t) - f^* + \beta\frac{1}{2}\left\|z_t - x^*\right\|^2\Big). \tag{5.16}$$

Then $E_t = E_0 + \int_0^t \dot{E}_s\, ds$, where

$$\begin{aligned}
\dot{E}_t &= \theta e^{\theta t}(f(x_t) - f^*) + \theta\beta\frac{e^{\theta t}}{2}\left\|z_t - x^*\right\|^2 + e^{\theta t}\left\langle \nabla f(x_t), \dot{x}_t\right\rangle + \beta e^{\theta t}\left\langle z_t - x^*, \dot{z}_t\right\rangle \\
&= \theta e^{\theta t}(f(x_t) - f^*) + \theta\beta\frac{e^{\theta t}}{2}\left\|z_t - x^*\right\|^2 + e^{\theta t}\left\langle \nabla f(x_t), \eta(z_t - x_t) - \gamma\nabla f(x_t)\right\rangle \\
&\qquad + \beta e^{\theta t}\left\langle z_t - x^*, \eta'(x_t - z_t) - \gamma'\nabla f(x_t)\right\rangle \\
&= \theta e^{\theta t}(f(x_t) - f^*) + \theta\beta\frac{e^{\theta t}}{2}\left\|z_t - x^*\right\|^2 - \gamma e^{\theta t}\left\|\nabla f(x_t)\right\|^2 \\
&\qquad + e^{\theta t}\eta\left\langle \nabla f(x_t), z_t - x_t\right\rangle + \beta\eta' e^{\theta t}\left\langle z_t - x^*, x_t - z_t\right\rangle - \beta\gamma' e^{\theta t}\left\langle \nabla f(x_t), z_t - x^*\right\rangle.
\end{aligned} \tag{5.17}$$

Using the decomposition

$$\left\langle \nabla f(x_t), z_t - x_t\right\rangle = \left\langle \nabla f(x_t), x^* - x_t\right\rangle + \left\langle \nabla f(x_t), z_t - x^*\right\rangle,$$

and the identity

$$\left\langle z_t - x^*, x_t - z_t\right\rangle = \frac{1}{2}\left\|x_t - x^*\right\|^2 - \frac{1}{2}\left\|z_t - x^*\right\|^2 - \frac{1}{2}\left\|x_t - z_t\right\|^2 \le \frac{1}{2}\left\|x_t - x^*\right\|^2 - \frac{1}{2}\left\|z_t - x^*\right\|^2,$$

we obtain from (5.17)

$$\begin{aligned}
e^{-\theta t}\dot{E}_t \le \theta(f(x_t) - f^*) &+ \frac{\beta}{2}\big(\theta - \eta'\big)\left\|z_t - x^*\right\|^2 + \frac{\beta\eta'}{2}\left\|x_t - x^*\right\|^2 \\
&+ \eta\left\langle \nabla f(x_t), x^* - x_t\right\rangle + (\eta - \beta\gamma')\left\langle \nabla f(x_t), z_t - x^*\right\rangle - \gamma\left\|\nabla f(x_t)\right\|^2.
\end{aligned} \tag{5.18}$$

We choose $\eta' = \theta$ to cancel the $\left\|z_t - x^*\right\|^2$ term, and we choose $\eta = \beta\gamma'$ to cancel the $\left\langle \nabla f(x_t), z_t - x^*\right\rangle$ term. Dropping the negative term $-\gamma\left\|\nabla f(x_t)\right\|^2$, we get

$$e^{-\theta t}\dot{E}_t \le \theta(f(x_t) - f^*) + \eta\left\langle \nabla f(x_t), x^* - x_t\right\rangle + \frac{\beta\theta}{2}\left\|x_t - x^*\right\|^2. \tag{5.19}$$

By Lemma 5.4.1,

$$\left\langle \nabla f(x_t), x^* - x_t\right\rangle \le -a\sqrt{\frac{\mu}{L_0}}(f(x_t) - f^*) - \frac{a}{2}\sqrt{\mu\mu_0}\left\|x_t - x^*\right\|^2.$$

Plugging into (5.19) yields

$$e^{-\theta t}\dot{E}_t \le \Big(\theta - \eta a\sqrt{\frac{\mu}{L_0}}\Big)(f(x_t) - f^*) + \frac{1}{2}\Big(\beta\theta - \eta a\sqrt{\mu\mu_0}\Big)\left\|x_t - x^*\right\|^2. \tag{5.20}$$

We impose the cancellation of both coefficients in (5.20):

$$\theta = \eta a\sqrt{\frac{\mu}{L_0}} \qquad \text{and} \qquad \beta\theta = \eta a\sqrt{\mu\mu_0}.$$

These two equalities are equivalent to $\beta = \sqrt{\mu_0 L_0}$ and $\theta = \eta a\sqrt{\mu/L_0}$. With the additional choice $\eta = \beta\gamma'$, we may take

$$\beta = \sqrt{\mu_0 L_0}, \qquad \gamma' = \beta^{-1/2} = (\mu_0 L_0)^{-1/4}, \qquad \eta = \beta\gamma' = (\mu_0 L_0)^{1/4}, \qquad \eta' = \theta = a\Big(\frac{\mu_0}{L_0}\Big)^{1/4}\sqrt{\mu}.$$

With these parameters, (5.20) gives $e^{-\theta t}\dot{E}_t \leq 0$, hence $E_t \leq E_0$. Therefore, using (5.16),

$$f(x_t) - f^* \leq e^{-\theta t} E_t \leq e^{-\theta t} E_0 = \exp\Big(-a\Big(\frac{\mu_0}{L_0}\Big)^{1/4}\sqrt{\mu}\, t\Big)\, E_0.$$

Finally, using $x_0 = z_0$ and $f \in \mathrm{QG}_-^{\mu_0}$,

$$\frac{\beta}{2}\left\|x_0 - x^*\right\|^2 \leq \beta\,\frac{1}{\mu_0}(f(x_0) - f^*) = \sqrt{\frac{L_0}{\mu_0}}(f(x_0) - f^*),$$

so $E_0 \leq \left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(x_0) - f^*)$, yields the claim.

**(ii)** The proof in the discrete case exploits the continuized framework. This implies that the following analysis differs from the previous ones. We refer to Appendix 5.F to a brief introduction of the tools needed when carrying such analysis, or Chapter 4 for a more detailed introduction. Also, we will prove the following more general result.

**Theorem 5.D.2.** *Let $f \in PL^\mu \cap AC^a \cap LS^L$. Denote $L_0 := \inf\{L' \leq L : f \in QG_+^{L'}\}$, $\mu_0 := \sup\{\mu' \geq \mu : f \in QG_-^{\mu'}\}$. Let $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ verify* (NM) *with the continuized parameterization and with $\nabla f(\tilde{x}_k)$ replaced by $\nabla f(\tilde{x}_k, \xi_k)$ for random variables $\xi_k$ independent of $\tilde{x}_k$ and under Assumption 5.D.1. Let $\eta' = \frac{a}{\rho}\left(\frac{\mu_0}{L_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}$, $\gamma = \frac{1}{\rho L}$, $\gamma' = \frac{1}{\rho}\frac{1}{(\mu_0 L_0)^{1/4}\sqrt{L}}$, and $\eta = \frac{1}{\rho}\frac{(\mu_0 L_0)^{1/4}}{\sqrt{L}}$. Then, with probability $1 - \frac{1}{c_0} - \exp(-(c_1 - 1 - \log(c_1))k)$, for some $c_0 > 1$ and $c_1 \in (0,1)$ we have*

$$f(\tilde{x}_k) - f^* \leq K_1 e^{-\frac{a}{\rho}\left(\frac{\mu_0}{L_0}\right)^{1/4}\sqrt{\frac{\mu}{L}}(1-c_1)k},$$

*where $K_1 := c_0\left(1 + \sqrt{\frac{L_0}{\mu_0}}\right)(f(x_0) - f^*)$.*

The proof of Theorem 5.5.3 (ii) is deduced in the case $\rho = 1$.

*Proof of Theorem 5.D.2.* Let $\overline{x}_t = (t, x_t, z_t)$. It satisfies $d\overline{x}_t = b(\overline{x}_t)dt + G(\overline{x}_t, \xi)dN(t,\xi)$, where

$$b(\overline{x}_t) = \begin{pmatrix} 1 \\ \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \qquad G(\overline{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma\nabla f(x_t, \xi) \\ -\gamma'\nabla f(x_t, \xi) \end{pmatrix}.$$

We apply Proposition 5.F.1 to $\phi(t) = \varphi(\overline{x}_t)$, where

$$\varphi(t, x, z) = A_t(f(x) - f^*) + \frac{B_t}{2}\left\|z - x^*\right\|^2.$$

Hence

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \langle\nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle\, ds + \int_0^t \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s, \xi)) - \varphi(\overline{x}_s)]\, ds + M_t, \tag{5.21}$$

where $M_t$ is a martingale.

We have

$$\frac{\partial\varphi}{\partial s}(\overline{x}_s) = \frac{dA_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{dB_s}{ds}\left\|z_s - x^*\right\|^2, \qquad \frac{\partial\varphi}{\partial x}(\overline{x}_s) = A_s\nabla f(x_s), \qquad \frac{\partial\varphi}{\partial z}(\overline{x}_s) = B_s(z_s - x^*).$$

So,

$$\begin{aligned}\langle\nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle = \frac{dA_s}{ds}(f(x_s) - f^*) + \frac{1}{2}\frac{dB_s}{ds}\left\|z_s - x^*\right\|^2 \\ + A_s\eta\,\langle\nabla f(x_s), z_s - x_s\rangle + B_s\eta'\,\langle z_s - x^*, x_s - z_s\rangle.\end{aligned} \tag{5.22}$$

Also,

$$\begin{aligned}\mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)] &= \mathbb{E}_\xi[A_s(f(x_s - \gamma\nabla f(x_s,\xi)) - f(x_s))] \\ &\quad + \frac{B_s}{2}\mathbb{E}_\xi\left[\|z_s - \gamma'\nabla f(x_s,\xi) - x^*\|^2 - \|z_s - x^*\|^2\right] \\ &= A_s\mathbb{E}_\xi[f(x_s - \gamma\nabla f(x_s,\xi)) - f(x_s)] + \frac{B_s\gamma'^2}{2}\mathbb{E}_\xi\left[\|\nabla f(x_s,\xi)\|^2\right] \qquad (5.23) \\ &\quad + B_s\gamma'\left\langle \nabla f(x_s), x^* - z_s\right\rangle, \qquad (5.24)\end{aligned}$$

where we used $\mathbb{E}_\xi[\nabla f(x_s,\xi)] = \nabla f(x_s)$.

As $f \in \mathrm{LS}^L$ and Assumption 5.D.1 holds, by Lemma 5.D.1 (with $\gamma = \frac{1}{\rho L}$),

$$A_s\mathbb{E}_\xi[f(x_s - \gamma\nabla f(x_s,\xi)) - f(x_s)] \le -\frac{A_s}{2\rho L}\|\nabla f(x_s)\|^2.$$

Moreover, by the strong growth condition, $\mathbb{E}_\xi\left[\|\nabla f(x_s,\xi)\|^2\right] \le \rho\|\nabla f(x_s)\|^2$. Hence

$$(5.23) \le \frac{1}{2}\Big(\rho B_s\gamma'^2 - \frac{A_s}{\rho L}\Big)\|\nabla f(x_s)\|^2. \qquad (5.25)$$

We use the identity

$$\langle z_s - x^*, x_s - z_s\rangle = \frac{1}{2}\|x_s - x^*\|^2 - \frac{1}{2}\|z_s - x^*\|^2 - \frac{1}{2}\|x_s - z_s\|^2 \le \frac{1}{2}\|x_s - x^*\|^2 - \frac{1}{2}\|z_s - x^*\|^2. \qquad (5.26)$$

Also,

$$A_s\eta\left\langle \nabla f(x_s), z_s - x_s\right\rangle + B_s\gamma'\left\langle \nabla f(x_s), x^* - z_s\right\rangle = A_s\eta\left\langle \nabla f(x_s), x^* - x_s\right\rangle + (A_s\eta - B_s\gamma')\left\langle \nabla f(x_s), z_s - x^*\right\rangle. \qquad (5.27)$$

By Lemma 5.4.1,

$$\langle \nabla f(x_s), x^* - x_s\rangle \le -a\sqrt{\frac{\mu}{L_0}}\big(f(x_s) - f^*\big) - \frac{a}{2}\sqrt{\mu\mu_0}\,\|x_s - x^*\|^2.$$

Gathering (5.22), (5.24), (5.25), (5.26), and (5.27), we obtain

$$\begin{aligned}&\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)] \\ &\le \frac{1}{2}\left(\frac{dB_s}{ds} - B_s\eta'\right)\|z_s - x^*\|^2 + \frac{1}{2}\left(B_s\eta' - A_s\eta\, a\sqrt{\mu\mu_0}\right)\|x_s - x^*\|^2 + \left(\frac{dA_s}{ds} - A_s\eta\, a\sqrt{\frac{\mu}{L_0}}\right)\big(f(x_s) - f^*\big) \\ &\quad + (A_s\eta - B_s\gamma')\left\langle \nabla f(x_s), z_s - x^*\right\rangle + \frac{1}{2}\left(\rho B_s\gamma'^2 - \frac{A_s}{\rho L}\right)\|\nabla f(x_s)\|^2.\end{aligned}$$

**Parameter tuning** Set $A_s = e^{\theta s}$ and $B_s = \sqrt{\mu_0 L_0}\, e^{\theta s}$, with

$$\theta = \eta' = \frac{a}{\rho}\Big(\frac{\mu_0}{L_0}\Big)^{1/4}\sqrt{\frac{\mu}{L}}, \quad \gamma = \frac{1}{\rho L}, \quad \gamma' = \frac{1}{\rho}\frac{1}{(\mu_0 L_0)^{1/4}\sqrt{L}}, \quad \eta = \frac{1}{\rho}\frac{(\mu_0 L_0)^{1/4}}{\sqrt{L}}.$$

These parameters imply the cancellations

$$A_s\eta - B_s\gamma' = 0, \qquad (5.28)$$

$$\rho B_s\gamma'^2 - \frac{A_s}{\rho L} = 0, \qquad (5.29)$$

$$\frac{dB_s}{ds} - B_s\eta' = 0, \qquad (5.30)$$

$$\frac{dA_s}{ds} - A_s\eta\, a\sqrt{\frac{\mu}{L_0}} = 0, \qquad (5.31)$$

$$B_s\eta' - A_s\eta\, a\sqrt{\mu\mu_0} = 0. \qquad (5.32)$$

Therefore,

$$\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi[\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)] \le 0,$$

so from (5.21),

$$\varphi(\overline{x}_t) \le \varphi(\overline{x}_0) + M_t.$$

We evaluate at $t = T_k$, take expectation, and use Theorem 5.F.1 and Proposition 5.F.2 to get

$$\mathbb{E}\left[e^{\theta T_k}\big(f(\tilde{x}_k) - f^*\big)\right] \le \varphi(\overline{x}_0).$$

Finally, as $x_0 = z_0$ and $f \in \mathrm{QG}_-^{\mu_0}$, we have

$$\varphi(\overline{x}_0) = f(\tilde{x}_0) - f^* + \frac{\sqrt{\mu_0 L_0}}{2}\left\|\tilde{x}_0 - x^*\right\|^2 \le \Big(1 + \sqrt{\frac{L_0}{\mu_0}}\Big)\left(f(\tilde{x}_0) - f^*\right),$$

where we used $\tilde{x}_0 = \tilde{z}_0$, and we obtain (5.52), namely

$$\mathbb{E}\left[e^{\beta T_k}(f(\tilde{x}_k) - f^*)\right] \le K_0,$$

with $\beta = \theta$ and $K_0 = \Big(1 + \sqrt{\frac{L_0}{\mu_0}}\Big)(f(\tilde{x}_0) - f^*)$. By Lemma 5.F.1, with probability $1 - \frac{1}{c_0} - \exp(-(c_1 - 1 - \log c_1)k)$,

$$f(\tilde{x}_k) - f^* \le c_0\Big(1 + \sqrt{\frac{L_0}{\mu_0}}\Big)(f(\tilde{x}_0) - f^*)\,\exp\Big(-\theta(1-c_1)k\Big),$$

which is the claim. □

# 5.E Trajectory-adapted convergence

In this section, we address a setting in which we assume the aiming condition is verified on average along the trajectory (Assumption 5.6.1). We start by proving Theorem 5.6.1, that addresses the dynamical system case. Then, we will state and prove similar results in the discrete case.

## 5.E.1 Proof of Theorem 5.6.1

**Proof with Gradient Flow** (GF) Let

$$E_t = e^{a_{avg}\sqrt{\mu\mu_0}\,t}\left[f(x_t) - f^* + \frac{\beta}{2}\left\|x_t - x^*\right\|^2\right].$$

Then

$$\begin{aligned}
\dot{E}_t &= a_{avg}\sqrt{\mu\mu_0}\,e^{a_{avg}\sqrt{\mu\mu_0}t}\left[f(x_t) - f^* + \frac{\beta}{2}\left\|x_t - x^*\right\|^2\right] \\
&\quad + e^{a_{avg}\sqrt{\mu\mu_0}t}\left[\langle \nabla f(x_t), \dot{x}_t\rangle + \beta\left\langle x_t - x^*, \dot{x}_t\right\rangle\right] \\
&\overset{(i)}{=} a_{avg}\sqrt{\mu\mu_0}\,e^{a_{avg}\sqrt{\mu\mu_0}t}\left[f(x_t) - f^* + \frac{\beta}{2}\left\|x_t - x^*\right\|^2\right] \\
&\quad + e^{a_{avg}\sqrt{\mu\mu_0}t}\left[-\left\|\nabla f(x_t)\right\|^2 - \beta\left\langle x_t - x^*, \nabla f(x_t)\right\rangle\right] \\
&\le a_{avg}\sqrt{\mu\mu_0}\,e^{a_{avg}\sqrt{\mu\mu_0}t}\left[f(x_t) - f^* + \frac{\beta}{2}\left\|x_t - x^*\right\|^2\right] - \beta e^{a_{avg}\sqrt{\mu\mu_0}t}\left\langle x_t - x^*, \nabla f(x_t)\right\rangle,
\end{aligned}$$

where (i) follows from $\dot{x}_t = -\nabla f(x_t)$.

Integrating from 0 to $t$ yields

$$E_t \le E_0 + \int_0^t a_{avg}\sqrt{\mu\mu_0}\,e^{a_{avg}\sqrt{\mu\mu_0}s}\left[f(x_s) - f^* + \frac{\beta}{2}\left\|x_s - x^*\right\|^2\right]ds - \int_0^t \beta e^{a_{avg}\sqrt{\mu\mu_0}s}\left\langle x_s - x^*, \nabla f(x_s)\right\rangle ds. \tag{5.33}$$

By Assumption 5.6.1 (with $\theta = a_{avg}\sqrt{\mu\mu_0}$), we have

$$\int_0^t e^{\theta s} \langle \nabla f(x_s), x_s - x^* \rangle \, ds \geq a_{avg} \int_0^t e^{\theta s} \|\nabla f(x_s)\| \, \|x_s - x^*\| \, ds,$$

hence

$$\int_0^t \beta e^{\theta s} \langle x_s - x^*, \nabla f(x_s) \rangle \, ds \geq \int_0^t \beta a_{avg} e^{\theta s} \|x_s - x^*\| \, \|\nabla f(x_s)\| \, ds.$$

Plugging into (5.33) gives

$$E_t \leq E_0 + \int_0^t a_{avg}\sqrt{\mu\mu_0}\, e^{\theta s} \left[ f(x_s) - f^* + \frac{\beta}{2} \|x_s - x^*\|^2 \right] ds - \int_0^t \beta a_{avg} e^{\theta s} \|x_s - x^*\| \, \|\nabla f(x_s)\| \, ds. \quad (5.34)$$

Using Lemma 5.C.1, for all $s \geq 0$,

$$- \|x_s - x^*\| \, \|\nabla f(x_s)\| \leq -\sqrt{\frac{\mu}{L_0}} (f(x_s) - f^*) - \frac{\sqrt{\mu\mu_0}}{2} \|x_s - x^*\|^2 .$$

Multiplying by $\beta a_{avg} e^{\theta s}$ and inserting into (5.34), we obtain

$$\begin{aligned} E_t &\leq E_0 + \int_0^t e^{\theta s} \Big[ a_{avg}\sqrt{\mu\mu_0} \, (f(x_s) - f^*) - \beta a_{avg} \sqrt{\frac{\mu}{L_0}} \, (f(x_s) - f^*) \\ &\qquad + \frac{\beta}{2} a_{avg}\sqrt{\mu\mu_0} \, \|x_s - x^*\|^2 - \frac{\beta}{2} a_{avg}\sqrt{\mu\mu_0} \, \|x_s - x^*\|^2 \Big] ds \\ &= E_0 + a_{avg} \int_0^t e^{\theta s} \Big( \sqrt{\mu\mu_0} - \beta \sqrt{\frac{\mu}{L_0}} \Big) \, (f(x_s) - f^*) \, ds. \end{aligned}$$

Choose $\beta = \sqrt{\mu_0 L_0}$, so that $\sqrt{\mu\mu_0} - \beta\sqrt{\mu/L_0} = 0$. Then $E_t \leq E_0$ for all $t \geq 0$. Since $f(x_t) - f^* \leq e^{-\theta t} E_t$, we get

$$f(x_t) - f^* \leq e^{-\theta t} E_0 = e^{-a_{avg}\sqrt{\mu\mu_0}\, t} \Big( f(x_0) - f^* + \frac{\beta}{2} \|x_0 - x^*\|^2 \Big).$$

Finally, using $f(x_0) - f^* \geq \frac{\mu_0}{2} \|x_0 - x^*\|^2$ and $\beta = \sqrt{\mu_0 L_0}$,

$$\frac{\beta}{2} \|x_0 - x^*\|^2 \leq \sqrt{\frac{L_0}{\mu_0}} \, (f(x_0) - f^*),$$

hence

$$f(x_t) - f^* \leq \Big( 1 + \sqrt{\frac{L_0}{\mu_0}} \Big) e^{-a_{avg}\sqrt{\mu\mu_0}\, t} \, (f(x_0) - f^*).$$

**Proof with Nesterov ODE momentum** (NMO) Let

$$E_t = e^{\theta t} \Big( f(x_t) - f^* \Big) + \frac{\beta}{2} e^{\theta t} \|z_t - x^*\|^2 .$$

As in the proof of Theorem 5.5.3 (i), choose $\eta' = \theta$ and $\eta = \beta\gamma'$. Then we get (see (5.19))

$$\dot{E}_t \leq e^{\theta t} \Big( \theta (f(x_t) - f^*) + \eta \langle \nabla f(x_t), x^* - x_t \rangle + \frac{\beta\theta}{2} \|x_t - x^*\|^2 \Big). \quad (5.35)$$

Integrating,

$$\begin{aligned} E_t &\leq E_0 + \int_0^t e^{\theta s} \Big( \theta (f(x_s) - f^*) + \frac{\beta\theta}{2} \|x_s - x^*\|^2 \Big) ds + \eta \int_0^t e^{\theta s} \langle \nabla f(x_s), x^* - x_s \rangle \, ds \\ &= E_0 + \int_0^t e^{\theta s} \Big( \theta (f(x_s) - f^*) + \frac{\beta\theta}{2} \|x_s - x^*\|^2 \Big) ds - \eta \int_0^t e^{\theta s} \langle \nabla f(x_s), x_s - x^* \rangle \, ds. \end{aligned} \quad (5.36)$$

By Assumption 5.6.1,

$$\int_0^t e^{\theta s} \langle \nabla f(x_s), x_s - x^* \rangle \, ds \geq a_{avg} \int_0^t e^{\theta s} \|\nabla f(x_s)\| \, \|x_s - x^*\| \, ds,$$

hence

$$E_t \leq E_0 + \int_0^t e^{\theta s} \Big( \theta (f(x_s) - f^*) + \frac{\beta\theta}{2} \|x_s - x^*\|^2 \Big) ds - \eta a_{avg} \int_0^t e^{\theta s} \|\nabla f(x_s)\| \, \|x_s - x^*\| \, ds. \tag{5.37}$$

Using Lemma 5.C.1,

$$- \|x_s - x^*\| \, \|\nabla f(x_s)\| \leq - \sqrt{\frac{\mu}{L_0}} (f(x_s) - f^*) - \frac{\sqrt{\mu\mu_0}}{2} \|x_s - x^*\|^2 ,$$

and inserting into (5.37), we obtain

$$E_t \leq E_0 + \int_0^t e^{\theta s} \left[ \Big( \theta - \eta a_{avg} \sqrt{\frac{\mu}{L_0}} \Big) (f(x_s) - f^*) + \frac{1}{2} \Big( \beta\theta - \eta a_{avg} \sqrt{\mu\mu_0} \Big) \|x_s - x^*\|^2 \right] ds. \tag{5.38}$$

Choose $\beta = \sqrt{\mu_0 L_0}$ and impose the cancellations

$$\theta = \eta a_{avg} \sqrt{\frac{\mu}{L_0}} \qquad \text{and} \qquad \beta\theta = \eta a_{avg} \sqrt{\mu\mu_0}.$$

These two equalities are equivalent to $\beta = \sqrt{\mu_0 L_0}$ and

$$\theta = a_{avg} \Big( \frac{\mu_0}{L_0} \Big)^{1/4} \sqrt{\mu}, \qquad \eta = (\mu_0 L_0)^{1/4}.$$

With $\eta' = \theta$ and $\eta = \beta\gamma'$, we can take

$$\gamma' = \frac{\eta}{\beta} = (\mu_0 L_0)^{-1/4}, \qquad \eta' = \theta = a_{avg} \Big( \frac{\mu_0}{L_0} \Big)^{1/4} \sqrt{\mu}, \qquad \gamma \geq 0.$$

Then the integrand in (5.38) is identically zero and we get $E_t \leq E_0$. Since $f(x_t) - f^* \leq e^{-\theta t} E_t$, we deduce

$$f(x_t) - f^* \leq e^{-\theta t} E_0 = e^{-\theta t} \Big( f(x_0) - f^* + \frac{\beta}{2} \|z_0 - x^*\|^2 \Big).$$

As $z_0 = x_0$, then

$$f(x_t) - f^* \leq e^{-\theta t} \Big( f(x_0) - f^* + \frac{\sqrt{\mu_0 L_0}}{2} \|x_0 - x^*\|^2 \Big).$$

Using $f(x_0) - f^* \geq \frac{\mu_0}{2} \|x_0 - x^*\|^2$, we obtain

$$f(x_t) - f^* \leq \Big( 1 + \sqrt{\frac{L_0}{\mu_0}} \Big) e^{-\theta t} \, (f(x_0) - f^*), \qquad \theta = a_{avg} \Big( \frac{\mu_0}{L_0} \Big)^{1/4} \sqrt{\mu}.$$

### 5.E.2 Discrete case

We address the discrete case. As in Appendix 5.D, we consider stochastic gradients under Assumption 5.D.1.

**Assumption 5.E.1.** *For some $\{A_i\}_{i=1,\dots,k}$ and $a_{avg} > 0$, $\{\tilde{y}_k\}_{k \geq 0}$ is such that*

$$\sum_{i=1}^{k} A_i \left[ \langle \nabla f(\tilde{y}_i, \xi_i), \tilde{y}_i - x^* \rangle - a_{avg} \|\nabla f(\tilde{y}_i, \xi_i)\| \, \|x^* - \tilde{y}_i\| \right] \geq 0.$$

We consider (NM), a similar and simpler analysis can be carried for (GD).

**Theorem 5.E.1.** *Let $f \in PL^\mu \cap LS^L$. Denote $L_0 := \inf\{L' \le L : f \in QG_+^{L'}\}$, $\mu_0 := \sup\{\mu' \ge \mu : f \in QG_-^{\mu'}\}$. Let $\{\tilde{x}_k\}_{k\in\mathbb{N}}$ verify* (NM) *with the continuized parameterization and with $\nabla f(\tilde{x}_k)$ replaced by $\nabla f(\tilde{x}_k, \xi_k)$ for random variables $\xi_k$ independent of $\tilde{x}_k$, under Assumption 5.D.1, and assume that Assumption 5.E.1 holds with $A_i = e^{\theta T_i}$, where*

$$\theta = \eta' = \frac{a_{avg}}{\rho}\Big(\frac{\mu_0}{L_0}\Big)^{1/4}\sqrt{\frac{\mu}{L}}.$$

*Let*

$$\gamma = \frac{1}{\rho L}, \quad \gamma' = \frac{1}{\rho(\mu_0 L_0)^{1/4}\sqrt{L}}, \quad \eta = \frac{1}{\rho}\frac{(\mu_0 L_0)^{1/4}}{\sqrt{L}}.$$

*Then, with probability $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k}$ (for some $c_0 > 1$, $c_1 \in (0,1)$),*

$$f(\tilde{x}_k) - f^* \le c_0\Big(1 + \sqrt{\frac{L_0}{\mu_0}}\Big)(f(\tilde{x}_0) - f^*)\,\exp\Big(-\theta(1-c_1)k\Big).$$

*Proof.* We use the same notations as in the proof of Theorem 5.D.2. Let $\overline{x}_t = (t, x_t, z_t)$ and define

$$\varphi(t,x,z) = A_t(f(x) - f^*) + \frac{B_t}{2}\|z - x^*\|^2, \qquad A_t = e^{\theta t}, \quad B_t = \sqrt{\mu_0 L_0}\, e^{\theta t}.$$

Applying Proposition 5.F.1, we have

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \Big[\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi\big(\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)\big)\Big]ds + M_t, \tag{5.39}$$

where $M_t$ is a martingale with $\mathbb{E}[M_t] = 0$.

From the same algebraic computations as in Theorem 5.5.3 (ii), we obtain

$$\begin{aligned}
&\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi\big(\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)\big) \\
&\le \frac{1}{2}\Big(\frac{dB_s}{ds} - B_s\eta'\Big)\|z_s - x^*\|^2 + \frac{B_s\eta'}{2}\|x_s - x^*\|^2 + \frac{dA_s}{ds}(f(x_s) - f^*) \\
&\quad + A_s\eta\,\langle \nabla f(x_s), x^* - x_s\rangle + (A_s\eta - B_s\gamma')\,\langle \nabla f(x_s), z_s - x^*\rangle \\
&\quad + \frac{1}{2}\Big(\rho B_s\gamma'^2 - \frac{A_s}{\rho L}\Big)\|\nabla f(x_s)\|^2.
\end{aligned} \tag{5.40}$$

With the parameter choices of the theorem and $B_s = \sqrt{\mu_0 L_0}A_s$, we have

$$A_s\eta - B_s\gamma' = 0, \qquad \rho B_s\gamma'^2 - \frac{A_s}{\rho L} = 0, \qquad \frac{dB_s}{ds} - B_s\eta' = 0.$$

Thus (5.40) reduces to

$$\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi\big(\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)\big) \le \frac{dA_s}{ds}(f(x_s) - f^*) + \frac{B_s\eta'}{2}\|x_s - x^*\|^2 + A_s\eta\,\langle \nabla f(x_s), x^* - x_s\rangle. \tag{5.41}$$

We now control the last term using the jump structure. For any predictable process $h_s$,

$$\int_0^t h_s\, dN_s = \int_0^t h_s\, ds + \int_0^t h_s\,(dN_s - ds),$$

where $\int_0^t h_s(dN_s - ds)$ is a martingale.

Applying this identity with

$$h_s = A_s\eta\Big[\langle \nabla f(x_s,\xi_s), x_s - x^*\rangle - a_{avg}\,\|\nabla f(x_s,\xi_s)\|\,\|x_s - x^*\|\Big],$$

and using Assumption 5.E.1 with $A_i = e^{\theta T_i}$, we obtain for $t = T_k$

$$\int_0^{T_k} A_s\eta\,\langle \nabla f(x_s), x^* - x_s\rangle\, dN_s \le -\eta a_{avg}\int_0^{T_k} A_s\,\|\nabla f(x_s)\|\,\|x_s - x^*\|\, dN_s. \tag{5.42}$$

This holds because by definition of the Poisson integrals, we have

$$\int_0^{T_k} A_s\eta \,\langle \nabla f(x_s), x^* - x_s\rangle \, dN_s = \sum_{i=1}^{k} A_k\eta \left\langle \nabla f(x_{T_k^-}), x^* - x_{T_k^-}\right\rangle = \sum_{i=1}^{k} A_k\eta \,\langle \nabla f(\tilde{y}_k), x^* - \tilde{y}_k\rangle \,,$$

where the last inequality follows from Proposition 5.F.2.

Using Lemma 5.C.1,

$$-\left\|\nabla f(x_s)\right\| \left\|x_s - x^*\right\| \leq -\sqrt{\frac{\mu}{L_0}}(f(x_s) - f^*) - \frac{\sqrt{\mu\mu_0}}{2}\left\|x_s - x^*\right\|^2,$$

and inserting this inequality into (5.42), then back into (5.41), we obtain

$$\langle \nabla\varphi(\overline{x}_s), b(\overline{x}_s)\rangle + \mathbb{E}_\xi\big(\varphi(\overline{x}_s + G(\overline{x}_s,\xi)) - \varphi(\overline{x}_s)\big) \leq 0.$$

Plugging this bound into (5.39) yields

$$\varphi(\overline{x}_t) \leq \varphi(\overline{x}_0) + M_t,$$

where $M_t$ is a martingale. Evaluating at $t = T_k$, taking expectations, and using Theorem 5.F.1, we get

$$\mathbb{E}\left[e^{\theta T_k}(f(\tilde{x}_k) - f^*)\right] \leq \varphi(\overline{x}_0).$$

Finally,

$$\varphi(\overline{x}_0) = f(\tilde{x}_0) - f^* + \frac{\sqrt{\mu_0 L_0}}{2}\left\|\tilde{z}_0 - x^*\right\|^2 \leq \Big(1 + \sqrt{\frac{L_0}{\mu_0}}\Big)(f(\tilde{x}_0) - f^*),$$

and Lemma 5.F.1 concludes the proof. □

### 5.E.3 Possible Extensions: Adapt to other Conditions and Exact Rates

We could consider other extensions of the proposed viewpoint. First, we could consider other averaged condition, such as an *averaged PL* condition:

$$\int_0^t e^{\theta t}\left\|\nabla f(x_s)\right\|^2 ds \geq 2\mu_{avg}\int_0^t (f(x_s) - f^*)ds,$$

for some $\mu_{avg} > 0$. This inequality is automatically verified if $f \in \mathrm{PL}^\mu$, with $\mu_{avg} \geq \mu$. Interestingly, $(x_t)_{t\geq 0}$ could cross a saddle point $\hat{x}$, such that $f \notin \mathrm{PL}^\mu$, while having that this inequality is verified. In particular, (NMO) could cross and escape such saddle point, and then converging to a global minima. This saddle point could be "averaged out".

Alternatively, we also could consider point-wise parameter, as illustrated next.

**Proposition 5.E.1.** *Define the point-wise Polyak-Lojasiewicz constant* $\mu(x) := 2\frac{\|\nabla f(x)\|^2}{f(x)-f^*}$. *If* $(x_t)_{t\geq 0}$ *follows* (GF), *we have*

$$f(x_t) - f^* = e^{-2\int_0^t \mu(x_s)ds}(f(x_0) - f^*).$$

We obtain an exact rate for the decrease of $f(x_t) - f^*$. If $f \in \mathrm{PL}^\mu$, we have in particular $2\int_0^t \mu(x_s)ds \geq 2\mu t$. We can carry similar analysis with (NMO), under a point-wise aiming condition.

**Proposition 5.E.2.** *Define the point-wise: Polyak-Lojasiewicz constant* $\mu(x) := 2\frac{\|\nabla f(x)\|^2}{f(x)-f^*}$, *aiming condition constant* $a(x) := \frac{\langle \nabla f(x), x - x^*\rangle}{\|\nabla f(x)\|\|x-x^*\|}$ *and quadratic growth constant* $\mu_0(x) = 0.5\cdot\frac{f(x)-f^*}{\|x-x^*\|^2}$. *If* $(x_t)_{t\geq 0}$ *follows* (NMO) *with* $\eta_t' = a(x_t)\sqrt{\mu(x_t)}$, $\quad \eta_t = \frac{L_0}{\sqrt{\mu_0(x_t)}}$, $\quad \gamma_t' = \frac{1}{\sqrt{\mu_0(x_t)}}$, $\quad \gamma_t \geq 0$, *we have*

$$f(x_t) - f^* \leq e^{-\int_0^t a(x_s)\sqrt{\mu(x_s)}ds} L_0\left\|x_0 - x^*\right\|^2,$$

*with* $L_0 = \sup_{x\in\mathbb{R}^d}\mu_0(x)$.

This rate is however purely conceptual, as it involves exact values of parameters we cannot compute. The averaged conditions such as Assumption 5.6.1 are more practical, as it is sufficient to use a small enough value of $a_{avg}$, and not an exact one.

**Proof of Proposition 5.E.1**

**(i) Case of gradient flow** (GF). Set

$$E_t := e^{2\int_0^t \mu(x_s)ds}(f(x_t) - f^*).$$

We have

$$\begin{aligned}\dot{E}_t &= 2\mu(x_t)e^{2\int_0^t \mu(x_s)ds}(f(x_t) - f^*) + e^{2\int_0^t \mu(x_s)ds}\langle \nabla f(x_t), \dot{x}_t\rangle \\ &\overset{\text{(GF)}}{=} 2\mu(x_t)e^{2\int_0^t \mu(x_s)ds}(f(x_t) - f^*) - e^{2\int_0^t \mu(x_s)ds}\|\nabla f(x_t)\|^2\end{aligned} \tag{5.43}$$

Recall $\mu(x) := 2\frac{\|\nabla f(x)\|^2}{f(x)-f^*}$. Plugging it in (5.43), we deduce

$$\dot{E}_t = 2\mu(x_t)e^{2\int_0^t \mu(x_s)ds}(f(x_t) - f^*) - 2\mu(x_t)e^{2\int_0^t \mu(x_s)ds}(f(x_t) - f^*) = 0. \tag{5.44}$$

By integrating we deduce

$$f(x_t) - f^* = e^{-2\int_0^t \mu(x_s)ds}(f(x_0) - f^*). \tag{5.45}$$

**Proof of Proposition 5.E.2**

Let

$$E_t = e^{\int_0^t a(x_s)\sqrt{\mu(x_s)}ds}\left(f(x_t) - f^* + \frac{\beta}{2}\|z_t - x^*\|^2\right). \tag{5.46}$$

Denote

$$\Delta_t := e^{\int_0^t a(x_s)\sqrt{\mu(x_s)}ds}, \qquad u_t := a(x_t)\sqrt{\mu(x_t)}.$$

We have

$$\begin{aligned}\Delta_t^{-1}\dot{E}_t &= u_t(f(x_t) - f^*) + \frac{u_t\beta}{2}\|z_t - x^*\|^2 + \langle \nabla f(x_t), \dot{x}_t\rangle + \beta\langle z_t - x^*, \dot{z}_t\rangle \\ &= u_t(f(x_t) - f^*) + \frac{u_t\beta}{2}\|z_t - x^*\|^2 \\ &\quad + \langle \nabla f(x_t), \eta_t(z_t - x_t) - \gamma_t\nabla f(x_t)\rangle \\ &\quad + \beta\langle z_t - x^*, \eta_t'(x_t - z_t) - \gamma_t'\nabla f(x_t)\rangle \\ &= u_t(f(x_t) - f^*) + \frac{u_t\beta}{2}\|z_t - x^*\|^2 \\ &\quad + \eta_t\langle \nabla f(x_t), x^* - x_t\rangle + (\eta_t - \beta\gamma_t')\langle \nabla f(x_t), z_t - x^*\rangle \\ &\quad + \beta\eta_t'\langle z_t - x^*, x_t - z_t\rangle - \gamma_t\|\nabla f(x_t)\|^2.\end{aligned} \tag{5.47}$$

We choose $\eta_t' = u_t$ to cancel the $\|z_t - x^*\|^2$ term, and $\eta_t = \beta\gamma_t'$ to cancel the mixed scalar product. Then

$$\Delta_t^{-1}\dot{E}_t = u_t(f(x_t) - f^*) + \beta\gamma_t'\langle \nabla f(x_t), x^* - x_t\rangle + \frac{\beta u_t}{2}\|x_t - x^*\|^2 - \gamma_t\|\nabla f(x_t)\|^2. \tag{5.48}$$

By Lemma 5.C.2,

$$\langle \nabla f(x_t), x^* - x_t\rangle = -a(x_t)\sqrt{\frac{\mu(x_t)}{\mu_0(x_t)}}(f(x_t) - f^*) - a(x_t)\sqrt{\mu(x_t)\mu_0(x_t)}\frac{1}{2}\|x_t - x^*\|^2.$$

Fixing $\gamma_t' = \frac{1}{\sqrt{\mu_0(x_t)}}$ and ignoring the nonpositive term $-\gamma_t\|\nabla f(x_t)\|^2$, we obtain

$$\Delta_t^{-1}\dot{E}_t = u_t\Big(1 - \frac{\beta}{\mu_0(x_t)}\Big)(f(x_t) - f^*). \tag{5.49}$$

Choosing $\beta = L_0 = \sup_{x\in\mathbb{R}^d}\mu_0(x)$ ensures $\dot{E}_t \le 0$ for all $t \ge 0$. Therefore $E_t \le E_0$, and

$$f(x_t) - f^* \le e^{-\int_0^t a(x_s)\sqrt{\mu(x_s)}ds}E_0.$$

Finally, since $\mu_0(x_0) \leq L_0$, we have

$$f(x_0) - f^* = 2\mu_0(x_0) \|x_0 - x^*\|^2 \leq 2L_0 \|x_0 - x^*\|^2 ,$$

which yields

$$f(x_t) - f^* \leq e^{-\int_0^t a(x_s)\sqrt{\mu(x_s)}ds} L_0 \|x_0 - x^*\|^2 .$$

## 5.F Continuized Toolbox

The continuized method has been discussed in Chapter 4. We restate here the tools used in the present chapter. We recall that the fundamental object are the continuized Nesterov equations:

$$\begin{cases} dx_t &= \eta(z_t - x_t)dt - \gamma\nabla f(x_t)dN_t \\ dz_t &= \eta'(x_t - z_t)dt - \gamma'\nabla f(x_t)dN_t \end{cases} \tag{CNE}$$

where $dN(t) = \sum_{k\geq 0} \delta_{T_k}(dt)$ is a Poisson point measure with intensity $dt$, with $T_1, T_2, \dots$ random times such that for all $i$, $T_{i+1} - T_i \overset{i.i.d}{\sim} \mathcal{E}(1)$. We can analyze it using Lyapunov-approaches, through Itô calculus.

**Proposition 5.F.1** ((Even et al., 2021), Proposition 2)**.** *Let $x_t \in \mathbb{R}^d$ be a solution of*

$$dx_t = b(x_t)dt + \int_\Xi G(x_t, \xi)dN(t, \xi)$$

*and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Then,*

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla\varphi(x_s), b(x_s)\rangle \, ds + \int_{[0,t]} \mathbb{E}_\xi \varphi(x_s + G(x_s, \xi)) - \varphi(x_s)ds + M_t, \tag{5.50}$$

*where $M_t$ is a martingale such that $\mathbb{E}[M_t] = 0$, $\forall t \geq 0$.*

Importantly, the following evaluations $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$ of (CNE) can be written as an algorithmic form.

**Proposition 5.F.2** (Chapter 4, Proposition 4.3.2)**.** *Let $(x_t, z_t)_{t\geq 0}$ follow* (CNE)*. Define $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$ as evaluations of this process. Then, $(\tilde{y}_k, \tilde{x}_k, \tilde{z}_k)$ writes as a Nesterov algorithm in the form of* (NM) *with stochastic parameters*

$$\begin{cases} \tilde{y}_k &= (\tilde{z}_k - \tilde{x}_k)\frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')(T_{k+1}-T_k)}\right) + \tilde{x}_k \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma\nabla f(\tilde{y}_k, \xi_k) \\ \tilde{z}_{k+1} &= \tilde{z}_k + \eta'\frac{\left(1-e^{-(\eta+\eta')(T_{k+1}-T_k)}\right)}{\eta'+\eta e^{-(\eta+\eta')(T_{k+1}-T_k)}}(\tilde{y}_k - \tilde{z}_k) - \gamma'\nabla f(\tilde{y}_k, \xi_k) \end{cases} \tag{5.51}$$

In order to transfer a result involving $(x_t, z_t)_{t\geq 0}$ to $(\tilde{x}_k, \tilde{z}_k)_{k\in\mathbb{N}}$, we use need a stopping theorem, such as the following.

**Theorem 5.F.1** (Stopping theorem, Chapter 4, Proposition 4.3.1)**.** *Let $(\varphi_t)_{t\in\mathbb{R}_+}$ be a nonnegative process with cadlag trajectories, such that it verifies*

$$\varphi_t \leq K_0 + M_t,$$

*for some positive random variable $K_0$, some martingale $(M_t)_{t\geq 0}$ with $M_0 = 0$. Then, for an almost surely finite stopping time $\tau$, one has*

$$\mathbb{E}[\varphi_\tau] \leq \mathbb{E}[K_0] .$$

Finally, because the continuized analysis leads to result in expectation, in the case where we obtain linear convergence rates, concentration inequalities such as Chernov inequality allows to deduce a result on the trajectory, with high probability.

**Lemma 5.F.1** (Chapter 4, Proposition 4.4.1)**.** *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d of law $\mathcal{E}(1)$, with convention $T_0 = 0$. If the iterations of (5.51) verify*

$$\mathbb{E}\left[e^{\beta T_k}(f(\tilde{x}_k) - f^*)\right] \leq K_0, \tag{5.52}$$

*then, with probability $1 - \frac{1}{c_0} - e^{-(c_1 - 1 - \log(c_1))k}$, for some $c_0 > 1$ and $c_1 \in (0,1)$ we have*

$$f(\tilde{x}_k) - f^* \leq c_0 K_0 e^{-\beta(1-c_1)k}.$$

# 5.G Proof of Proposition 5.4.1

**Computation of $\nabla h(x)$.** We have

$$\begin{aligned}\nabla h(x) &= \frac{x}{\|x\|} f'(\|x\|) g\left(\frac{x}{\|x\|}\right) + \left(\frac{1}{\|x\|} I_d - \frac{x x^T}{\|x\|^3}\right) f(\|x\|) \nabla g\left(\frac{x}{\|x\|}\right) \\ &= \frac{x}{\|x\|} f'(\|x\|) g\left(\frac{x}{\|x\|}\right) + \frac{1}{\|x\|}\left(I_d - \frac{x x^T}{\|x\|^2}\right) f(\|x\|) \nabla g\left(\frac{x}{\|x\|}\right)\end{aligned} \tag{5.53}$$

Noting $r := \|x\|$ and $u := x/\|x\|$, this can be rewritten

$$\nabla h(x) = \nabla h(ru) = f'(r) g(u) u + \frac{1}{r} f(r) \nabla_{\mathbb{S}} g(u), \tag{5.54}$$

where $\nabla_{\mathbb{S}} g(u) = (I_d - uu^T)\nabla g(u)$ is the Riemannian gradient of $g$ on the sphere. $\nabla_{\mathbb{S}} g(u)$ belongs to the tangent space at $u$, namely $\nabla_{\mathbb{S}} g(u) \in \mathcal{T}_u = \{v, \langle u, v\rangle = 0\}$. In particular, it implies that $u$ and $\nabla_{\mathbb{S}} g(u)$ are orthogonal, so

$$\|\nabla h(x)\|^2 = f'(r)^2 g(u)^2 + \frac{1}{r^2} f(r)^2 \|\nabla_{\mathbb{S}} g(u)\|^2,$$

and

$$\langle u, \nabla_{\mathbb{S}} g(u)\rangle = 0.$$

So, we have

$$\langle \nabla h(x), x - x^*\rangle = \left\langle f'(r) g(u) u + \frac{1}{r} f(r) \nabla_{\mathbb{S}} g(u), ru \right\rangle = r f'(r) g(u),$$

and

$$\|\nabla h(x)\| \|x - x^*\| = \sqrt{f'(r)^2 g(u)^2 + \frac{1}{r^2} f(r)^2 \|\nabla_{\mathbb{S}} g(u)\|^2}\, r,$$

such that

$$\frac{\langle \nabla h(x), x - x^*\rangle}{\|\nabla h(x)\| \|x - x^*\|} = \frac{f'(r) g(u)}{\sqrt{f'(r)^2 g(u)^2 + \frac{1}{r^2} f(r)^2 \|\nabla_{\mathbb{S}} g(u)\|^2}}.$$

So, $h \in \mathrm{AC}^a$ for $a$ such that

$$a := \inf_{r>0, u \in \mathbb{S}^{d-1}} \frac{f'(r) g(u)}{\sqrt{f'(r)^2 g(u)^2 + \frac{1}{r^2} f(r)^2 \|\nabla_{\mathbb{S}} g(u)\|^2}} = \inf_{r>0, u \in \mathbb{S}^{d-1}} \frac{1}{\sqrt{1 + \left(\frac{f(r)}{r f'(r)}\right)^2 \left(\frac{\|\nabla_{\mathbb{S}} g(u)\|}{g(u)}\right)^2}}$$

## 5.G.1 Proof of Proposition 5.4.2

For the converse relation, we define a weaker relaxation of strong convexity than $\mathrm{SQC}^{\mu}_{\tau}$, defined by the Restricted Secant Inequality (RSI) (Zhang and Yin, 2013).

**Definition 5.G.1.** *For $\nu > 0$, we call RSI$^{\nu}$ the set of functions $f$ that satisfies the Restricted Secant Inequality, namely $\forall x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x^*\rangle \geq \nu \|x - x^*\|^2$.*

A function in $\mathrm{SQC}_\tau^\mu$ is also in $\mathrm{RSI}^{\frac{2\mu}{2-\gamma}}$, such that the RSI condition is weaker, see Chapter 3.

**Proposition 5.G.1.** *Let $f \in RSI^\nu$. If $f$ also satisfies*

$$\forall x \in \mathbb{R}^d,\ \|\nabla f(x)\| \leq L \|x - x^*\|, \tag{5.55}$$

*then $f \in AC^{\frac{\nu}{L}}$.*

*Proof.* We have

$$\langle \nabla f(x), x - x^* \rangle \geq \nu \|x - x^*\|^2 \geq \frac{\nu}{L} \|x - x^*\| \|\nabla f(x)\|.$$

□

The condition (5.55) is sometimes named $\mathrm{EB}^+$, as the upper version of the Error-Bound condition (Guille-Escuret et al., 2021). It can also be interpreted as being Lipschitz with respect to the minimizers. The result follows from the previous proposition.

# Chapter 6

# Application of the Continuized Nesterov Momentum Method to Functions with Lipschitz Gradient and Hessian: achieving the $\mathcal{O}(\varepsilon^{-7/4})$ Complexity without Momentum-Control Mechanisms

We recall that the content of this chapter is based on the following paper.

> J. Hermant, J-F. Aujol, C. Dossal, L. Huang and A. Rondepierre, I. Waldspurger.
> **Continuized Nesterov Momentum Achieves the $\mathcal{O}(\varepsilon^{7/4})$ Complexity in Smooth Nonconvex Optimization**.
> *arXiv preprint*, 2026

## 6.1 Introduction

We recall that we consider the following Nesterov Momentum algorithm (Nesterov, 2013).

$$\begin{cases} \tilde{y}_k &= \alpha_k \tilde{x}_k + (1-\alpha_k)\tilde{z}_k \\ \tilde{x}_{k+1} &= \tilde{y}_k - s\nabla f(\tilde{y}_k) \\ \tilde{z}_{k+1} &= \beta_k \tilde{z}_k + (1-\beta_k)\tilde{y}_k - \eta_k \nabla f(\tilde{y}_k) \end{cases} \tag{NM}$$

Following our discussion in Section 1.1.1, we recall that in the case of functions with Lipschitz gradient and Hessian, existing first-order algorithms achieving the accelerated $\mathcal{O}(\varepsilon^{-7/4})$ complexity rely on an algorithm of the form

**Algorithm 15** Structure of existing algorithms achieving the $\mathcal{O}(\varepsilon^{-7/4})$ complexity

1: **if** Security check succeed **then**
2: Run update rule (NM')
3: **else**
4: Erase accumulated momentum (+ potentially trigger alternative mechanisms)
5: **end if**

where

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \alpha_k(\tilde{x}_k - \tilde{x}_{k-1}), \\ \tilde{x}_{k+1} = \tilde{x}_k - \gamma \nabla f(\tilde{y}_k). \end{cases} \tag{NM'}$$

These algorithms are enforcing control on momentum. We then raised the following question.

**Are momentum-control mechanisms fundamentally necessary to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity?**

In a recent result, (Okamura et al., 2024) shows that the time needed for the solution of the vanilla Heavy Ball ordinary differential equation (ODE) to find an $\varepsilon$-stationary solution is at most $\mathcal{O}(\varepsilon^{-7/4})$. Importantly, the Heavy Ball ODE can be viewed as a continuous-time analogue of momentum algorithms (Su et al., 2016; Siegel, 2019). It means that this continuous-time momentum dynamic does not require such mechanisms, and naturally raises the question of whether a similar result can be established in the discrete setting. To the best of our knowledge, no such discrete counterpart is currently known. Following our point of view of Chapter 4, it is natural to wonder whether the continuized approach can help transfer this result to the discrete setting.

**Contribution** We draw inspiration from the ODE-based analysis of Okamura et al. (2024) to analyze the solution of (CNE), which yields our main result, stated as follows (see Theorem 6.4.1 for the formal version).

**Theorem 6.1.1** (Informal version)**.** *Let $f$ be with Lipschitz gradient and Hessian. For any $\tilde{\varepsilon} \in (0, 1/2)$ and for a number of iterations $n$ large enough,* (NM) *with a suitable stochastic parameterization outputs a point $\tilde{x}$*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \|\nabla f(\tilde{x})\|\right] = \mathcal{O}(n^{-4/7}),$$

*where $\mathcal{A}_n$ is a subset of the realizations that satisfies $\mathbb{P}(\mathcal{A}_n) \geq 1 - \tilde{\varepsilon}$.*

It further implies that a stochastic parameterization of (NM) is able to achieve a point that satisfies $\mathbb{E}\left[\|\mathbb{1}_{\mathcal{A}_n} \nabla f(\tilde{x})\|\right] \leq \varepsilon$ in at most $\mathcal{O}(\varepsilon^{-7/4})$ gradient evaluations, without any momentum-reset mechanisms. We note that for $n$ large enough, the hidden numerical factor of our bound improves over the best existing bound from Li and Lin (2023). Conceptually, our result indicates that (NM) by itself can achieve the complexity $\mathcal{O}(\varepsilon^{-7/4})$, at the cost of randomizing parameters, and thus, of a result in expectation. We also show that under Lipschitz gradient only, our algorithm recovers in expectation the best existing complexity $\mathcal{O}(\varepsilon^{-2})$.

### 6.1.1 Related works

**First-order acceleration under Lipschitz gradient and Hessian** The $\mathcal{O}(\varepsilon^{-7/4})$ complexity with a first-order algorithm was first achieved by (Carmon et al., 2017). The proposed nested-loop algorithm combines (NM') with negative-curvature exploitation and the minimization of a regularized surrogate function. This approach is subsequently simplified in (Jin et al., 2018), where the algorithm reduces to (NM') combined with negative-curvature steps. Replacing the negative curvature exploitation with a restart mechanism, (Li and Lin, 2023) also achieves the $\mathcal{O}(\varepsilon^{-7/4})$ bound, additionally improving by a logarithmic factor. This restart-based line of work has since been extended to a parameter-free algorithm (Marumo and Takeda, 2022), a forward-backward version (Renaud et al., 2026) and a universal algorithm under Hölder continuous Hessian (Marumo and Takeda, 2025). A different line of work, based on online-learning techniques rather than momentum, was initiated by Cutkosky et al. (2023), who achieved the complexity bound $\tilde{\mathcal{O}}(\varepsilon^{-7/4})$ under the supplementary assumption that $f$ is Lipschitz. Their idea was later extended by Jiang et al. (2025), improving the dependence on $\varepsilon$ to $\mathcal{O}(d^{1/4}\varepsilon^{-13/8})$ at the cost of an explicit dimension dependence, without this supplementary Lipschitzness assumption on $f$. These online-learning based methods are somewhat intricate, relying on nested algorithms and/or nested loops, but they show that in some settings, alternatives to momentum can achieve this accelerated bound. The first known lower-bound on the number of gradient evaluations needed to find an $\varepsilon$-stationary point in this setting was $\mathcal{O}(\varepsilon^{-12/7})$ Carmon et al. (2021b), and was recently improved to $\mathcal{O}(\varepsilon^{-7/4})$ Zhou (2026), closing the gap with the upper-bound.

**Second-order stationary points** In contrast to our setting, several works have focused on second-order stationary points, which under suitable assumptions correspond to local minimizers, see a formal definition in Appendix 6.A.1. With probability $1 - \delta$, such a point is found in at most $\mathcal{O}(\varepsilon^{-7/4} \log(\frac{d}{\delta\varepsilon}))$ gradient and Hessian-vector product evaluations (Agarwal et al., 2017; Carmon et al., 2018), or with gradient evaluation only (Jin et al., 2018; Li and Lin, 2023). Procedures based on first-order information have been designed that transform algorithms for finding first-order stationary points into ones that find second-order stationary

points Xu et al. (2018); Allen-Zhu and Li (2018). Even though designing methods that achieve second-order stationary points efficiently is an important avenue of research, we do not consider this direction in this work.

## 6.2 Background

In this work, we fix an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We use the notation $\Delta_f := f(x_0) - f^*$.

We recall our assumptions.

**Assumption 6.2.1.** *$f$ is such that* $\forall x, y \in \mathbb{R}^d,\ \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$.

**Assumption 6.2.2.** *$f$ is such that* $\forall x, y \in \mathbb{R}^d,\ \left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{spec} \leq L_2 \|x - y\|$.

For functions satisfying Assumption 6.2.1 and/or Assumption 6.2.2, a critical point could be a non-global minimizer, a saddle point, or a maximizer. We recall that Assumption 6.2.1 alone is sufficient to find an $\varepsilon$-stationary point with first-order methods (Nesterov, 2013; Ghadimi and Lan, 2016), but the complexity of gradient descent cannot be improved (Carmon et al., 2020). Considering Assumptions 6.2.1-6.2.2 together allows for such improvement. In existing analysis (Jin et al., 2018; Li and Lin, 2023), Assumption 6.2.2 enables working with local quadratic approximations, on which momentum is known to be effective (O'Neill and Wright, 2019). Our analysis does not rely on such approximations.

### 6.2.1 Continuized Nesterov Algorithm

The process considered in this chapter has been extensively described in Chapter 4. As a reminder or for readers only interested in this chapter, we provide a brief description of the process and of its main properties. An intuitive way to present the process is as follows: consider a sequence of random times $\{T_k\}_{k\in\mathbb{N}^*}$ such that $T_0 = 0$ and $T_{k+1} - T_k \overset{i.i.d}{\sim} \mathcal{E}(1)$ for $k \in \mathbb{N}$. For some constant parameters $\eta, \eta', \gamma, \gamma'$, we define a continuous-time process $(x_t, z_t)_{t\geq 0}$ on each interval $(T_k, T_{k+1})$ as the solution of the following equation

$$\left\{ \begin{array}{ll} dx_t &= \eta(z_t - x_t)dt, \\ dz_t &= \eta'(x_t - z_t)dt. \end{array} \right.$$

At $t = T_k$, the process jumps by performing gradient steps,

$$x_{T_k} = x_{T_k^-} - \gamma \nabla f(x_{T_k^-}), \quad z_{T_k} = z_{T_k^-} - \gamma' \nabla f(x_{T_k^-}).$$

This defines the *continuized Nesterov equation*, where we assume $x_0 = z_0$. It can be written in a more compact form as the following stochastic differential equation

$$\left\{ \begin{array}{ll} dx_t &= \eta(z_t - x_t)dt - \gamma \nabla f(x_{t^-})dN_t, \\ dz_t &= \eta'(x_t - z_t)dt - \gamma' \nabla f(x_{t^-})dN_t, \end{array} \right. \tag{CNE}$$

where $dN_t = \sum_{k\geq 0} \delta_{T_k}(dt)$ is a Poisson point measure with intensity $dt$. It mixes the continuous component, the $dt$ factor, with the gradient component that acts at discrete random times, the $dN_t$ factor.

A remarkable property of (CNE) is that, upon defining the sequences $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_k := x_{T_k}$ and $\tilde{z}_k := z_{T_k}$, $k \in \mathbb{N}$, it satisfies a recursive relation that takes the form of a Nesterov momentum algorithm, with stochastic parameters depending on the random times $\{T_k\}_{k\in\mathbb{N}^*}$.

**Proposition 6.2.1** (Chapter 4, Proposition 4.3.2)**.** *Let $(x_t, z_t)_{t\geq 0}$ follow* (CNE) *with $\eta + \eta' > 0$ and with underlying jumping times $\{T_k\}_{k\in\mathbb{N}^*}$. Define $\tilde{y}_k := x_{T_{k+1}^-}$, $\tilde{x}_{k+1} := x_{T_{k+1}}$ and $\tilde{z}_{k+1} := z_{T_{k+1}}$ as evaluations of this process. Then, $(\tilde{y}_k, \tilde{x}_k, \tilde{z}_k)$ writes as a specific parameterization of* (NM)*, that we call the continuized Nesterov algorithm* (CNA)*, namely*

$$\left\{ \begin{array}{ll} \tilde{y}_k &= \tilde{x}_k + \frac{\eta}{\eta+\eta'}(1 - e_k)(\tilde{z}_k - \tilde{x}_k) \\ \tilde{x}_{k+1} &= \tilde{y}_k - \gamma \nabla f(\tilde{y}_k) \\ \tilde{z}_{k+1} &= \tilde{z}_k + \eta' \frac{1-e_k}{\eta'+\eta e_k}(\tilde{y}_k - \tilde{z}_k) - \gamma' \nabla f(\tilde{y}_k) \end{array} \right. \tag{CNA}$$

*where $e_k = e^{-(\eta+\eta')(T_{k+1} - T_k)}$.*

In words, (CNA) is an exact evaluation of a realization of (CNE). Intuitively, exact access to the continuous process (CNE) is possible because gradient information is only queried at discrete times. This direct correspondence between (CNA) and (CNE) is powerful: it enables convergence guarantees for a discrete algorithm to be derived via continuous-time analysis. This comes at the cost that even with deterministic gradients, the resulting algorithm is inherently stochastic, thus leading to non-deterministic convergence guarantees.

**Practical implementation of** (CNA) From a practical point of view, compared to a deterministic parameterization of (NM), the randomization procedure in (CNA) does not introduce any significant additional difficulty. This is because, for any $k \in \mathbb{N}$, the increments $T_{k+1} - T_k$ are independent and identically distributed following an exponential law of parameter 1, and can therefore be easily and independently simulated. As a consequence, to run (CNA) with $n \in \mathbb{N}^*$ iterations, we can first generate at once $n$ random variable $E_k \overset{i.i.d}{\sim} \mathcal{E}(1)$, and save the vector $\{E_k\}_{k\in\{1,\cdots,n\}}$. Then, for fixed parameters $\eta, \eta', \gamma$, and $\gamma'$, the algorithm (CNA) can be implemented in the same way as (NM), using the sequences

$$\alpha_k = \frac{\eta}{\eta+\eta'}\left(1 - e^{-(\eta+\eta')E_k}\right), \qquad \beta_k = \eta' \frac{1 - e^{-(\eta+\eta')E_k}}{\eta' + \eta e^{-(\eta+\eta')E_k}}, \quad k = 1, \dots, n.$$

**Link between** (CNA) **and** (NM') Former results that achieved the $\mathcal{O}(\varepsilon^{-7/4})$ complexity used a two-sequence formulation (NM') as their momentum component. As a reminder, we proved in Chapter 3, Proposition 3.B.1 that we can rewrite (CNA) as a two-sequence algorithm:

$$\begin{cases} \tilde{y}_k = \tilde{x}_k + \dfrac{(1-e_k)e_{k-1}}{1-e_{k-1}}(\tilde{x}_k - \tilde{x}_{k-1}) + \dfrac{\eta}{\eta+\eta'}(1-e_k)\left(\dfrac{\gamma'}{\gamma} - 1 - \dfrac{(\eta+\eta')e_{k-1}}{\eta(1-e_{k-1})}\right)(\tilde{x}_k - \tilde{y}_{k-1}), \\ \tilde{x}_{k+1} = \tilde{y}_k - \gamma\nabla f(\tilde{y}_k), \end{cases}$$

with $e_k = e^{-(\eta+\eta')(T_{k+1}-T_k)}$. Compared with (NM'), there is a supplementary term, factor of $\tilde{x}_k - \tilde{y}_{k-1} = -\gamma\nabla f(\tilde{y}_{k-1})$. This supplementary term appears in algorithms that get an improved constant in the convergence bound in the smooth convex setting, such as the Optimized Gradient Method (OGM) (Kim and Fessler, 2016). With our choice of parameter in Theorem 6.4.1, this supplementary factor does not reduce to zero. It may indicate that algorithms of the form (NM') are unable to achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity by their own, even if randomizing the parameter.

## 6.3 Convergence Result without Lipschitz Hessian Assumption

We derive a convergence result under Assumption 6.2.1 only. This intermediate result serves as a key building block for proving our main theorem in the Lipschitz Hessian setting (Theorem 6.4.1).

**Lemma 6.3.1.** *If $(x_t, z_t)_{t\geq 0}$ is solution of (CNE) under Assumption 6.2.1 with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \gamma' - \gamma$ and $\eta' \geq -\eta$, we have*

$$\mathbb{E}\left[\int_0^t (\eta+\eta')\left\|x_s - z_s\right\|^2 + \frac{\gamma}{4}\left\|\nabla f(x_s)\right\|^2 ds\right] \leq \mathbb{E}\left[f(x_0) - f^*\right]. \tag{6.1}$$

*Sketch of Proof.* We set $\overline{x}_t = (x_t, z_t)$ as a concatenation of $(x_t)_{t\geq 0}$ and $(z_t)_{t\geq 0}$, which satisfies $d\overline{x}_t = \zeta(\overline{x}_t)dt + G(\overline{x}_t)dN(t)$, where

$$\zeta(\overline{x}_t) = \begin{pmatrix} \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\overline{x}_t) = \begin{pmatrix} -\gamma_t\nabla f(x_t) \\ -\gamma_t'\nabla f(x_t) \end{pmatrix}.$$

We set the Lyapunov function

$$\varphi(x, z) = f(x) + \frac{1}{2}\left\|x - z\right\|^2.$$

Intuitively, one can stochastically derivate $\varphi(\overline{x}_t)$ using an Itô formula, yielding the following relation

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \langle\nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle\, ds + \int_0^t \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) ds + M_t,$$

where $(M_t)_{t\geq 0}$ is a martingale verifying $\mathbb{E}[M_t] = 0$ for all $t \geq 0$. Using Assumption 6.2.1 and the parameter choice, we compute that

$$\langle \nabla\varphi(\overline{x}_s), \xi(\overline{x}_s)\rangle + \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) \leq -(\eta+\eta')\|x_s - z_s\|^2 - \frac{\gamma}{4}\|\nabla f(x_s)\|^2,$$

which yields

$$\int_0^t (\eta+\eta')\|x_s - z_s\|^2 + \frac{\gamma}{4}\|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t. \tag{6.2}$$

One can conclude using $\mathbb{E}[M_t] = 0$ for all $t \geq 0$. □

See the complete proof in Appendix 6.B. From Lemma 6.3.1, it follows that $\mathbb{E}\left[\min_{s\in[0,t]}\|\nabla f(x_s)\|^2\right] = \mathcal{O}(1/t)$. Perhaps surprisingly, the parameter $\eta'$ can be chosen freely in $(-\eta, +\infty)$, without affecting the result. A similar result is obtained under Assumption 6.2.1 for a continuized system in (Nabli et al., 2023, Proposition 6), but with a different dynamic from (CNE), specifically adapted to asynchronous algorithms.

Because $\tilde{x}_k$ in (CNA) is exactly $x_t$ in (CNE) evaluated at the time $T_k$, one can deduce a convergence result in the discrete case.

**Proposition 6.3.1.** *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d. of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption 6.2.1, iterations of (CNA) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' > -\eta$ verify*

$$\mathbb{E}\left[\min_{0\leq i\leq k-1}\|\nabla f(\tilde{y}_i)\|^2\right] \leq \frac{4}{\gamma}\frac{\Delta_f}{k}.$$

See the proof in Appendix 6.B. Intuitively, the idea is to evaluate (6.2) at $t = T_n$, from which will follow a convergence result for (CNA). Proposition 6.3.1 implies that we find, in expectation, an $\varepsilon$-stationary point in at most $O(\varepsilon^{-2})$ gradient evaluations. This is a new result for the continuized algorithm (CNA), which shows it recovers the known bound of the classical Nesterov momentum algorithms in this setting, see (Ghadimi and Lan, 2016).

## 6.4 Main Result under Lipschitz Hessian Assumption

The authors of (Okamura et al., 2024) show that under Assumptions 6.2.1 and 6.2.2, if $(x_t)_{t\geq 0}$ satisfies the Heavy Ball equation (HB)

$$\ddot{x}_t + \alpha\dot{x}_t + \nabla f(x_t) = 0, \tag{HB}$$

with $\alpha = \mathcal{O}(T^{-1/7})$, then defining for $t \geq 0$ the following average

$$\overline{x}_t = \int_0^t w_t(s) x_s ds \tag{6.3}$$

with $w_t(s) := \frac{\alpha e^{\alpha s}}{e^{\alpha t} - 1}$, one has

$$\min_{0\leq t\leq T}\|\nabla f(\overline{x}_t)\| = \mathcal{O}(T^{-4/7}). \tag{6.4}$$

Fixing $\varepsilon > 0$, (6.4) implies that we find a point $\hat{x}$ such that $\|\nabla f(\hat{x})\| \leq \varepsilon$ in at most $\mathcal{O}(\varepsilon^{-7/4})$ running time. This can be seen as a continuous version of the $\mathcal{O}(\varepsilon^{-7/4})$ bound on the number of iterations of a given algorithm. Inspired by their analysis, we state in the next section a result that holds for an algorithm of the form (NM), using the continuized method.

### 6.4.1 Convergence Result

We state our main result.

**Theorem 6.4.1.** *Let $\{T_k\}_{k\in\mathbb{N}^*}$ such that $T_0 = 0$, $T_{k+1} - T_k \overset{i.i.d}{\sim} \mathcal{E}(1)$. Let $\tilde{\varepsilon} \in (0, 1/2]$, and $c_{upp} \geq 1$ such that it satisfies $1 - \tilde{\varepsilon} - \log(1-\tilde{\varepsilon}) = c_{upp} - \log(c_{upp})$, which implies $c_{upp} \leq 1 + 2\tilde{\varepsilon}$. Under Assumptions 6.2.1-6.2.2, consider the iterations of (CNA) with $\gamma = \frac{1}{L}$, $\gamma' = \frac{1}{L} + \sqrt{\frac{1}{2L}}$, $\eta = \sqrt{\frac{1}{2L}}$, $\eta' = \alpha - \sqrt{\frac{1}{2L}}$, where*

$$\alpha = \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{(1-\tilde{\varepsilon})^2}\right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L^3 n}\right)^{\frac{1}{7}}.$$

*Then, if $n$ is large enough to ensure $2((1+2\tilde{\varepsilon})n)^{-1} < \alpha \leq 1$, we have*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{k\in\{1,\dots n\}} \|\nabla f(\overline{x}_k)\|\right] \leq \mathcal{B} L^{\frac{2}{7}} L_2^{\frac{1}{7}} {\Delta_f}^{\frac{4}{7}} n^{-4/7},$$

*where $\overline{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k} \tilde{y}_i$ with $\lambda_{i,k} = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^{k} e^{\alpha T_j}}$, $\mathbb{P}(\mathcal{A}_n) \geq 1 - \tilde{\varepsilon}$, and $\mathcal{B}$ is such that it is uniformly bounded with $n$ and $\lim_{n\to+\infty} \mathcal{B} \leq 9.2(1+2\tilde{\varepsilon})^2 \left(\frac{1+2\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2}\right)^{1/7}$.*

A more general statement can be found in Theorem 6.5.2, in which we provide a result while relaxing the constraint of the value of $n$, to the cost of a less clean bound. From Theorem 6.4.1, we deduce that (CNA) outputs a point $\tilde{x} := \operatorname{argmin}_{k\in\{1,\dots n\}} \|\nabla f(\overline{x}_k)\|$ that verifies $\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \|\nabla f(\tilde{x})\|\right] \leq \varepsilon$ in at most $\mathcal{O}\left(\frac{L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}\right)$ gradient evaluations, as long as $\varepsilon$ is small enough. We may remove the expectation and obtain a result with a given probability, the Markov property being a direct way to do so. The stochasticity induced by the process created significant challenges in the analysis, that we solved by deriving new concentration inequalities, based on an idea by Irène Waldspurger. One of these establishes that for any $\tilde{\varepsilon} \in (0,1)$ and $n \in \mathbb{N}^*$, there exists $\mathcal{C}_n$ such that with probability $1 - \tilde{\varepsilon}$

$$\forall\, i \in \{1,\dots,n\}, \quad \sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} \leq \mathcal{C}_n\, \mathbb{E}\left[\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)}\right],$$

where $\mathcal{C}_n$ is uniformly upper-bounded with $n$ and $T_i - T_j \overset{i.i.d}{\sim} \Gamma(i-j, 1)$, see details in Theorem 6.5.1.

**Numerical factor and asymptotic improvement** The factor $\mathcal{B}$, explicitly stated in (6.23), has a somewhat intricate finite-time expression. Choosing a $\tilde{\varepsilon}$ value close to zero increases the size of the set $\mathcal{A}_n$, on which our result holds, at the cost of a poorer constant $\mathcal{B}$ in *finite time*. However, asymptotically with $n$, the factors that deteriorate the bound when $\tilde{\varepsilon}$ is small vanish. So, in this asymptotic regime, we can choose $\tilde{\varepsilon}$ small enough to ensure for instance $\lim_{n\to+\infty} \mathcal{B} \leq 10$. Then for $\varepsilon$ small enough, from Theorem 6.4.1 we deduce that we reach in expectation a $\varepsilon$-solution in at most $\frac{10^{7/4} L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}} \approx \frac{56 L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}$ gradient evaluations. The best existing bound Li and Lin (2023) is $\frac{82^{7/4} L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}} \approx \frac{2234 L^{1/2} L_2^{1/4} \Delta_f}{\varepsilon^{7/4}}$, such that we considerably improve the numerical constant factor in the asymptotic regime.

**Horizon-dependent parameter** We note that as the analysis of Okamura et al. (2024) requires knowledge of the terminal time $T$ to tune parameters in the continuous setting, our result uses the final iteration $n$ to tune $\eta'$ in Theorem 6.4.1. There is a possibility that this dependence can be removed by using an iteration-dependent schedule of the form $\eta'_k = \mathcal{O}(k^{-1/7}) - \sqrt{\gamma/2}$, at the cost of significant additional technical complexity. In a similar way, prior works achieving the $\mathcal{O}(\varepsilon^{-7/4})$ complexity involve $\varepsilon$-depending parameters Carmon et al. (2017); Jin et al. (2018); Li and Lin (2023), with the notable exception of the parameter-free restart method of Marumo and Takeda (2022).

**Experiment: the behavior of** (CNA) **in the small stepsize regime** Empirically, we observe that in the small stepsize regime, the behavior of (CNA) is very close to a deterministic version. On Figure 6.1, we run (CNA) on a matrix factorization problem, that satisfies Assumptions 6.2.1 and 6.2.2 on a subspace of the domain, see (Jin et al., 2017, Lemma 6) for admissible values of $L$ and $L_2$. We also run a deterministic instance

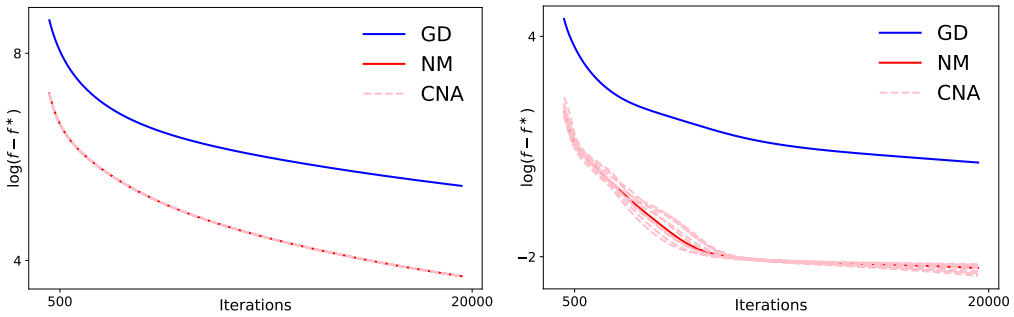



Figure 6.1: We plot one run of gradient descent, one run of a deterministic instance of (NM), and ten runs of (CNA) on a matrix factorization problem. On the left, we choose $\gamma$ according to the theoretical result of (Jin et al., 2017, Lemma 6). On the right, we chose a more aggressive stepsize and momentum parameter $\alpha$, i.e. smaller. We note that because we pushed the parameters closer to the stability limit, one of the 10 runs of (CNA) diverged and is not plotted.

of (NM) by approximating the parameters of (CNA), designed as follows. We have $\frac{\eta}{\eta+\eta'}\left(1-e^{-(\eta+\eta')\Delta_k}\right) \approx \eta(T_{k+1}-T_k)$ if $\eta+\eta'$ is small, while $\eta'\frac{1-e^{-(\eta+\eta')\Delta_k}}{\eta'+\eta e^{-(\eta+\eta')\Delta_k}} \approx \frac{\eta'(\eta+\eta)\Delta_k}{\eta'+\eta(1-(\eta+\eta'))\Delta_k}$. Replacing roughly $\Delta_k$ by its expectation 1, we then chose $\alpha_k=\eta$ and $\beta_k=\frac{\eta'(\eta+\eta)}{\eta'+\eta(1-(\eta+\eta'))}$ as our deterministic sequences of parameters for (NM). We observe on Figure 6.1 that in the small stepsize regime, the continuized version behaves very similarly to the deterministic version, with function values curves being superposed. Behaviors start differentiating when increasing stepsize and choosing a more aggressive momentum parameter.

## 6.5 Overview of the proof of Theorem 6.4.1

This section provides a proof outline of Theorem 6.4.1, with most technical details deferred to the appendix. We recall that $(x_t,z_t)_{t\geq 0}$ are defined as follows

$$\left\{\begin{array}{ll} dx_t & =\eta(z_t-x_t)dt-\gamma\nabla f(x_{t^-})dN_t,\\ dz_t & =\eta'(x_t-z_t)dt-\gamma'\nabla f(x_{t^-})dN_t,\end{array}\right. \tag{CNE}$$

with $dN_t=\sum_{k\geq 1}\delta_{T_k}(dt)$ for $\{T_k\}_{k\in\mathbb{N}^*}$ a sequence of random variables satisfying $T_{k+1}-T_k \overset{i.i.d}{\sim} \mathcal{E}(1)$, with convention $T_0=0$. Throughout the proof, we note $\alpha:=\eta+\eta'$ and use the shorthand notation $\int_0^t := \int_{[0,t]}$ to save space. Inspired from the averaging (Okamura et al., 2024, Equations (4)-(5)), we consider a Poisson-average of the trajectory $x_t$

$$\overline{x}_t := \int_0^t w_t(s)x_{s^-}dN_s, \tag{6.5}$$

where we set $w_t(s)=\frac{\alpha e^{\alpha s}}{\int_0^t \alpha e^{\alpha s'}dN(s')}$. To readers unfamiliar with Poisson processes, we emphasize that Poisson integrals are stochastic sums. Namely, for a measurable function $\phi$,

$$\int_0^t \phi(s)dN_s=\sum_{i=1}^{N_t}\phi(T_i)$$

where $N_t=\sum_{k\geq 1}\mathbb{1}_{T_k\leq t}\sim\mathcal{P}(t)$. So, we can also write $\overline{x}_t=\sum_{i=1}^{N_t}w_t(T_i^-)x_{T_i^-}$. In particular, at $t=T_n$, by definition $N_{T_n}=n$ and the number of terms in the sum is fixed, becoming $\sum_{i=1}^{n}w_t(T_i^-)x_{T_i^-}$. To ensure $\int_0^t w_t(s)dN_s=1$, we assume $N_t\geq 1$ (otherwise the integral is zero). This is not a restriction, because $N_t<1$ would imply no jumps occurred, *i.e.* the underlying algorithm has not started yet.

**Plan of the Proof** Our proof strategy is the following, divided into seven steps.

I. We establish some upper-bounds on trajectory-depending quantities, that follow from the Lyapunov control of Section 6.3, which uses the Lipschitz gradient property. These controls are used in step III and IV.

II. We upper-bound by three terms an expected weighed average $\mathbb{E}\left[\int_0^{T_n} \Lambda_t \left\|\nabla f(\overline{x}_t)\right\| dN_t\right]$, for some weighting parameter $\Lambda_t$. It uses the Hessian Lipschitz property, and a stochastic integration by part formula.

III. We bound the first of these three terms, by a quantity of the order $\mathcal{O}(\sqrt{\alpha n})$. The main argument is the Cauchy-Schwartz inequality and our Lyapunov control from I.

IV. We bound with high probability the two remaining terms, by a quantity of the order $\mathcal{O}(\alpha^{-3})$. This part is the most tedious of the proof. Apart from our Lyapunov control from I, it requires to use a concentration inequality on the underlying jump times $\{T_k\}_{k\in\mathbb{N}^*}$, thus a result holding on a subset of realization $\mathcal{A}_n$.

V. We combine the bounds derived in III and IV, and show that with the choice $\alpha = \mathcal{O}(n^{-1/7})$, we obtain a result of the form

$$\mathbb{E}\left[\min_{t\in\{T_1,\dots,T_n\}} \left\|\nabla f(\overline{x}_t)\right\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] = \mathcal{O}(n^{3/7}).$$

VI. We obtain a lower-bound of the form $\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t \geq Cn$ for some constant $C < 1$, that holds with high probability.

VII. We state our final convergence results. The final key argument it that although it involves the continuous-time process $\overline{x}_t$, $\arg\min_{t\in\{T_1,\dots,T_n\}} \left\|\nabla f(\overline{x}_t)\right\|$, can in fact be computed by (CNA).

### 6.5.1 Part I - Lyapunov Controls from Section 6.3

In this first part, we establish bounds on trajectory-depending quantities that we can deduce from the analysis carried in Section 6.3. Under Assumption 6.2.1, we proved in Lemma 6.3.1 inequality (6.2), namely

$$\int_0^t (\eta+\eta') \left\|x_s - z_s\right\|^2 + \frac{\gamma}{4} \left\|\nabla f(x_s)\right\|^2 ds \leq f(x_0) - f^* + M_t,$$

with $M_t$ a martingale such that $\mathbb{E}[M_t] = 0$. From this, we deduce

$$\int_0^t (\eta+\eta') \left\|x_s - z_s\right\|^2 ds \leq f(x_0) - f^* + M_t,$$

$$\int_0^t \frac{\gamma}{4} \left\|\nabla f(x_{s^-})\right\|^2 dN_s \leq f(x_0) - f^* + U_t, \tag{6.6}$$

$$\int_0^t (\eta+\eta') \left\|x_{s^-} - z_{s^-}\right\|^2 dN_s \leq f(x_0) - f^* + V_t, \tag{6.7}$$

where for some martingales $U_t$, $V_t$ with mean zero, (6.6) and (6.7) hold because $\int_0^t (\eta+\eta') \left\|x_s - z_s\right\|^2 ds = \int_0^t (\eta+\eta') \left\|x_{s^-} - z_{s^-}\right\|^2 dN_s$ and $\int_0^t \frac{\gamma}{4} \left\|\nabla f(x_s)\right\|^2 ds = \int_0^t \frac{\gamma}{4} \left\|\nabla f(x_{s^-})\right\|^2 dN_s$, up to martingales additive terms whose expectations are zero. Now, using Theorem 6.B.1, evaluating these bounds at $t = T_n$, they become

$$\mathbb{E}\left[\int_0^{T_n} (\eta+\eta') \left\|x_s - z_s\right\|^2 ds\right] \leq \mathbb{E}\left[f(x_0) - f^*\right], \tag{6.8}$$

$$\mathbb{E}\left[\int_0^{T_n} \frac{\gamma}{4} \left\|\nabla f(x_{s^-})\right\|^2 dN_s\right] \leq \mathbb{E}\left[f(x_0) - f^*\right], \tag{6.9}$$

$$\mathbb{E}\left[\int_0^{T_n} (\eta+\eta') \left\|x_{s^-} - z_{s^-}\right\|^2 dN_s\right] \leq \mathbb{E}\left[f(x_0) - f^*\right]. \tag{6.10}$$

The three inequalities (6.8), (6.9), (6.10) will play a central role. A large part of the remainder of the proof is devoted to preparing the ground for the application of these inequalities.

### 6.5.2 Part II - Using the Hessian Lipschitz Property to Obtain a Three Terms Upper-Bound

We adapt the analysis of (HB) carried by (Okamura et al., 2024) to derive an analysis of (CNE). Ultimately, our goal is to obtain a control of the norm of the gradient of the average (6.5), namely of $\|\nabla f(\overline{x}_t)\|$. The Lipschitz Hessian property (Assumption 6.2.2) allows us to control the distance of $\nabla f(\overline{x}_t)$ to a quantity of interest, namely the average of the gradient.

**Lemma 6.5.1.** *Let $(x_t, z_t)_{t\geq 0}$ be solution of* (CNE)*. Under Assumption 6.2.2, we have*

$$\begin{aligned}
&\left\| \int_0^t w_t(s) \nabla f(x_{s^-}) dN_s - \nabla f(\overline{x}_t) \right\| \\
&\leq L_2 \int_0^t \eta \left\| z_s - x_s \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \qquad (6.11) \\
&+ L_2 \int_0^t \gamma^2 \left\| \nabla f(x_{s^-}) \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s. \qquad (6.12)
\end{aligned}$$

See the proof in Appendix 6.C.1. The quantities $\|z_s - x_s\|^2$ and $\|\nabla f(x_{s^-})\|^2$ appearing in (6.11) and (6.12) are interesting. Indeed, as showed in Section 6.5.1, we can bound the expectation of integrals of these quantities.

In contrast, we do not have a direct control of $\int_0^t w_t(s) \nabla f(x_{s^-}) dN_s$. However, because of the choice of averaging $w_s(\cdot)$ and the definition of (CNE), a stochastic integration by parts formula reduces it to a more tractable quantity.

**Lemma 6.5.2.** *If $(x_t, z_t)_{t\geq 0}$ is solution of* (CNE)*, for $t > 0$ we have*

$$\int_0^t w_t(s) \nabla f(x_{s^-}) dN_s = -\frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma}. \qquad (6.13)$$

See the proof in Appendix 6.C.2. Combining Lemmas 6.5.1 and 6.5.2 and using a triangular inequality yields the following upper-bound

$$\begin{aligned}
\|\nabla f(\overline{x}_t)\| \leq{}& \frac{w_t(t) \left\| z_t - x_t \right\|}{\gamma' - \gamma} \\
&+ L_2 \int_0^t \eta \left\| z_s - x_s \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\
&+ L_2 \int_0^t \gamma^2 \left\| \nabla f(x_{s^-}) \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s.
\end{aligned}$$

This bound is nonetheless not sufficient to use our controls of Section 6.5.1; in particular, we do not have a bound involving the non-integrated quantity $\|z_t - x_t\|$. So, we integrate between $t = 0$ and $t = T_n$ and take the expectation on the above inequality, after multiplication of each side by a suitable quantity.

**Lemma 6.5.3.** *Noting $\Lambda_t := \left( \int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2$, for $(x_t, z_t)_{t\geq 0}$ solution of* (CNE) *under Assumption 6.2.2,*

*we have*

$$\begin{aligned}\mathbb{E}\left[\int_0^{T_n}\Lambda_t\left\|\nabla f(\overline{x}_t)\right\|dN_t\right] &\leq \frac{1}{\gamma'-\gamma}\underbrace{\mathbb{E}\left[\int_0^{T_n}\Lambda_t w_t(t)\left\|z_t-x_t\right\|dN_t\right]}_{Term\ 1}\\ &+\alpha^2\eta^2L_2\underbrace{\mathbb{E}\left[\int_0^{T_n}\left\|z_s-x_s\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)ds\right]}_{Term\ 2}\\ &+\alpha^2\gamma^2L_2\underbrace{\mathbb{E}\left[\int_0^{T_n}\left\|\nabla f(x_{s^-})\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\right]}_{Term\ 3}.\end{aligned} \tag{6.14}$$

See the proof in Appendix 6.C.3. Section 6.5.3 is devoted to the control of Term 1. Section 6.5.4 is devoted to the control of Term 2 and Term 3, these two terms sharing a similar analysis.

### 6.5.3 Part III: Control of Term 1

We obtain the following control of Term 1 in (6.14).

**Lemma 6.5.4.** *Under Assumptions 6.2.1 and 6.2.2, let $(x_t,z_t)_{t\geq 0}$ be solution of* (CNE) *with $\gamma\leq\frac{1}{L}$, $\gamma'=\gamma+\sqrt{\frac{\gamma}{2}}$, $\eta=\sqrt{\frac{\gamma}{2}}$ and $\eta'=\alpha-\eta$ for $\alpha>0$. Then, we have*

$$\mathbb{E}\left[\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2\left\|\frac{w_t(t)(z_t-x_t)}{\gamma'-\gamma}\right\|dN_t\right]\leq\sqrt{\alpha n}\sqrt{\frac{A_n}{\gamma}}\sqrt{\Delta_f},$$

*with $A_n=12\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)$.*

*Sketch of Proof.* The main idea is to use Cauchy-Schwartz inequality and our controls from Section 6.5.1. Indeed, using the notation $\Lambda_t:=\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2$ and the definition $w_t(t)=\frac{\alpha e^{\alpha t}}{\int_0^t\alpha e^{\alpha s}dN_s}$, Term 1 becomes

$$\begin{aligned}\mathbb{E}\left[\int_0^{T_n}\Lambda_t\left\|\frac{w_t(t)(z_t-x_t)}{\gamma'-\gamma}\right\|dN_t\right]&=\mathbb{E}\left[\int_0^{T_n}\frac{\alpha}{\gamma'-\gamma}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)\left\|z_t-x_t\right\|dN_t\right]\\ &\leq\frac{\alpha}{\gamma'-\gamma}\sqrt{\mathbb{E}\left[\int_0^{T_n}\Lambda_t dN_t\right]}\sqrt{\mathbb{E}\left[\int_0^{T_n}\left\|z_t-x_t\right\|^2dN_t\right]},\end{aligned} \tag{6.15}$$

where we used the Cauchy-Schwarz inequality. $\sqrt{\mathbb{E}\left[\int_0^{T_n}\Lambda_t dN_t\right]}$ can be computed explicitly. For the other term, we still cannot use our controls from Section 6.5.1. Before, we use that

$$\begin{aligned}\int_0^{T_n}\left\|z_t-x_t\right\|^2dN_t&=\int_0^{T_n}\left\|z_{t^-}-x_{t^-}-(\gamma'-\gamma)\nabla f(x_{t^-})\right\|^2dN_t\\ &\leq 2\int_0^{T_n}\left\|z_{t^-}-x_{t^-}\right\|^2dN_t+2(\gamma'-\gamma)^2\int_0^{T_n}\left\|\nabla f(x_{t^-})\right\|^2dN_t\end{aligned}$$

We note here that this step is necessary because our control (6.7) holds when integrating $\left\|z_{t^-}-x_{t^-}\right\|^2$, not $\left\|z_t-x_t\right\|^2$. The reason is that the relation $\int_0^t\phi(s)dN_s=\int_0^t\phi(s)ds+M_t$ for some centered martingale $(M_t)_{t\geq 0}$ holds as long as $\phi$ is a *previsible* process with respect to $N$, namely if $\phi(t)$ is measurable with respect to the filtration $\{N_s\}_{0\leq s<t}$. While $\left\|z_{t^-}-x_{t^-}\right\|^2$ is previsible, $\left\|z_t-x_t\right\|^2$ is not. □

See Appendix 6.C.4 for the full proof.

### 6.5.4 Part IV: Control of Term 2 and Term 3

Compared with Term 1, the remaining Terms 2 and 3 from (6.14) are more challenging. When using the classic deterministic (HB) equation, the analogue of the triple integrals inside the parentheses are deterministic, and can be directly computed Okamura et al. (2024). In our case, if the expectation of these triple integrals can be computed, we have to deal with the expectation of the full expression. To do so, we are faced with the lack of independence between the trajectory-dependent factors and these triple integrals.

To illustrate this, we consider Term 3, namely

$$\int_0^{T_n} \|\nabla f(x_{s^-})\|^2 \left( \int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau-t)} e^{\alpha(\sigma-t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t \right) dN_s.$$

It can be rearranged as

$$\int_0^{T_n} \left( \int_s^{T_n} \int_s^t e^{\alpha(\tau-t)} e^{\alpha(s-t)} \|\nabla f(x_{s^-})\|^2 \int_0^s e^{\alpha(\sigma-s)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t \right) dN_s.$$

Writing it as a sum, it becomes

$$\sum_{i=1}^{n} \sum_{l=i}^{n} \sum_{k=i}^{l} e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^{i} e^{\alpha(T_j - T_i)} (k-j). \tag{6.16}$$

Taking the expectation, we have

$$\begin{aligned}
\mathbb{E}\left[(6.16)\right] &= \mathbb{E}\left[ \sum_{i=1}^{n} \sum_{l=i}^{n} \sum_{k=i}^{l} e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^{i} e^{\alpha(T_j - T_i)} (k-j) \right] \\
&= \sum_{i=1}^{n} \sum_{l=i}^{n} \sum_{k=i}^{l} \mathbb{E}\left[ e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^{i} e^{\alpha(T_j - T_i)} (k-j) \right]
\end{aligned}$$

Importantly, the increments of $\{T_k\}_{k\in\mathbb{N}}$ are independent, namely for any integers $1 \le \ell_0 < \ell_1$, we have $T_{\ell_1} - T_{\ell_0} \perp\!\!\!\perp \phi(T_1, \cdots, T_{\ell_0})$, for any measurable function $\phi$. In our case, as $l, k \ge i$, we have $T_k - T_l \perp\!\!\!\perp \left\| \nabla f(x_{T_i^-}) \right\|$ and $T_i - T_l \perp\!\!\!\perp \left\| \nabla f(x_{T_i^-}) \right\|$. So, we deduce

$$\mathbb{E}\left[(6.16)\right] = \sum_{i=1}^{n} \sum_{l=i}^{n} \sum_{k=i}^{l} \mathbb{E}\left[ e^{\alpha(T_k - T_l)} e^{\alpha(T_i - T_l)} \right] \mathbb{E}\left[ \left\| \nabla f(x_{T_i^-}) \right\|^2 \sum_{j=1}^{i} e^{\alpha(T_j - T_i)} (k-j) \right]$$

The challenge we face here is that $\nabla f(x_{T_i^-})$ and $\sum_{j=1}^{i} e^{\alpha(T_j - T_i)} (k-j)$ are not independent. We address this difficulty in the next section.

#### Splitting the expectations

In the previous section, we showed that we arrive at a point where trajectory-dependent factors are multiplied with some trajectory-independent terms, from which they are no independent, which prevents to separate the expectations. Precisely, the problematic trajectory-independent factors we obtain are

$$\sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)}, \ \sum_{j=1}^{i} e^{-\alpha(T_i - T_j)}, \ \sum_{j=1}^{i} (i-j) e^{-\alpha(T_i - T_j)}, \quad i \in \{1, \ldots, n\}.$$

Our solution is to show that with high probability, these terms can be upper-bounded by their expectation multiplied by a constant factor. The trickiest part is to find a probability and a constant that do not become trivial as $n$ grows. In particular, it cannot work by directly using the Markov property.

**Theorem 6.5.1.** *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0,1)$, $\alpha \in (0,1)$. Let $K = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil$, and $c_{upp} \geq 1$ be chosen such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let*

$$\mathcal{C}_n = 32(4 + 4C_n + C_n^2)\frac{c_{upp}e}{c_{inf}^2}, \text{ where } C_n = c_{inf}\alpha K.$$

*There exists a set $\mathcal{A}_n \subset \Omega$ of realizations satisfying $\mathbb{P}(\mathcal{A}_n) \geq 1 - \varepsilon$ such that the three following assertions hold*

$$\begin{aligned}
&(1) \quad \forall\, i \in \{K+1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)}\right],\\
&(2) \quad \forall\, i \in \{1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)}\right],\\
&(3) \quad \forall\, i \in \{1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i} (i - j) e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i} (i - j) e^{-\alpha(T_i - T_j)}\right].
\end{aligned}$$

Section 6.D is devoted to the complete proof of Theorem 6.5.1, which is based on an idea by Irène Waldspurger. One of its key step is to use concentration inequalities to obtain that with a high probability that does not depend on $n$, we ensure both an upper and lower-bound on the increments $T_i - T_j$, of the form

$$c_{inf}\mathbb{E}\left[T_i - T_j\right] - c_{inf}K \leq T_i - T_j \leq c_{upp}\mathbb{E}\left[T_i - T_j\right] + c_{upp}K,$$

where $K = \mathcal{O}(\log(n/\varepsilon))$. Precisely, the set $\mathcal{A}_n$ is the set on which the above holds. These bounds then allow to relate the sums with their expectations.

Thanks to Theorem 6.5.1, with high probability one can fully separate the expectation of the factor that depend on the trajectory from those who do not. This is formalized in the following result.

**Lemma 6.5.5.** *We assume the same setting as Theorem 6.5.1.*

1. *Noting $A_k := \int_{[T_k, T_n]} \int_{[T_k, t]} \int_{[0, T_k]} e^{\alpha(T_k - t)} e^{\alpha(\tau - t)} e^{\alpha(\sigma - T_k)} (\tau - \sigma) dN_\sigma dN_\tau dN_t$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \int_0^{T_n} \eta^2 \left\|z_s - x_s\right\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t\right) ds\right]\\
&\leq \mathcal{C}_n \max_{k \in \{1, \dots, n\}} \mathbb{E}\left[A_k\right] \mathbb{E}\left[\int_0^{T_n} \eta^2 \left\|z_s - x_s\right\|^2 ds\right]
\end{aligned} \tag{6.17}$$

2. *Noting $B_k := \int_{T_k}^{T_n} \int_{T_k}^{t} \int_0^{T_k} e^{\alpha(\tau - t)} e^{\alpha(T_k - t)} e^{\alpha(\sigma - T_k)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \int_0^{T_n} \gamma^2 \left\|\nabla f(x_{s^-})\right\|^2 \left(\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t\right) dN_s\right]\\
&\leq \mathcal{C}_n \max_{k \in \{1, \dots, n\}} \mathbb{E}\left[B_k\right] \mathbb{E}\left[\int_0^{T_n} \gamma^2 \left\|\nabla f(x_{s^-})\right\|^2 dN_s\right]
\end{aligned} \tag{6.18}$$

The proof of Lemma 6.5.5 uses the ideas introduced in the beginning of Section 6.5.4, such as writing the integrals as sums (see (6.16)). See Appendix 6.C.5 for the complete proof.

### Final Control

Thanks to Lemma 6.5.5, we see that we can control separately the expectations of trajectory-dependent factors, and the factors $\max_{k \in \{1, \dots, n\}} \mathbb{E}\left[A_k\right]$ and $\max_{k \in \{1, \dots, n\}} \mathbb{E}\left[B_k\right]$. The first can be dealt with using the Lyapunov controls of Section 6.5.1. For the two other terms, we note that we can write the quantities $A_k$ and $B_k$ as sums, for instance in the case of $B_k$

$$\int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t = \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} e^{2\alpha(T_l - T_i)} e^{\alpha(T_j - T_l)} (l - j).$$

The expectation of the terms in this sum can be computed using basic properties of Gamma laws. Indeed, as $T_l - T_i \sim \Gamma(l-i,1)$, we recognize that $\mathbb{E}\left[e^{2\alpha(T_l-T_i)}\right]$ is the characteristic function of a gamma law evaluated at $2\alpha$, whose value is $(1+2\alpha)^{-(l-i)}$. The terms appearing in $\mathbb{E}[A_k]$ are not significantly more difficult. Then, $\max_{k\in\{2,\dots,n\}}\mathbb{E}[A_k]$ and $\max_{k\in\{2,\dots,n\}}\mathbb{E}[B_k]$ can be upper-bounded with basic algebraic manipulations, yielding the following bounds.

**Lemma 6.5.6.** *Consider $A_k$ and $B_k$ as defined in Lemma 6.5.5. Then*

1. $\max_{k\in\{2,\dots,n\}}\mathbb{E}[A_k] \leq \frac{(1+\alpha)(1+2\alpha)}{\alpha^4}$.
2. $\max_{k\in\{1,\dots,n\}}\mathbb{E}[B_k] \leq \frac{(1+\alpha)^2(1+2\alpha)}{\alpha^4}$.

See the proof in Appendix 6.C.6. Combining Lemma 6.5.6 with our controls from Section 6.5.1, we finally obtain the following bound on Terms 2 and 3.

**Lemma 6.5.7.** *Assume the same setting as in Theorem 6.5.1. Under Assumptions 6.2.1 and 6.2.2, let $(x_t,z_t)_{t\geq 0}$ be solution of* (CNE) *with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' = \alpha - \eta$ for $\alpha \in (0,1)$. Then, we have*

$$\begin{aligned}
&\alpha^2\eta^2 L_2 \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\int_0^{T_n}\|z_s - x_s\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)ds\right]\\
&+\alpha^2\gamma^2 L_2\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\int_0^{T_n}\|\nabla f(x_{s^-})\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\right]\\
&\leq \frac{\mathcal{C}_n L_2\gamma}{\alpha^3}B_n\Delta_f,
\end{aligned}$$

*with $B_n := \frac{1}{2}(1+\alpha)(1+2\alpha)\left(1+\alpha 8(1+\alpha)\right)$.*

See the proof in Appendix 6.C.7.

### 6.5.5 Part V: Tuning $\alpha$

Combining Lemma 6.5.3, and the controls provided by Lemma 6.5.4 (Term 1) and Lemma 6.5.7 (Terms 2 and 3), we obtain that under the statement of Theorem 6.4.1, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2\|\nabla f(\overline{x}_t)\|\,dN_t\right] &\leq \sqrt{\alpha n}\sqrt{\frac{A_n}{\gamma}}\sqrt{\Delta_f}\\
&\quad+\frac{\mathcal{C}_n L_2\gamma}{\alpha^3}B_n\Delta_f
\end{aligned} \tag{6.19}$$

with, $A_n := 12\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)$ and $B_n := \frac{1}{2}(1+\alpha)(1+2\alpha)\left(1+\alpha 8(1+\alpha)\right)$.

Because we integrate with respect to $dN_t$ on $[0,T_n]$, we have

$$\min_{t\in\{T_1,\dots,T_n\}}\|\nabla f(\overline{x}_t)\|\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t \leq \int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2\|\nabla f(\overline{x}_t)\|\,dN_t.$$

Therefore, from (6.19), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\min_{t\in\{T_1,\dots,T_n\}}\|\nabla f(\overline{x}_t)\|\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right]\\
&\leq \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2\|\nabla f(\overline{x}_t)\|\,dN_t\right]\\
&\overset{(6.19)}{\leq} \sqrt{\alpha n}\sqrt{\frac{A_n}{\gamma}}\sqrt{\Delta_f}+\frac{\mathcal{C}_n L_2\gamma}{\alpha^3}B_n\Delta_f\\
&= n^{3/7}\left(\sqrt{A_n C_\alpha L^{1+a_1}L_2^{a_2}{\Delta_f}^{1+a_3}}+B_n {C_\alpha}^{-3}\mathcal{C}_n L^{-1-3a_1}L_2^{1-3a_2}{\Delta_f}^{1-3a_3}\right),
\end{aligned} \tag{6.20}$$

where in the last inequality, we chose $\gamma = 1/L$ and defined

$$\alpha = C_\alpha n^{-1/7} L^{a_1} L_2^{a_2} \Delta_f^{a_3},$$

where $C_\alpha > 0$, $a_1, a_2, a_3 \in \mathbb{R}$ are defined such that $\alpha \leq 1$. The choice $\alpha = \mathcal{O}(n^{-1/7})$ is derived from a trade-off between $\sqrt{\alpha}$ and $\alpha^{-3}$. It is consistent with the choice made in Okamura et al. (2024), in the case of studying the Heavy Ball (HB) ODE.

We exhibit two admissible choices of $\alpha$, subject to the constrain $\alpha \leq 1$. We note that in general, this forces to set a condition on the value of the number of iteration $n$, which may depend on $L$, $L_2$ or $\Delta_f$. If one wants to avoid such conditions, one can choose $C_\alpha = 1$, $a_1 = a_2 = a_3 = 0$.

**Corollary 6.5.1** (Choice I)**.** *Let $\alpha = n^{-1/7} \leq 1$. In this case, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] \\ &\leq 16(1+2n^{-1/7})^4 \left(1+8(1+2n^{-1/7})\right)(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2}\left(\sqrt{L\Delta_f}+L^{-1}L_2\Delta_f\right) n^{3/7},\end{aligned} \tag{6.21}$$

*with $C_n = c_{inf} n^{-1/7} \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$.*

Otherwise, if we set a condition of $n$, we can obtain a cleaner result.

**Corollary 6.5.2** (Choice II)**.** *Let*

$$\alpha = \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{c_{inf}^2}\right)^{2/7} \left(\frac{L_2^2 \Delta_f}{Ln}\right)^{\frac{1}{7}}.$$

*If $\alpha \leq 1$, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] \\ &\leq (1+C_n+C_n^2/4) 7\, 6^{-6/7}\, 12^{3/7} 64^{1/7} \left(\frac{c_{upp}e}{c_{inf}^2}\right)^{1/7} \left(\frac{c_{upp}e}{c_{inf}^2}\right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} \Delta_f^{\frac{4}{7}} n^{3/7},\end{aligned}$$

*with $C_n = c_{inf}\alpha \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$ and $= (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$.*

The proofs of Corollary 6.5.1 and 6.5.2 are in Appendix 6.C.8.

### 6.5.6 Part VI: Control of the double Poisson integral factor

Up to now, we managed to get an upper-bound of the form

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] = \mathcal{O}(n^{3/7}), \tag{6.22}$$

where we recall that by Theorem 6.5.1, $\mathcal{A}_n$ is such that $\mathbb{P}(\mathcal{A}_n) \geq 1-\varepsilon$. As we chose $\alpha = \mathcal{O}(n^{-1/7})$, we have (see Lemma 6.E.4-$(ii)$)

$$\mathbb{E}\left[\int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] = \mathcal{O}(n),$$

which heuristically suggests that $\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\|\right] = \mathcal{O}(n^{-4/7})$. However, the two factors inside the expectations in (6.22) are not independent, as they both depend on the $n$ first jump times. We therefore cannot separate the expectation. To circumvent this difficulty, we use the following result.

**Lemma 6.5.8.** *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0,1)$, $\alpha \in (0,1]$ and $K = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$. Let $c_{upp} \geq 1$ be chosen such that $c_{inf} - \log(c_{inf}) = c_{upp} - \log(c_{upp})$ and $C_n = c_{inf}\alpha K$. Then, with the same set $\mathcal{A}_n$ as in Theorem 6.5.1, we have*

$$\mathbb{1}_{\mathcal{A}_n} \int_0^{T_n} \left( \int_0^t \alpha e^{\alpha(s-t)} dN_s \right)^2 dN_t \geq \mathbb{1}_{\mathcal{A}_n} \frac{n}{c_{upp}^2} e^{-2\frac{c_{upp}}{c_{inf}} C_n} \left(1 - \frac{2}{c_{upp}\alpha n}\right).$$

The proof of Lemma 6.5.8 borrows arguments used to prove Theorem 6.5.1, see full proof in Appendix 6.C.9.

### 6.5.7 Part VII: Final Results

We now state our final result. There is a result with our first choice $\alpha = n^{-1/7}$, that (almost) does not require any restriction on $n$, at the cost of a less clean bound. A second result holds with a more careful choice of $\alpha$, but it induces a contraint on $n$ that has to be large enough.

**Theorem 6.5.2** (Complete Version)**.** *Let $\{T_k\}_{k\in\mathbb{N}^*}$ such that $T_0 = 0$, $T_{k+1} - T_k \overset{i.i.d}{\sim} \mathcal{E}(1)$. Let $c_{inf}, \varepsilon \in (0,1)$, and $c_{upp} \geq 1$ such that it satisfies $c_{inf} - \log(c_{inf}) = c_{upp} - \log(c_{upp})$.*

*Under Assumptions 6.2.1-6.2.2, consider the iterations of (CNA) with $\gamma = \frac{1}{L}$, $\gamma' = \frac{1}{L} + \sqrt{\frac{1}{2L}}$, $\eta = \sqrt{\frac{1}{2L}}$, $\eta' = \alpha - \sqrt{\frac{1}{2L}}$.*

*(i) Let $\alpha = n^{-1/7}$. If $n \geq 3$, we have*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{k\in\{1,\dots n\}} \|\nabla f(\overline{x}_k)\|\right] \leq \mathcal{A}\left(\sqrt{L\Delta_f} + L^{-1}L_2\Delta_f\right) n^{-4/7},$$

*where*

$$\mathcal{A} := 16(1+2n^{-1/7})^4 \left(1 + 8(1+2n^{-1/7})\right)(4 + 4C_n + C_n^2) \left(1 - \frac{2}{c_{upp}n^{6/7}}\right)^{-1} \frac{c_{upp}^3}{c_{inf}^2} e^{1+2\frac{c_{upp}}{c_{inf}}C_n},$$

*with $C_n = c_{inf}\alpha \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$.*

*(ii) Let $\alpha = \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{c_{inf}^2}\right)^{2/7} \left(\frac{L_2^2\Delta_f}{L^3 n}\right)^{\frac{1}{7}}$. If $n$ is large enough to ensure $2(c_{upp}n)^{-1} < \alpha \leq 1$, then we have*

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \min_{k\in\{1,\dots n\}} \|\nabla f(\overline{x}_k)\|\right] \leq \mathcal{B} L^{\frac{2}{7}} L_2^{\frac{1}{7}} {\Delta_f}^{\frac{4}{7}} n^{-4/7},$$

*with*

$$\mathcal{B} := (1 + C_n + C_n^2/4) 7\, 6^{-6/7}\, 12^{3/7} (64e)^{1/7} \left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} c_{upp}^2 \left(\frac{c_{upp}}{c_{inf}^2}\right)^{1/7} e^{2\frac{c_{upp}}{c_{inf}}C_n},$$

*with $C_n = c_{inf}\alpha \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$ and $= (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$.*

*In both statements, $\overline{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k}\tilde{y}_i$ with $\lambda_{i,k} = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^k e^{\alpha T_j}}$, and $\mathbb{P}(\mathcal{A}_n) \geq 1 - \varepsilon$.*

While the constant $\varepsilon$ determines the size of the set $\mathcal{A}_n$, the role of $c_{inf}$ and $c_{upp}$ might be more difficult to grasp. These two constants appear during the proof of Theorem 6.5.1. Choosing $c_{inf}$ close to one, which makes $c_{upp}$ close to 1, makes $C_n$ increase. Choosing $c_{inf} \ll 1$, inducing $c_{upp} \gg 1$, reduces $C_n$ but still hurts the bound because of factors such as $c_{upp}^2$.

*Proof.* Corollary 6.5.1 and 6.5.2 provide bounds that involve $\min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\|$, which is about the continuous-time process $\overline{x}_t$. We have to make sure that the bound will be achievable by our algorithm, precisely, we have to show that the point $\arg\min_{t=\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\|$ can be outputted by (CNA). We have

$$\overline{x}_k := \overline{x}_{T_k} = \int_{[0,T_k]} w_{T_k}(s) x_{s^-} dN_s = \sum_{i=1}^{k} w_{T_k}(T_i) x_{T_i^-} = \sum_{i=0}^{k-1} w_{T_k}(T_{i+1}) \tilde{y}_i,$$

using $x_{T_i^-} = \tilde{y}_{i-1}$ by Proposition 6.2.1. Also, recall that $w_t(s) = \frac{e^{\alpha s}}{\int_0^t e^{\alpha s} dN_s}$, such that we fix

$$\lambda_{i,k} := w_{T_k}(T_{i+1}) = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^{k} e^{\alpha T_j}}.$$

So, for $\overline{x}_k = \sum_{i=0}^{k-1} \lambda_{i,k} \tilde{y}_i$, $\lambda_{i,k} = \frac{e^{\alpha T_{i+1}}}{\sum_{j=1}^{k} e^{\alpha T_j}}$, we have

$$\min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\| = \min_{k\in\{1,\dots n\}} \|\nabla f(\overline{x}_k)\|.$$

Then, the result is a direct combination of the bounds provided Corollary 6.5.1 and 6.5.2 with Lemma 6.5.8. We just have to make sure that $\left(1 - \frac{2}{c_{upp}\alpha n}\right) > 0$. In the case $\alpha = n^{-1/7}$, it holds as long as $n \geq 3$. More generally it is verified as long as $\alpha > \frac{2}{c_{upp} n}$. In the case $\alpha = \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{c_{inf}^2}\right)^{2/7} \left(\frac{L_2^2 \Delta_f}{Ln}\right)^{\frac{1}{7}}$, it induces the following constraint on $n$

$$n > \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{c_{inf}^2}\right)^{-1/3} \left(\frac{L_2^2 \Delta_f}{L}\right)^{-\frac{1}{6}} 2^{7/6} c_{upp}^{-7/6}.$$

□

We can conclude with the proof of Theorem 6.4.1, which can be seen as a corollary of Theorem 6.5.2.

*Proof of Theorem 6.4.1.* It is a direct application of $(ii)$ in Theorem 6.5.2, fixing $c_{inf} = 1 - \tilde{\varepsilon}$. Then, recalling the expression of $\mathcal{B}$

$$\mathcal{B} := (1 + C_n + C_n^2/4) 7\, 6^{-6/7}\, 12^{3/7} (64e)^{1/7} \left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7} e^{2\frac{c_{upp}}{1-\tilde{\varepsilon}} C_n}, \tag{6.23}$$

with $C_n = (1-\tilde{\varepsilon})\alpha \left\lceil (-\tilde{\varepsilon} - \log(1-\tilde{\varepsilon}))^{-1} \log\left(2\frac{n}{\tilde{\varepsilon}}\right)\right\rceil$ and $= (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$. As $\alpha = \mathcal{O}(1/n^{1/7})$, we have that $C_n \to_{n\to+\infty} 0$, $\to_{n\to+\infty} 1$, and $\left(1 - \frac{2}{c_{upp}\alpha n}\right)^{-1} \to_{n\to+\infty} 1$. So,

$$\lim_{n\to+\infty} \mathcal{B} = 7\, 6^{-6/7}\, 12^{3/7} (64e)^{1/7} c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7} \leq 9.2 c_{upp}^2 \left(\frac{c_{upp}}{(1-\tilde{\varepsilon})^2}\right)^{1/7}.$$

It remains to show the bound $c_{upp} \leq 1 + 2\tilde{\varepsilon}$, which is done in Lemma 6.E.5.

□

# 6.A Complementary details to Section 6.1

## 6.A.1 Second-order stationary points

We provide a rigorous statement of the notion of second-order $\varepsilon$-stationary point, mentioned in our introductory section. We then briefly comment on it.

**Definition 6.A.1** (Second-order $\varepsilon$-stationary point, (Nesterov and Polyak, 2006))**.** *For a $L_2$-Lipschitz Hessian function $f$, $x$ is an $\varepsilon$-second-order stationary point if*

$$\|\nabla f(x)\| \leq \varepsilon, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\varepsilon L_2}.$$

A second-order $\varepsilon$-stationary point approximates a second-order stationary point, namely a point $x^*$ such that

$$\|\nabla f(x^*)\| = 0, \quad \lambda_{\min}(\nabla^2 f(x^*)) \geq 0.$$

In particular, such points are ensured to avoid *strict saddle-points*, that is, stationary points such that $\lambda_{\min}(\nabla^2 f(x^*)) < 0$. The 1d function $x \mapsto x^3$ is a simple example of a function that has a second-order stationary point (in $x = 0$) that is not a local minimizer, thus called *non-strict saddle-points*. Hopefully, several works indicate that many applications are prevented from such pathological stationary points (Kawaguchi, 2016; Bandeira et al., 2016; Mei et al., 2017; Boumal et al., 2016; Bhojanapalli et al., 2016; Ge et al., 2016, 2017).

### 6.A.2 Detailed formulations of existing security check and alternative mechanisms

For completeness, we give detailed versions of existing algorithms that achieve the complexity $\mathcal{O}(e^{-7/4})$. Algorithm 16 is a detailed version of an algorithm that performs a negative curvature step (Jin et al., 2018). Algorithm 17 is a detailed version of an algorithm that uses a restart mechanism (Li and Lin, 2023). We make a few remarks:

- We chose to state Algorithm 16 as the version from (Jin et al., 2018) rather than the earlier one from (Carmon et al., 2017), because of the significant length and trickiness of the latter.
- In its original statement, Algorithm 16 also involves a "stochastic perturbation step", used to find an $\varepsilon$-second-order stationary point (Definition 6.A.1). Because we do not study this concept, we avoided this step in our presentation.
- In the convergence result associated with Algorithm 17 (Li and Lin, 2023, Theorem 1), the restart threshold is set to $B = \sqrt{\varepsilon/L_2}$, where $\varepsilon$ denotes the target accuracy and $L_2$ is the Lipschitz constant of the Hessian. As noted by the authors, this value of $B$ can be very small in practice, causing the algorithm to restart at nearly every iteration. In such cases, the method effectively reduces to standard gradient descent, which eliminates the practical benefit of momentum. This behavior is confirmed by numerical experiments, showing little to no empirical acceleration due to the near absence of momentum steps.

  To address this issue, the authors propose an alternative restart scheme (Li and Lin, 2023, Algorithm 2), in which restarts are allowed to occur less frequently. While keeping $B = \sqrt{\varepsilon/L_2}$, the user is permitted to initialize a larger value $B_0$, potentially satisfying $B_0 \gg B$. The restart condition is then replaced by

$$k \sum_{t=0}^{k-1} \|x^{t+1} - x^t\|^2 > \max\{B^2, B_0^2\},$$

  and $B_0$ is gradually decreased until $B_0 \leq B$, at which point the algorithm reduces to Algorithm 17. It is shown (Li and Lin, 2023, Theorem 2) that this modification introduces an additional $\mathcal{O}\Big(L^{1/2}\varepsilon^{-1/4}L_2^{1/4}\log\big(\frac{L_2 B_0}{\varepsilon}\big)\Big)$ term in the gradient complexity, which does not significantly degrade the overall rate. This modification shows great empirical behaviour, enabling acceleration over gradient descent. However, this alternative scheme requires additional function evaluations, leading to a total complexity of $\mathcal{O}(\varepsilon^{-3/2})$ function queries.

## 6.B Proof of Lemma 6.3.1 and Proposition 6.3.1

Under Assumption 6.2.1, we have this classical descent lemma (**?**).

**Lemma 6.B.1.** *Under Assumption 6.2.1, if $\gamma \leq \frac{1}{L}$, we have*

$$f(x - \gamma\nabla f(x)) - f(x) \leq \gamma\left(\frac{L}{2}\gamma - 1\right)\|\nabla f(x)\|^2.$$

**Algorithm 16** Negative curvature exploitation and Nesterov momentum – Detailed version (Jin et al., 2018)

**Require:** $x_0, \eta, \theta, \gamma, s$
$v_0 \leftarrow 0$
**for** $t = 0, 1, \ldots$ **do**
  $y_t \leftarrow x_t + (1-\theta)v_t$
  $x_{t+1} \leftarrow y_t - \eta \nabla f(y_t)$
  $v_{t+1} \leftarrow x_{t+1} - x_t$
  **if** $f(x_t) \leq f(y_t) + \langle \nabla f(y_t), x_t - y_t \rangle - \frac{\gamma}{2} \|x_t - y_t\|^2$ **then**
    $(x_{t+1}, v_{t+1}) \leftarrow$ Negative-Curvature-Exploitation$(x_t, v_t, s)$
  **end if**
**end for**

**Negative-Curvature-Exploitation**$(x_t, v_t, s)$
**if** $\|v_t\| \geq s$ **then**
  $x_{t+1} \leftarrow x_t$
**else**
  $\delta \leftarrow s \cdot v_t / \|v_t\|$
  $x_{t+1} \leftarrow \arg\min_{x \in \{x_t + \delta,\, x_t - \delta\}} f(x)$
**end if**
**return** $(x_{t+1}, 0)$

To perform Lyapunov analysis, our tool is the following Itô formula, which intuitively allows stochastic derivation.

**Proposition 6.B.1** (Even et al. (2021), Proposition 2)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Let $x_t \in \mathbb{R}^d$ be a solution of*

$$dx_t = \zeta(x_t)dt + G(x_{t^-})dN_t$$

*where $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ is a locally Lipschitz function and $G : \mathbb{R}^d \to \mathbb{R}^d$ is measurable. Then,*

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla\varphi(x_s), \zeta(x_s) \rangle \, ds + \int_{[0,t]} \varphi(x_s + G(x_s)) - \varphi(x_s) ds + M_t, \tag{6.24}$$

*where $M_t$ is a martingale such that $\mathbb{E}[M_t] = 0$, $\forall t \geq 0$.*

We first restate Lemma 6.3.1.

**Lemma 6.3.1.** *If $(x_t, z_t)_{t \geq 0}$ is solution of (CNE) under Assumption 6.2.1 with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \gamma' - \gamma$ and $\eta' \geq -\eta$, we have*

$$\mathbb{E}\left[\int_0^t (\eta + \eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds\right] \leq \mathbb{E}[f(x_0) - f^*]. \tag{6.1}$$

*Proof.* Let $\overline{x}_t = (x_t, z_t)$ where $(x_t, z_t)_{t \geq 0}$ satisfies (CNE). It satisfies $d\overline{x}_t = \zeta(\overline{x}_t)dt + G(\overline{x}_t)dN(t)$, where

$$\zeta(\overline{x}_t) = \begin{pmatrix} \eta(z_t - x_t) \\ \eta'(x_t - z_t) \end{pmatrix}, \quad G(\overline{x}_t) = \begin{pmatrix} -\gamma_t \nabla f(x_t) \\ -\gamma'_t \nabla f(x_t) \end{pmatrix}$$

We apply Proposition 6.B.1 to $\varphi(\overline{x}_t)$, where

$$\varphi(x, z) = f(x) + \frac{1}{2} \|x - z\|^2,$$

inducing

$$\varphi(\overline{x}_t) = \varphi(\overline{x}_0) + \int_0^t \langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s) \rangle \, ds + \int_0^t \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) ds + M_t, \tag{6.25}$$

where $M_t$ is a martingale.

**Algorithm 17** Restarted Nesterov Momentum – Detailed version (Li and Lin, 2023)

**Require:** $x^{-1} = x^0$, $B > 0$, $K$, $\eta$, $\theta$
$k \leftarrow 0$
**while** $k < K$ **do**
  $y^k \leftarrow x^k + (1-\theta)(x^k - x^{k-1})$
  $x^{k+1} \leftarrow y^k - \eta \nabla f(y^k)$
  $k \leftarrow k+1$
  **if** $k \sum_{t=0}^{k-1} \left\|x^{t+1} - x^t\right\|^2 > B^2$ **then**
    $x^{-1} \leftarrow x^k$
    $x^0 \leftarrow x^k$
    $k \leftarrow 0$
  **end if**
**end while**
$K_0 \leftarrow \arg\min_{\lfloor K/2 \rfloor \leq k \leq K-1} \left\|x^{k+1} - x^k\right\|$
$\hat{y} \leftarrow \dfrac{1}{K_0+1} \sum_{k=0}^{K_0} y^k$
**return** $\hat{y}$

**Computations** We compute $\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle$ and $\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s)$. We have

$$\frac{\partial \varphi}{\partial x} = \nabla f(x) + x - z, \quad \frac{\partial \varphi}{\partial z} = z - x.$$

So,

$$\begin{aligned}\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle &= \langle \nabla f(x_s) + x_s - z_s, \eta(z_s - x_s)\rangle + \langle z_s - x_s, \eta'(x_s - z_s)\rangle \\ &= \eta \langle \nabla f(x_s), z_s - x_s\rangle - (\eta + \eta') \|x_s - z_s\|^2 . \end{aligned} \tag{6.26}$$

Also,

$$\begin{aligned}\varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) &= f(x_s - \gamma\nabla f(x_s)) - f(x_s) \\ &\quad + \frac{1}{2}\left(\|x_s - \gamma\nabla f(x_s) - (z_s - \gamma'\nabla f(x_s)\|^2 - \|x_s - z_s\|^2\right) \\ &= f(x_s - \gamma\nabla f(x_s)) - f(x_s) + \frac{(\gamma'-\gamma)^2}{2}\|\nabla f(x_s)\|^2 \\ &\quad + (\gamma' - \gamma)\langle \nabla f(x_s), x_s - z_s\rangle \\ &\leq \left((\gamma'-\gamma)^2 - \gamma(2 - L\gamma)\right)\frac{1}{2}\|\nabla f(x_s)\|^2 \\ &\quad + (\gamma' - \gamma)\langle \nabla f(x_s), x_s - z_s\rangle . \end{aligned} \tag{6.27}$$

The last inequality uses Lemma 6.B.1, assuming $\gamma \leq \frac{1}{L}$. We combine (6.26) and (6.27)

$$\begin{aligned}&\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) \\ &\leq (\eta - (\gamma' - \gamma))\langle \nabla f(x_{s^-}), z_s - x_s\rangle - (\eta + \eta')\|x_s - z_s\|^2 \\ &\quad + \left((\gamma'-\gamma)^2 - \gamma(2-L\gamma)\right)\frac{1}{2}\|\nabla f(x_s)\|^2 \end{aligned} \tag{6.28}$$

**Parameter tuning** We now choose the parameters driving (CNE) such that we obtain the desired result. We set $\eta = \gamma' - \gamma$ to cancel scalar product. Then fixing $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$ ensures

$$(\gamma'-\gamma)^2 - \gamma(2 - L\gamma) \leq \frac{\gamma}{2} - \gamma = -\frac{\gamma}{2}.$$

The latter holds because $\gamma \leq \frac{1}{L} \Rightarrow -\gamma(2 - L\gamma) \leq -\gamma$. We thus have

$$\langle \nabla\varphi(\overline{x}_s), \zeta(\overline{x}_s)\rangle + \varphi(\overline{x}_s + G(\overline{x}_s)) - \varphi(\overline{x}_s) \leq -(\eta + \eta')\|x_s - z_s\|^2 - \frac{\gamma}{4}\|\nabla f(x_s)\|^2 . \tag{6.29}$$

**Conclusion** Combining (6.29), (6.28) and (6.25), we get

$$\varphi(\overline{x}_t) \leq \varphi(\overline{x}_0) + \int_0^t -(\eta+\eta') \|x_s - z_s\|^2 - \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds + M_t$$
$$\Rightarrow \int_0^t (\eta+\eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t, \tag{6.30}$$

using $\varphi(\overline{x}_0) - \varphi(\overline{x}_t) \leq f(x_0) - \min_x f(x)$ and that $x_0 = z_0$. It remains to take expectation on (6.30)

$$\mathbb{E}\left[\int_0^t (\eta+\eta') \|x_s - z_s\|^2 + \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds\right] \leq \mathbb{E}\left[f(x_0) - f^*\right],$$

where we used $\mathbb{E}[M_t] = \mathbb{E}[M_0] = 0$. □

We now prove the discrete version of the result, which we first restate.

**Proposition 6.3.1.** *Assume $T_1, \dots, T_k$ are random variables such that $T_{i+1} - T_i$ are i.i.d. of law $\mathcal{E}(1)$, with convention $T_0 = 0$. Under Assumption 6.2.1, iterations of (CNA) with $\gamma \leq \frac{1}{L}$, $\gamma' = \gamma + \sqrt{\frac{\gamma}{2}}$, $\eta = \sqrt{\frac{\gamma}{2}}$ and $\eta' > -\eta$ verify*

$$\mathbb{E}\left[\min_{0\leq i\leq k-1} \|\nabla f(\tilde{y}_i)\|^2\right] \leq \frac{4}{\gamma}\frac{\Delta_f}{k}.$$

*Proof.* From (6.30), we have

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds \leq f(x_0) - f^* + M_t.$$

Also, we have

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_s)\|^2 ds = \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s^-})\|^2 dN_s + \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s^-})\|^2 (ds - dN_s), \tag{6.31}$$

where $\int_0^t \frac{\gamma}{4} \|\nabla f(x_{s^-})\|^2 (ds - dN_s)$ is a centered martingale. Then, there exists a centered martingale $U_t$ such that

$$\int_0^t \frac{\gamma}{4} \|\nabla f(x_{s^-})\|^2 dN_s \leq \Delta_f + U_t.$$

We use the following stopping theorem.

**Theorem 6.B.1** (Hermant et al. (2025a), Theorem 6)**.** *Let $(\varphi_t)_{t\in\mathbb{R}_+}$ be a nonnegative process with cadlag trajectories, such that it verifies*

$$\varphi_t \leq K_0 + M_t,$$

*for some positive random variable $K_0$, some martingale $(M_t)_{t\geq 0}$ with $M_0 = 0$. Then, for an almost surely finite stopping time $\tau$, one has*

$$\mathbb{E}\left[\varphi_\tau\right] \leq \mathbb{E}\left[K_0\right].$$

Applying Theorem 6.B.1 with $\tau = T_k$, $\varphi_t = \int_0^t \frac{\gamma}{4} \|\nabla f(x_{s^-})\|^2 dN_s$, $K_0 = \Delta_f$ and $M_t = U_t$, we obtain

$$\mathbb{E}\left[\int_0^{T_k} \|\nabla f(x_{s^-})\|^2 dN_s\right] \leq \frac{4}{\gamma}\Delta_f.$$

Finally, we have

$$\int_0^{T_k} \|\nabla f(x_{s^-})\|^2 dN_s = \sum_{i=1}^k \left\|\nabla f(x_{T_i^-})\right\|^2 = \sum_{i=1}^k \|\nabla f(\tilde{y}_{i-1})\|^2 \geq k \min_{0\leq i\leq k-1} \|\nabla f(\tilde{y}_i)\|^2,$$

thus the result. □

## 6.C Proof of the Lemmas stated in Section 6.5

We prove formally all the Lemmas used in Section 6.5 to prove Theorem 6.4.1.

### 6.C.1 Proof of Lemma 6.5.1

The proof follows similar steps as for Lemma 1 in Okamura et al. (2024). We show the following inequalities

$$\begin{aligned}
&\left\| \int_0^t w_t(s) \nabla f(x_{s^-}) dN_s - \nabla f(\overline{x}_t) \right\| \\
&\leq \frac{L_2}{2} \int_0^t w_t(s) \left\| x_{s^-} - \overline{x}_t \right\|^2 dN_s \qquad (6.32) \\
&= \frac{L_2}{2} \int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) \left\| x_{\tau^-} - x_{\sigma^-} \right\|^2 dN_\sigma dN_\tau \qquad (6.33) \\
&\leq L_2 \int_0^t \eta^2 \left\| z_s - x_s \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (\tau - \sigma) dN_\sigma dN_\tau \right) ds \\
&+ L_2 \int_0^t \gamma^2 \left\| \nabla f(x_{s^-}) \right\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau) (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s. \qquad (6.34)
\end{aligned}$$

**Inequality** (6.32)**.** Using Taylor expansion of the gradient, we have

$$\begin{aligned}
&\nabla f(x_{s^-}) - \nabla f(\overline{x}_t) \\
&= \int_0^1 \nabla^2 f((1-\sigma)\overline{x}_t + \sigma x_{s^-})(x_{s^-} - \overline{x}_t) d\sigma \\
&= \nabla^2 f(\overline{x}_t)(x_{s^-} - \overline{x}_t) + \int_0^1 (\nabla^2 f((1-\sigma)\overline{x}_t + \sigma x_{s^-}) - \nabla^2 f(\overline{x}_t))(x_{s^-} - \overline{x}_t) d\sigma.
\end{aligned}$$

Multiply both sides by $w_t(s)$ and integrate with respect to $dN(\cdot)$ on $s \in [0, t]$

$$\begin{aligned}
&\int_0^t w_t(s)(\nabla f(x_{s^-}) - \nabla f(\overline{x}_t)) dN_s \\
&= \int_0^t w_t(s) \nabla^2 f(\overline{x}_t)(x_{s^-} - \overline{x}_t) dN_s \qquad (6.35) \\
&+ \int_0^t w_t(s) \int_0^1 (\nabla^2 f((1-\sigma)\overline{x}_t + \sigma x_{s^-}) - \nabla^2 f(\overline{x}_t))(x_{s^-} - \overline{x}_t) d\sigma dN_s.
\end{aligned}$$

Because $\int_0^t w_t(s) dN_s = 1$, we have

$$\int_0^t w_t(s)(\nabla f(x_{s^-}) - \nabla f(\overline{x}_t)) dN_s = \int_0^t w_t(s) \nabla f(x_{s^-}) dN_s - \nabla f(\overline{x}_t),$$

and because $\overline{x}_t = \int_0^t w_t(s) x_{s^-} dN_s$, we get

$$\int_0^t w_t(s) \nabla^2 f(\overline{x}_t)(x_{s^-} - \overline{x}_t) dN_s = 0.$$

Also,

$$\begin{aligned}
&\left\| \int_0^t w_t(s) \int_0^1 (\nabla^2 f((1-\sigma)\overline{x}_t + \sigma x_{s^-}) - \nabla^2 f(\overline{x}_t))(x_{s^-} - \overline{x}_t) d\sigma dN_s \right\| \\
&\leq \int_0^t w_t(s) \int_0^1 \left\| \nabla^2 f((1-\sigma)\overline{x}_t + \sigma x_{s^-}) - \nabla^2 f(\overline{x}_t) \right\|_2 \|x_{s^-} - \overline{x}_t\| \, d\sigma dN_s \\
&\leq \int_0^t w_t(s) \int_0^1 \sigma L_2 \|x_{s^-} - \overline{x}_t\|^2 \, d\sigma dN_s \quad \text{(Assumption 6.2.2)} \\
&= \frac{L_2}{2} \int_0^t w_t(s) \|x_{s^-} - \overline{x}_t\|^2 \, dN_s.
\end{aligned}$$

Finally, taking the norm, (6.35) becomes

$$\left\| \int_0^t w_t(s) \nabla f(x_{s^-}) dN_s - \nabla f(\overline{x}_t) \right\| \leq \frac{L_2}{2} \int_0^t w_t(s) \|x_{s^-} - \overline{x}_t\|^2 \, dN_s.$$

**Equality** (6.33)**.** We have

$$\begin{aligned}
\left\| \int_0^t w_t(\sigma) x_{\sigma^-} dN_\sigma \right\|^2 &= \left\langle \int_0^t w_t(\sigma) x_{\sigma^-} dN_\sigma, \int_0^t w_t(\tau) x_{\tau^-} dN_\tau \right\rangle \\
&= \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \langle x_{\sigma^-}, x_{\tau^-} \rangle \, dN_\sigma dN_\tau. \qquad (6.36)
\end{aligned}$$

Also

$$\begin{aligned}
&\int \int_{0 \leq \sigma \leq \tau \leq t} w_t(\sigma) w_t(\tau) \|x_{\tau^-} - x_{\sigma^-}\|^2 \, dN_\sigma dN_\tau \\
&= \frac{1}{2} \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \|x_{\tau^-} - x_{\sigma^-}\|^2 \, dN_\sigma dN_\tau \\
&= \int_0^t w_t(\sigma) \|x_{\sigma^-}\|^2 \, dN_\sigma - \int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \langle x_{\sigma^-}, x_{\tau^-} \rangle \, dN_\sigma dN_\tau
\end{aligned}$$

where the last equality is obtained by developing the squared norm, and because

$$\int \int_{[0,t]^2} w_t(\sigma) w_t(\tau) \|x_{\tau^-}\|^2 \, dN_\sigma dN_\tau = \int_0^t w_t(\tau) \|x_{\tau^-}\|^2 \, dN_\tau \int_0^t w_t(\sigma) dN_\sigma = \int_0^t w_t(\tau) \|x_{\tau^-}\|^2 \, dN_\tau.$$

Then,

$$
\begin{aligned}
&\int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - \int\int_{[0,t]^2} w_t(\sigma) w_t(\tau) \left\langle x_{\sigma^-}, x_{\tau^-} \right\rangle dN_\sigma dN_\tau \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - \left\| \int_0^t w_t(\sigma) x_{\sigma^-} dN_\sigma \right\|^2, \quad \text{using (6.36)} \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - \left\|\overline{x}_t\right\|^2 \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - 2\left\|\overline{x}_t\right\|^2 + \left\|\overline{x}_t\right\|^2 \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - 2\left\langle \int_0^t w_t(\sigma) x_{\sigma^-} dN_\sigma, \underbrace{\int_0^t w_t(\sigma) x_{\sigma^-} dN_\sigma}_{=\overline{x}_t} \right\rangle \\
&+ \left\|\overline{x}_t\right\|^2 \underbrace{\int_0^t w_t(\sigma) dN_\sigma}_{=1} \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-}\right\|^2 dN_\sigma - 2\int_0^t w_t(\sigma) \left\langle x_{\sigma^-}, \overline{x}_t \right\rangle dN_\sigma + \int_0^t w_t(\sigma) \left\|\overline{x}_t\right\|^2 dN_\sigma \\
&= \int_0^t w_t(\sigma) \left\|x_{\sigma^-} - \overline{x}_t\right\|^2 dN_\sigma.
\end{aligned}
$$

**Inequality** (6.34)**.** By definition of $(x_t)_{t\geq 0}$, for all $\sigma \in [0,t]$ we have

$$
x_t = x_{\sigma^-} + \int_\sigma^t \eta(z_s - x_s) ds - \int_{[\sigma,t]} \gamma \nabla f(x_{s^-}) dN_s
$$

Taking the left-limit at $t$, it becomes

$$
x_{t^-} = x_{\sigma^-} + \int_\sigma^t \eta(z_s - x_s) ds - \int_{[\sigma,t)} \gamma \nabla f(x_{s^-}) dN_s
$$

For $0 \leq \sigma \leq \tau \leq t$, using $\left\|a-b\right\|^2 \leq 2\left\|a\right\|^2 + 2\left\|b\right\|^2$ for $a, b \in \mathbb{R}^d$ and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\left\|x_{\tau^-} - x_{\sigma^-}\right\|^2 &= \left\| \int_\sigma^\tau \eta(z_s - x_s) ds - \int_{[\sigma,\tau)} \gamma \nabla f(x_{s^-}) dN_s \right\|^2 \\
&\leq 2\left\| \int_\sigma^\tau \eta(z_s - x_s) ds \right\|^2 + 2\left\| \int_{[\sigma,\tau)} \gamma \nabla f(x_{s^-}) dN_s \right\|^2 \\
&\leq 2\left( \int_\sigma^\tau \left\|\eta(z_s - x_s)\right\| ds \right)^2 + 2\left( \int_{[\sigma,\tau)} \left\|\gamma \nabla f(x_{s^-})\right\| dN_s \right)^2 \\
&\leq 2\left( \int_\sigma^\tau \left\|\eta(z_s - x_s)\right\|^2 ds \right)\left( \int_\sigma^\tau ds \right) + 2\left( \int_{[\sigma,\tau)} dN_s \right)\left( \int_{[\sigma,\tau)} \left\|\gamma \nabla f(x_{s^-})\right\|^2 dN_s \right) \\
&\leq 2(\tau - \sigma)\left( \int_\sigma^\tau \eta^2 \left\|(z_s - x_s)\right\|^2 ds \right) + 2(N_{\tau^-} - N_{\sigma^-}) \int_\sigma^\tau \gamma^2 \left\|\nabla f(x_{s^-})\right\|^2 dN_s.
\end{aligned}
\tag{6.37}
$$

In the last inequality, we used $\int_{[\sigma,\tau)} \|\gamma\nabla f(x_{s^-})\|^2 dN_s \leq \int_{[\sigma,\tau]} \|\gamma\nabla f(x_{s^-})\|^2 dN_s$, recalling our notation $\int_\sigma^\tau dN_s = \int_{[\sigma,\tau]}$. Then, we have

$$
\begin{aligned}
&\int\int_{0\leq\sigma\leq\tau\leq t} w_t(\sigma)w_t(\tau)\|x_{\tau^-}-x_{\sigma^-}\|^2 dN_\sigma dN_\tau \\
&\overset{(6.37)}{\leq} \int\int_{0\leq\sigma\leq\tau\leq t} w_t(\sigma)w_t(\tau)2(\tau-\sigma)\left(\int_\sigma^\tau \eta^2\|z_s-x_s\|^2 ds\right) dN_\sigma dN_\tau \\
&+\int\int_{0\leq\sigma\leq\tau\leq t} w_t(\sigma)w_t(\tau)2(N_{\tau^-}-N_{\sigma^-})\left(\int_\sigma^\tau \gamma^2\|\nabla f(x_{s^-})\|^2 dN_s\right) dN_\sigma dN_\tau \\
&=\int\int_{0\leq\sigma\leq s\leq\tau\leq t} w_t(\sigma)w_t(\tau)2(\tau-\sigma)\eta^2\|z_s-x_s\|^2 ds dN_\sigma dN_\tau \\
&+\int\int_{0\leq\sigma\leq s\leq\tau\leq t} w_t(\sigma)w_t(\tau)2(N_{\tau^-}-N_{\sigma^-})\gamma^2\|\nabla f(x_{s^-})\|^2 dN_s dN_\sigma dN_\tau \\
&=2\int_0^t \eta^2\|z_s-x_s\|^2\left(\int_s^t\int_0^s w_t(\sigma)w_t(\tau)(\tau-\sigma)dN_\sigma dN_\tau\right) ds \\
&+2\int_0^t \gamma^2\|\nabla f(x_{s^-})\|^2\left(\int_s^t\int_0^s w_t(\sigma)w_t(\tau)(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau\right) dN_s.
\end{aligned}
$$

### 6.C.2 Proof of Lemma 6.5.2

We set $\tilde{w}(s) = \alpha e^{\alpha s}$, such that $s \mapsto \tilde{w}(s)$ is differentiable, satisfying

$$\frac{d\tilde{w}(s)}{ds} = \alpha\tilde{w}(s).$$

Summing the two lines of (CNE) and using $\eta + \eta' = \alpha$, we get

$$dx_s - dz_s = \alpha(z_s - x_s)ds - (\gamma-\gamma')\nabla f(x_{s^-})dN_s. \tag{6.38}$$

We multiply each side of (6.38) by $\tilde{w}(s)$ and use $d\tilde{w}(s) = \alpha\tilde{w}(s)ds$ to get

$$
\begin{aligned}
\tilde{w}(s)(dx_s - dz_s) &= \tilde{\alpha}w(s)(z_s-x_s)ds - \tilde{w}(s)(\gamma-\gamma')\nabla f(x_{s^-})dN_s \\
&= (z_s-x_s)d\tilde{w}(s) - \tilde{w}(s)(\gamma-\gamma')\nabla f(x_{s^-})dN_s.
\end{aligned}
$$

By rearranging, we obtain

$$\tilde{w}(s)(dx_s - dz_s) + (x_s - z_s)d\tilde{w}(s) = -\tilde{w}(s)(\gamma-\gamma')\nabla f(x_{s^-})dN_s. \tag{6.39}$$

We write (6.39) in an integral form, such that

$$\int_0^t \tilde{w}(s)(dx_s - dz_s) + \int_0^t (x_s - z_s)d\tilde{w}(s) = -(\gamma-\gamma')\int_0^t \tilde{w}(s)\nabla f(x_{s^-})dN_s \tag{6.40}$$

We use the following integration by part formula for stochastic integrals, see a precise statement in (Protter, 2012, p.83).

**Corollary 6.C.1** (Protter (2012))**.** *Let $X$, $Y$ be semimartingales with at least one of $X$ or $Y$ continuous. Then*

$$X_tY_t = X_0Y_0 + \int_0^t X_{s^-}dY_s + \int_0^t Y_{s^-}dX_s.$$

We apply Corollary 6.C.1 to $X_t = x_t - z_t$ and $Y_t = \tilde{w}(t)$, such that $X_0Y_0 = 0$, and so

$$\int_0^t \tilde{w}(s)(dx_s - dz_s) + \int_0^t (x_s - z_s)d\tilde{w}(s) = \tilde{w}(t)(x_t - z_t). \tag{6.41}$$

We combine (6.40) with (6.41), and multiply both sides by $\left(\int_0^t \alpha e^{\alpha s} dN_s\right)^{-1}$ to conclude. Note that we can use Corollary 6.C.1 as $t \mapsto \tilde{w}(t)$ is deterministic and continuous, and $x_t - z_t$ is a semimartingale as a difference of semimartingales. To see that $(x_t)_{t\geq 0}$ is a semi martingale, remark that it writes

$$\begin{aligned} x_t &= x_0 + \eta \int_0^t (z_s - x_s) ds - \gamma \int_0^t \nabla f(x_{s^-}) dN_s \\ &= x_0 + \underbrace{\int_0^t (\eta(z_s - x_s) - \gamma \nabla f(x_s)) ds}_{\text{adapted + continuous + bounded variation}} \underbrace{- \gamma \int_0^t \nabla f(x_{s^-})(dN_s - ds)}_{\text{martingale}}, \end{aligned}$$

which is a semimartingale by definition. The same argument holds for $(z_t)_{t\geq 0}$.

### 6.C.3 Proof of Lemma 6.5.3

From Lemma 6.5.1, we have

$$\begin{aligned} \|\nabla f(\overline{x}_t)\| &\leq \left\| \int_0^t w_t(s) \nabla f(x_{s^-}) dN_s \right\| \\ &\quad + L_2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ &\quad + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s^-})\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s \\ &= \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| + L_2 \int_0^t \eta^2 \|z_s - x_s\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(\tau - \sigma) dN_\sigma dN_\tau \right) ds \\ &\quad + L_2 \int_0^t \gamma^2 \|\nabla f(x_{s^-})\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s, \end{aligned}$$

where the equality is because of Lemma 6.5.2. Multiplying both sides by $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2$ and integrating on $[0, T_n]$ with respect to $dN_t$, we get

$$\begin{aligned} &\int_0^{T_n} \Lambda_t \|\nabla f(\overline{x}_t)\| dN_t \leq \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &+ L_2 \int_0^{T_n} \Lambda_t \int_0^t \eta^2 \|z_s - x_s\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(\tau - \sigma) dN_\sigma dN_\tau \right) ds dN_t \\ &+ L_2 \int_0^{T_n} \Lambda_t \int_0^t \gamma^2 \|\nabla f(x_{s^-})\|^2 \left( \int_s^t \int_0^s w_t(\sigma) w_t(\tau)(N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s dN_t \\ &\overset{\text{(i)}}{=} \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &+ \alpha^2 L_2 \int_0^{T_n} \int_0^t \eta^2 \|z_s - x_s\|^2 \left( \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (\tau - \sigma) dN_\sigma dN_\tau \right) ds dN_t \\ &+ \alpha^2 L_2 \int_0^{T_n} \int_0^t \gamma^2 \|\nabla f(x_{s^-})\|^2 \left( \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau \right) dN_s dN_t \\ &\overset{\text{(ii)}}{=} \int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t \\ &+ \alpha^2 L_2 \int_0^{T_n} \eta^2 \|z_s - x_s\|^2 \left( \int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (\tau - \sigma) dN_\sigma dN_\tau dN_t \right) ds \\ &+ \alpha^2 L_2 \int_0^{T_n} \gamma^2 \|\nabla f(x_{s^-})\|^2 \left( \int_s^{T_n} \int_s^t \int_0^s e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t \right) dN_s. \end{aligned}$$

Equality (i) holds as $\Lambda_t = \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 = \left(\int_0^t \alpha e^{\alpha s} dN_s\right)^2 e^{-2\alpha t}$, such that

$$\Lambda_t w_t(\sigma) w_t(\tau) = \left(\int_0^t \alpha e^{\alpha s} dN_s\right)^2 e^{-2\alpha t} \frac{\alpha e^{\alpha\sigma}}{\int_0^t \alpha e^{\alpha s} dN_s} \frac{\alpha e^{\alpha\tau}}{\int_0^t \alpha e^{\alpha s} dN_s} = \alpha^2 e^{-2\alpha t} e^{\alpha\sigma} e^{\alpha\tau}.$$

Equality (ii) holds by Tonelli, enabling to swap the integration order of $s$ and $t$. We conclude by taking the expectation.

### 6.C.4 Proof of Lemma 6.5.4

Recalling $w_t(t) = \frac{\alpha e^{\alpha t}}{\int_0^t \alpha e^{\alpha s} dN_s}$, one has

$$\begin{aligned}\left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| &= \frac{\alpha e^{\alpha t}}{(\gamma' - \gamma)} \frac{1}{\int_0^t \alpha e^{\alpha s} dN_s} \|z_t - x_t\| \\ &= \frac{\alpha}{(\gamma' - \gamma)} \frac{1}{\int_0^t \alpha e^{\alpha(s-t)} dN_s} \|z_t - x_t\|\end{aligned}$$

So, using the notation $\Lambda_t := \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2$, we have

$$\begin{aligned}\mathbb{E}\left[\int_0^{T_n} \Lambda_t \left\| \frac{w_t(t)(z_t - x_t)}{\gamma' - \gamma} \right\| dN_t\right] &= \mathbb{E}\left[\int_0^{T_n} \frac{\alpha}{\gamma' - \gamma} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right) \|z_t - x_t\| \, dN_t\right] \\ &\leq \frac{\alpha}{\gamma' - \gamma} \sqrt{\mathbb{E}\left[\int_0^{T_n} \Lambda_t dN_t\right]} \sqrt{\mathbb{E}\left[\int_0^{T_n} \|z_t - x_t\|^2 \, dN_t\right]},\end{aligned} \tag{6.42}$$

where we used the Cauchy-Schwarz inequality. We have

$$\begin{aligned}\int_0^{T_n} \|z_t - x_t\|^2 \, dN_t &= \int_0^{T_n} \|z_{t^-} - x_{t^-} - (\gamma' - \gamma)\nabla f(x_{t^-})\|^2 \, dN_t \\ &\leq 2\int_0^{T_n} \|z_{t^-} - x_{t^-}\|^2 \, dN_t + 2(\gamma' - \gamma)^2 \int_0^{T_n} \|\nabla f(x_{t^-})\|^2 \, dN_t\end{aligned}$$

Then, from (6.9), we have

$$\mathbb{E}\left[\int_0^{T_n} \|\nabla f(x_{t^-})\|^2 \, dN_t\right] \leq \frac{4}{\gamma} \Delta_f, \tag{6.43}$$

and from (6.10), we have

$$\mathbb{E}\left[\int_0^{T_n} \|z_{t^-} - x_{t^-}\|^2 \, dN_t\right] \leq \frac{1}{\eta + \eta'} \Delta_f = \frac{1}{\alpha} \Delta_f. \tag{6.44}$$

Combining (6.43) and (6.44), we obtain

$$\begin{aligned}\sqrt{\mathbb{E}\left[\int_0^{T_n} \|z_t - x_t\|^2 \, dN_t\right]} &\leq \sqrt{2\Delta_f} \sqrt{\frac{1}{\alpha} + \frac{4(\gamma' - \gamma)^2}{\gamma}} \\ &\leq \sqrt{2\alpha^{-1}\Delta_f} \sqrt{1 + \frac{4(\gamma' - \gamma)^2}{\gamma}},\end{aligned} \tag{6.45}$$

the last inequality using $\alpha^{-1} \geq 1$. Note that we assume $\gamma' = \gamma + \sqrt{\gamma/2}$, such that $\frac{(\gamma' - \gamma)^2}{\gamma} = \frac{1}{2}$, and $\sqrt{1 + \frac{4(\gamma' - \gamma)^2}{\gamma}} = \sqrt{1 + 2} = \sqrt{3}$. Combining this with (6.42), Lemma 6.E.4-(i) and (6.45), we deduce

$$\mathbb{E}\left[\int_0^{T_n}\Lambda_t\left\|\frac{w_t(t)(z_{t^-}-x_{t^-})}{\gamma'-\gamma}\right\|dN_t\right]\le\frac{\alpha}{\gamma'-\gamma}\sqrt{3\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)n}\sqrt{2\frac{\Delta_f}{\alpha}}$$
$$=\sqrt{2\alpha n}\sqrt{\frac{6\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)}{\gamma}}\sqrt{\Delta_f}$$
$$=\sqrt{\alpha n}\sqrt{\frac{A_n}{\gamma}}\sqrt{\Delta_f}, \tag{6.46}$$

where we used $\gamma'=\gamma+\sqrt{\frac{\gamma}{2}}$ and defined $A_n=12\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)$.

### 6.C.5 Proof of Lemma 6.5.5

We prove each statement separately.

**Equation** (6.17) We have

$$\int_0^{T_n}\eta^2\left\|z_s-x_s\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)ds$$
$$=\sum_{i=0}^{n-1}\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)ds,$$

where we used that integration with respect to $ds$ on a singleton is zero.

Then, noting $U_i:=\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2ds$, we have

$$\sum_{i=0}^{n-1}\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)ds$$
$$\overset{(i)}{=}\sum_{i=0}^{n-1}\left(\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2ds\right)\left(\int_{T_{i+1}}^{T_n}\int_{T_{i+1}}^t\int_0^{T_i} e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)$$
$$\overset{(ii)}{=}\sum_{i=1}^{n-1}\left(\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2ds\right)\left(\int_{T_{i+1}}^{T_n}\int_{T_{i+1}}^t\int_0^{T_i} e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)$$
$$=\sum_{i=1}^{n-1}U_i\left(\int_{T_{i+1}}^{T_n}\int_{T_{i+1}}^t\int_0^{T_i} e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right)$$
$$=\sum_{i=1}^{n-1}\left(\int_{T_{i+1}}^{T_n}\int_{T_{i+1}}^t\int_0^{T_i} e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)U_i dN_\sigma dN_\tau dN_t\right), \tag{6.47}$$

where (i) uses the fact that $\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t$ is constant for $s\in(T_i,T_{i+1})$, as no new jump occurs on this interval. (ii) uses that $\int_0^{T_0}\cdot\,dN_t=\int_0^0\cdot\,dN_t=0$, as almost surely no jumps occur at $t=0$. Because we integrate over $[0,T_n]$ with Poisson integrals, we can write (6.47) as a sum

$$(6.47)=\sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\sum_{j=1}^{i}e^{\alpha(T_k-T_l)}e^{\alpha(T_j-T_l)}(T_k-T_j)U_i. \tag{6.48}$$

We multiply each side of (6.48) by $\mathbb{1}_{\mathcal{A}_n}$, and take the expectation

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.48)\right] &= \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\sum_{j=1}^{i}e^{\alpha(T_k-T_l)}e^{\alpha(T_j-T_l)}(T_k-T_j)U_i\right]\\
&= \sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}\sum_{j=1}^{i}e^{\alpha(T_j-T_{i+1})}(T_k-T_j)U_i\right]\\
&\overset{(i)}{\leq} \sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_k-T_j)U_i\right]\\
&= \sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}U_i\mathbb{1}_{\mathcal{A}_n}\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_k-T_{i+1})\right]\\
&+ \sum_{i=K}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}U_i\mathbb{1}_{\mathcal{A}_n}\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_{i+1}-T_j)\right]\\
&+ \sum_{i=1}^{K-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}U_i\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_{i+1}-T_j)\right].
\end{aligned}$$

(i) uses that $\sum_{j=1}^{i}a_j \leq \sum_{j=1}^{i+1}a_j$, as long as $a_j \geq 0$ for any $j \in \{1,\dots,i+1\}$. The last line uses $\mathbb{1}_{\mathcal{A}_n} \leq 1$. Now, we use Proposition 6.D.4 to deal with the third term, and Theorem 6.5.1 to deal with the first and second term, such that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.48)\right] &\leq \sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}(T_k-T_{i+1})U_i\mathcal{C}_n\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}\right]\right]\\
&+ \sum_{i=K}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}U_i\mathcal{C}_n\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_{i+1}-T_j)\right]\right]\\
&+ \sum_{i=1}^{K-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}U_ic_{upp}K^2\right]\\
&\leq \mathcal{C}_n\sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}(T_k-T_{i+1})U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}\right]\\
&+ \mathcal{C}_n\sum_{i=K}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_{i+1}-T_j)\right]\\
&+ \sum_{i=1}^{K-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_{i+1}-T_l)}U_i\right]c_{upp}K^2\\
&\overset{(i)}{\leq} \mathcal{C}_n\sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}(T_k-T_{i+1})U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}\right]\\
&+ \mathcal{C}_n\sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l}\mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k-T_l)}U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1}e^{\alpha(T_j-T_{i+1})}(T_{i+1}-T_j)\right].
\end{aligned} \tag{6.49}$$

(i) uses that for $i \geq K$ and by Lemma 6.D.4, we have

$$\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right] \geq \frac{e^{-1}}{32}\min\left\{(i+1)^2, \frac{1}{\alpha^2}\right\} \geq \frac{e^{-1}}{32}\min\left\{K^2, \frac{1}{\alpha^2}\right\} \geq \frac{c_{inf}^2 e^{-1}}{32}K^2,$$

the last inequality using the assumption $\alpha^{-1} \geq c_{inf}K$. So, we deduce

$$\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right] \geq \frac{c_{inf}^2 e^{-1}}{32 c_{upp}} c_{upp} K^2 \geq \mathcal{C}_n^{-1} c_{upp} K^2,$$

which implies

$$c_{upp}K^2 \leq \mathcal{C}_n \mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right].$$

Then, $U_i$ depends on $\{T_k\}_{k\in\{1,\dots,i\}}$, as it depends on $x_s$ and $z_s$ for $s \in (T_i, T_{i+1})$, which depend on $\{T_k\}_{k\in\{1,\dots,i\}}$. Also, $U_i$ depends on the length of the interval $(T_i, T_{i+1})$, such that $U_i$ also depends on $T_{i+1}$. So, we have that $U_i$ only depends on $\{T_k\}_{k\in\{1,\dots,i+1\}}$. Because the increments of $\{T_k\}_{k\in\mathbb{N}}$ are independent, because $k, l \geq i+1$, we have $T_{i+1} - T_l \perp\!\!\!\perp U_i$ and $T_k - T_l \perp\!\!\!\perp U_i$. So, (6.49) becomes

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n} \cdot (6.48)\right] \leq\ & \mathcal{C}_n \sum_{i=1}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}(T_k - T_{i+1})\right]\mathbb{E}\left[U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}\right] \\
& + \mathcal{C}_n \sum_{i=K}^{n-1}\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}\right]\mathbb{E}\left[U_i\right]\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right] \\
=\ & \mathcal{C}_n \sum_{i=1}^{n-1}\mathbb{E}\left[U_i\right]\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}(T_k - T_{i+1})\right]\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}\right] \\
& + \mathcal{C}_n \sum_{i=K}^{n-1}\mathbb{E}\left[U_i\right]\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}\right]\mathbb{E}\left[\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right] \qquad (6.50) \\
\overset{(i)}{=}\ & \mathcal{C}_n \sum_{i=1}^{n-1}\mathbb{E}\left[U_i\right]\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}(T_k - T_{i+1})\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}\right] \\
& + \mathcal{C}_n \sum_{i=1}^{n-1}\mathbb{E}\left[U_i\right]\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} \mathbb{E}\left[e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_{i+1} - T_j)\right] \\
=\ & \mathcal{C}_n \sum_{i=1}^{n-1}\mathbb{E}\left[U_i\right]\mathbb{E}\left[\sum_{l=i+1}^{n}\sum_{k=i+1}^{l} e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_k - T_j)\right]
\end{aligned}$$

where in (i) we used the independence of the increments of the $\{T_k\}_{k\in\mathbb{N}}$. Finally, writing the sums as Poisson integrals, noting

$$\begin{aligned}
A_{i+1} &:= \sum_{l=i+1}^{n}\sum_{k=i+1}^{l} e^{\alpha(T_{i+1}-T_l)}e^{\alpha(T_k - T_l)}\sum_{j=1}^{i+1} e^{\alpha(T_j - T_{i+1})}(T_k - T_j) \\
&= \int_{[T_{i+1},T_n]}\int_{[T_{i+1},t]} e^{\alpha(T_{i+1}-t)}e^{\alpha(\tau - t)}\int_{[0,T_{i+1}]} e^{\alpha(\sigma - T_{i+1})}(\tau - \sigma)dN_\sigma dN_\tau dN_t,
\end{aligned}$$

and recalling $U_i := \int_{(T_i,T_{i+1})} \eta^2 \left\|z_s - x_s\right\|^2 ds$, (6.50) becomes

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.48)\right] &\leq \mathcal{C}_n\sum_{i=1}^{n-1}\mathbb{E}\left[\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2ds\right]\mathbb{E}\left[A_{i+1}\right]\\ &\leq \mathcal{C}_n\max_{i\in\{1,\dots,n-1\}}\mathbb{E}\left[A_{i+1}\right]\sum_{i=1}^{n-1}\mathbb{E}\left[\int_{(T_i,T_{i+1})}\eta^2\left\|z_s-x_s\right\|^2ds\right]\\ &\leq \mathcal{C}_n\max_{i\in\{1,\dots,n\}}\mathbb{E}\left[A_i\right]\mathbb{E}\left[\int_0^{T_n}\eta^2\left\|z_s-x_s\right\|^2ds\right].\end{aligned}$$

**Equation** (6.18)

$$\begin{aligned}&\int_0^{T_n}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\\ &=\int_0^{T_n}\left(\int_s^{T_n}\int_s^t e^{\alpha(\tau-t)}e^{\alpha(s-t)}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2\int_0^s e^{\alpha(\sigma-s)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\end{aligned}\tag{6.51}$$

Because we integrate over $[0,T_n]$, (6.51) can be written as a sum

$$(6.51)=\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\sum_{j=1}^i e^{\alpha(T_j-T_i)}(k-j).\tag{6.52}$$

We used that $N_{T_k^-}-N_{T_j^-}=(k-1)-(j-1)$. Multiplying by $\mathbb{1}_{\mathcal{A}_n}$ and taking expectation, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.51)\right]&=\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\sum_{j=1}^i e^{\alpha(T_j-T_i)}(k-j)\right]\\ &=\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\sum_{j=1}^i e^{\alpha(T_j-T_i)}(k-j)\right]\\ &=\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2(k-i)\mathbb{1}_{\mathcal{A}_n}\sum_{j=1}^i e^{\alpha(T_j-T_i)}\right]\\ &+\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\mathbb{1}_{\mathcal{A}_n}\sum_{j=1}^i e^{\alpha(T_j-T_i)}(i-j)\right]\end{aligned}$$

Now, Theorem 6.5.1 ensures that there exists a constant $\mathcal{C}_n$, such that we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.51)\right]&\leq\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\mathcal{C}_n(k-i)\mathbb{E}\left[\sum_{j=1}^i e^{\alpha(T_j-T_i)}\right]\right]\\ &+\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\mathcal{C}_n\mathbb{E}\left[\sum_{j=1}^i e^{\alpha(T_j-T_i)}(i-j)\right]\right]\\ &\leq\mathcal{C}_n\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\right](k-i)\mathbb{E}\left[\sum_{j=1}^i e^{\alpha(T_j-T_i)}\right]\\ &+\mathcal{C}_n\sum_{i=1}^n\sum_{l=i}^n\sum_{k=i}^l\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\left\|\nabla f(x_{T_i^-})\right\|^2\right]\mathbb{E}\left[\sum_{j=1}^i e^{\alpha(T_j-T_i)}(i-j)\right].\end{aligned}\tag{6.53}$$

Now, because the increments of $\{T_k\}_{k\in\mathbb{N}}$ are independent, as $l,k\geq i$, we have $T_k - T_l \perp\!\!\!\perp \left\|\nabla f(x_{T_i^-})\right\|$ and $T_i - T_l \perp\!\!\!\perp \left\|\nabla f(x_{T_i^-})\right\|$. So, (6.53) becomes

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.51)\right] &\leq \mathcal{C}_n\sum_{i=1}^{n}\sum_{l=i}^{n}\sum_{k=i}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\right]\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right](k-i)\mathbb{E}\left[\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}\right]\\
&+\mathcal{C}_n\sum_{i=1}^{n}\sum_{l=i}^{n}\sum_{k=i}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\gamma^2\right]\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\mathbb{E}\left[\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}(i-j)\right]\\
&=\mathcal{C}_n\sum_{i=1}^{n}\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\sum_{l=i}^{n}\sum_{k=i}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\right](k-i)\mathbb{E}\left[\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}\right] \qquad (6.54)\\
&+\mathcal{C}_n\sum_{i=1}^{n}\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\sum_{l=i}^{n}\sum_{k=i}^{l}\mathbb{E}\left[e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\right]\mathbb{E}\left[\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}(i-j)\right]\\
&\overset{(i)}{=}\mathcal{C}_n\sum_{i=1}^{n}\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\mathbb{E}\left[\sum_{l=i}^{n}\sum_{k=i}^{l}e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}(k-j)\right],
\end{aligned}$$

where (i) used the independence property of the increments of $\{T_k\}_{k\in\mathbb{N}^*}$. Rewriting (6.54) as Poisson integrals, we note

$$\begin{aligned}
B_i &:= \sum_{l=i}^{n}\sum_{k=i}^{l}e^{\alpha(T_k-T_l)}e^{\alpha(T_i-T_l)}\sum_{j=1}^{i}e^{\alpha(T_j-T_i)}(k-j)\\
&=\int_{T_i}^{T_n}\int_{T_i}^{t}e^{\alpha(\tau-t)}e^{\alpha(T_i-t)}\int_{0}^{T_i}(N_{\tau^-}-N_{\sigma^-})e^{\alpha(\sigma-T_i)}dN_\sigma dN_\tau dN_t,
\end{aligned}$$

such that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\cdot(6.51)\right] &\leq \mathcal{C}_n\sum_{i=1}^{n}\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\mathbb{E}\left[B_i\right]\\
&\leq \mathcal{C}_n\max_{i\in\{1,\dots,n\}}\mathbb{E}\left[B_i\right]\sum_{i=1}^{n}\gamma^2\mathbb{E}\left[\left\|\nabla f(x_{T_i^-})\right\|^2\right]\\
&=\mathcal{C}_n\max_{i\in\{1,\dots,n\}}\mathbb{E}\left[B_i\right]\mathbb{E}\left[\int_0^{T_n}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2dN_s\right]
\end{aligned}$$

### 6.C.6 Proof of Lemma 6.5.6

We recall

$$\begin{aligned}
A_k &:= \int_{[T_k,T_n]}\int_{[T_k,t]}\int_{[0,T_k]}e^{\alpha(T_k-t)}e^{\alpha(\tau-t)}e^{\alpha(\sigma-T_k)}(\tau-\sigma)dN_\sigma dN_\tau dN_t,\\
B_k &:= \int_{T_k}^{T_n}\int_{T_k}^{t}\int_{0}^{T_k}e^{\alpha(\tau-t)}e^{\alpha(T_k-t)}e^{\alpha(\sigma-T_k)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t.
\end{aligned}$$

We will split the proof in several sub lemmas. We first bound $\max_{k\in\{1,\dots,n\}}\mathbb{E}\left[B_k\right]$.

**Lemma 6.C.1.** *We have*

$$\max_{k\in\{1,\dots,n\}}\mathbb{E}\left[B_k\right]\leq\frac{(1+\alpha)^2(1+2\alpha)}{\alpha^4}$$

*Proof.* Let $k \leq n$.

$$\begin{aligned} B_k &= \int_{T_k}^{T_n} \int_{T_k}^{t} \int_{0}^{T_k} e^{\alpha(\tau - t)} e^{\alpha(\sigma - t)} (N_{\tau^-} - N_{\sigma^-}) dN_\sigma dN_\tau dN_t \\ &= \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} e^{\alpha(T_l - T_i)} e^{\alpha(T_j - T_i)} (l - j) \\ &= \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} e^{2\alpha(T_l - T_i)} e^{\alpha(T_j - T_l)} (l - j). \end{aligned}$$

We take expectation, use $T_l - T_i \perp\!\!\!\perp T_j - T_l$ and Lemma 6.E.3

$$\begin{aligned} \mathbb{E}[B_k] =& \mathbb{E}\left[ \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} e^{2\alpha(T_l - T_i)} e^{\alpha(T_j - T_l)} (l - j) \right] && (6.55) \\ =& \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} \mathbb{E}\left[ e^{2\alpha(T_l - T_i)} \right] \mathbb{E}\left[ e^{\alpha(T_j - T_l)} \right] (l - j) && (6.56) \\ =& \sum_{i=k}^{n} \sum_{j=1}^{k} \sum_{l=k}^{i} (1 + 2\alpha)^{l-i} (1 + \alpha)^{j-l} (l - j) \\ =& \sum_{\substack{i,j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq i \leq n}} (1 + 2\alpha)^{l-i} (1 + \alpha)^{j-l} (l - j) \\ =& \sum_{\substack{j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq n}} (1 + \alpha)^{j-l} (l - j) \sum_{i=l}^{n} (1 + 2\alpha)^{l-i} \end{aligned}$$

We have $\sum_{i=l}^{n} (1 + 2\alpha)^{l-i} \leq \sum_{i=l}^{+\infty} (1 + 2\alpha)^{l-i} = 1 + \frac{1}{2\alpha}$. We plug this inequality in the above, such that

$$\mathbb{E}[B_k] \leq \left(1 + \frac{1}{2\alpha}\right) \sum_{\substack{j,l \text{ s.t.} \\ 1 \leq j \leq k \leq l \leq n}} (1 + \alpha)^{j-l} (l - j) \tag{6.57}$$

Now, we do the change of variable $(j, l) \to (j, l - j) := (j, d)$, such that

$$\mathbb{E}[B_k] \leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} \sum_{j=\max\{1, k-d\}}^{\min\{k, n-d\}} (1 + \alpha)^{-d} d. \tag{6.58}$$

Indeed, for a fixed $k$, we both have $1 \leq j \leq k$, and $k \leq l \leq n \Rightarrow k - d \leq l - d \overset{(i)}{=} j \leq n - d$, where $(i)$ is due to the change of variable. Then, as $\sum_{j=\max\{1,k-d\}}^{\min\{k,n-d\}} (1 + \alpha)^{-d} d \leq \sum_{j=k-d}^{k} (1 + \alpha)^{-d} d$, we deduce

$$\begin{aligned} \mathbb{E}[B_k] &\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} \sum_{j=k-d}^{k} (1 + \alpha)^{-d} d \\ &\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{n-1} (1 + \alpha)^{-d} d(d + 1) \\ &\leq \left(1 + \frac{1}{2\alpha}\right) \sum_{d=0}^{+\infty} (1 + \alpha)^{-d} d(d + 1) \\ &= \left(1 + \frac{1}{2\alpha}\right) \frac{2}{\alpha^3} (1 + \alpha)^2 \\ &= \frac{(1 + \alpha)^2 (1 + 2\alpha)}{\alpha^4}. \end{aligned} \tag{6.59}$$

□

We now turn to $\max_{k\in\{1,\dots,n\}} \mathbb{E}[A_k]$. The computations are similar, thanks to the following elementary result.

**Lemma 6.C.2.** *If $T_n \sim \Gamma(n,1)$, then*

$$\mathbb{E}\left[T_n e^{-\alpha T_n}\right] = \frac{n}{(1+\alpha)^{n+1}}$$

*Proof.* Let $\varphi(\alpha) = \mathbb{E}\left[e^{-\alpha T_n}\right]$ the moment-generating function of $T_n$. As $T_n \sim \Gamma(n,1)$, we have $\phi(\alpha) = (1+\alpha)^{-n} \Rightarrow \phi'(\alpha) = -n(1+\alpha)^{-n-1}$. So

$$\phi(\alpha) = -\mathbb{E}\left[T_n e^{-\alpha T_n}\right] \Rightarrow \mathbb{E}\left[T_n e^{-\alpha T_n}\right] = \frac{n}{(1+\alpha)^{n+1}}.$$

□

**Lemma 6.C.3.** *We have*

$$\max_{k\in\{1,\dots,n\}} \mathbb{E}[A_k] \leq \frac{(1+\alpha)(1+2\alpha)}{\alpha^4}$$

*Proof.* The proof is almost identical to the one of Lemma 6.C.1. The only difference lies in (6.56), where $l-j$ is replaced by $T_l - T_j$. Then, a direct application of Lemma 6.C.2 gives $\mathbb{E}\left[T_n e^{-\alpha T_n}\right] = (1+\alpha)^{-1}\mathbb{E}\left[n e^{-\alpha T_n}\right]$ for some $T_n \sim \Gamma(n,1)$. □

## 6.C.7 Proof of Lemma 6.5.7

**1. Bounding Term 2.** Using Lemma 6.5.5 and Lemma 6.5.6, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\alpha^2 L_2 \int_0^{T_n} \eta^2 \left\|z_s - x_s\right\|^2 \left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right) ds\right] \\
&\leq \mathcal{C}_n L_2 \frac{(1+\alpha)(1+2\alpha)}{\alpha^2}\mathbb{E}\left[\int_0^{T_n} \eta^2 \left\|z_s - x_s\right\|^2 ds\right].
\end{aligned} \tag{6.60}$$

From (6.8), we have

$$\mathbb{E}\left[\int_0^{T_n} \left\|z_s - x_s\right\|^2 ds\right] \leq \frac{1}{\eta+\eta'}\Delta_f = \frac{1}{\alpha}\Delta_f. \tag{6.61}$$

So, combining (6.60) and (6.61), we deduce

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\alpha^2 L_2 \int_0^{T_n} \eta^2 \left\|z_s - x_s\right\|^2 \left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(\tau-\sigma)dN_\sigma dN_\tau dN_t\right) ds\right] \\
&\leq \mathcal{C}_n L_2 \eta^2 \frac{(1+\alpha)(1+2\alpha)}{\alpha^3}\Delta_f \\
&= \mathcal{C}_n L_2 \gamma \frac{(1+\alpha)(1+2\alpha)}{2\alpha^3}\Delta_f,
\end{aligned} \tag{6.62}$$

where the last equality uses $\eta = \sqrt{\frac{\gamma}{2}}$.

**2. Bounding Term 3.** Using Lemma 6.5.5 and Lemma 6.5.6, we have

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\alpha^2 L_2\int_0^{T_n}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\right]\\&\leq \mathcal{C}_n L_2\frac{(1+\alpha)^2(1+2\alpha)}{\alpha^2}\mathbb{E}\left[\int_0^{T_n}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2 dN_s\right]\end{aligned}\tag{6.63}$$

From (6.9), we have

$$\mathbb{E}\left[\int_0^{T_n}\left\|\nabla f(x_{s^-})\right\|^2 dN_s\right]\leq\frac{4}{\gamma}\Delta_f.\tag{6.64}$$

So, combining (6.63) and (6.64), we deduce

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\alpha^2 L_2\int_0^{T_n}\gamma^2\left\|\nabla f(x_{s^-})\right\|^2\left(\int_s^{T_n}\int_s^t\int_0^s e^{\alpha(\tau-t)}e^{\alpha(\sigma-t)}(N_{\tau^-}-N_{\sigma^-})dN_\sigma dN_\tau dN_t\right)dN_s\right]\\&\leq 4\mathcal{C}_n L_2\gamma\frac{(1+\alpha)^2(1+2\alpha)}{\alpha^2}\Delta_f.\end{aligned}\tag{6.65}$$

Combining (6.62) and (6.65), we obtain the desired result.

## 6.C.8 Proofs of Section 6.5.5

**Proof of Corollary 6.5.1** From (6.20)

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\min_{t\in\{T_1,\dots,T_n\}}\left\|\nabla f(\overline{x}_t)\right\|\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right]\\&=n^{3/7}\left(\sqrt{A_n L\Delta_f}+B_n\mathcal{C}_n L^{-1}L_2\Delta_f\right)\\&\leq 4\sqrt{3}B_n\mathcal{C}_n\left(\sqrt{L\Delta_f}+L^{-1}L_2\Delta_f\right)n^{3/7}\end{aligned}\tag{6.66}$$

where we use $\mathcal{C}_n\geq 1$ and $\sqrt{A_n}=2\sqrt{3}\sqrt{(1+\frac{3}{2}\alpha)(1+2\alpha)}\leq 2\sqrt{3}(1+2\alpha)\leq 2\sqrt{3}(1+2\alpha)(1+\alpha)\left(1+\alpha 8(1+\alpha)\right)=4\sqrt{3}B_n$. Using $\alpha=n^{-1/7}$, we have

$$B_n\mathcal{C}_n=2\sqrt{3}(1+2n^{-1/7})(1+n^{-1/7})\left(1+n^{-1/7}8(1+n^{-1/7})\right)32(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2},$$

with $C_n=c_{inf}n^{-1/7}K$.

**Proof of Corollary 6.5.2** From (6.20)

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\min_{t\in\{T_1,\dots,T_n\}}\left\|\nabla f(\overline{x}_t)\right\|\int_0^{T_n}\left(\int_0^t\alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right]\\&\leq n^{3/7}\left(\sqrt{A_n C_\alpha L^{1+a_1}L_2^{a_2}{\Delta_f}^{1+a_3}}+B_n {C_\alpha}^{-3}\mathcal{C}_n L^{-1-3a_1}L_2^{1-3a_2}{\Delta_f}^{1-3a_3}\right),\end{aligned}\tag{6.67}$$

with, $A_n:=12\left(1+\frac{3}{2}\alpha\right)(1+2\alpha)$ and $B_n:=\frac{1}{2}(1+\alpha)(1+2\alpha)\left(1+\alpha 8(1+\alpha)\right)$. We want to find $a_1,a_2,a_3$ such that the exponents in the left and right term are the same. Therefore, we solve

$$\begin{cases}\dfrac{1+a_1}{2}=-1-3a_1,\\[2mm]\dfrac{a_2}{2}=1-3a_2,\\[2mm]\dfrac{1+a_3}{2}=1-3a_3,\end{cases}\qquad\Longrightarrow\qquad(a_1,a_2,a_3)=\left(-\frac{3}{7},\frac{2}{7},\frac{1}{7}\right).$$

Then, (6.67) becomes

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_n}\min_{t\in\{T_1,\dots,T_n\}}\|\nabla f(\overline{x}_t)\|\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right]\\
&\leq n^{3/7}\left(\sqrt{A_n\left(1+\frac{3}{2}\right)C_\alpha}+B_n{C_\alpha}^{-3}\mathcal{C}_n\right)L^{\frac{2}{7}}L_2^{\frac{1}{7}}{\Delta_f}^{\frac{4}{7}}\\
&\overset{(i)}{=}\left(\sqrt{A_nC_\alpha}+B_n{C_\alpha}^{-3}\Big(32\frac{c_{upp}e}{c_{inf}^2}(4+4C_n+C_n^2)\Big)\right)L^{\frac{2}{7}}L_2^{\frac{1}{7}}{\Delta_f}^{\frac{4}{7}}n^{3/7} \qquad (6.68)\\
&\overset{(ii)}{\leq}(1+C_n+C_n^2/4)\left(\sqrt{A_nC_\alpha}+128B_n\frac{c_{upp}e}{c_{inf}^2}{C_\alpha}^{-3}\right)L^{\frac{2}{7}}L_2^{\frac{1}{7}}{\Delta_f}^{\frac{4}{7}}n^{3/7}\\
&\overset{(iii)}{\leq}(1+C_n+C_n^2/4)\left(\sqrt{12C_\alpha}+64\frac{c_{upp}e}{c_{inf}^2}{C_\alpha}^{-3}\right)L^{\frac{2}{7}}L_2^{\frac{1}{7}}{\Delta_f}^{\frac{4}{7}}n^{3/7}.
\end{aligned}
$$

In $(i)$ we use the definition of $\mathcal{C}_n$, namely $\mathcal{C}_n = 32(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2}$. In $(ii)$ we use $(4+4C_n+C_n^2) = 4(1+C_n+C_n^2/4)$ and $(1+C_n+C_n^2/4)\geq 1$. In $(iii)$, replace $A_n$ and $B_n$ by their values and we set $:= \max\{\sqrt{(1+\frac{3}{2}\alpha)(1+2\alpha)}, (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))\} = (1+\alpha)(1+2\alpha)(1+\alpha 8(1+\alpha))$. We then optimize on $C_\alpha$. Note that we keep implicit the dependence on $C_\alpha$ in , and do not consider it in the optimization of $C_\alpha$. Denoting $a=12$, $b=64\frac{c_{upp}e}{c_{inf}^2}$, we want to minimize

$$\sqrt{aC_\alpha}+b{C_\alpha}^{-3}.$$

The function $\phi : t\mapsto \sqrt{at}+bt^{-3}$ is derivable on $\mathbb{R}_+^*$, and $\phi'(t) = \frac{\sqrt{a}}{2\sqrt{t}} - \frac{3b}{t^4}$. It follows that $\phi$ reaches its minimum on $\mathbb{R}_+^*$ at the point

$$\phi'(t)=0 \Leftrightarrow t=\left(\frac{6b}{\sqrt{a}}\right)^{2/7}.$$

Then, the minimum of $\phi$, denoted $\phi_{\min}$, is

$$
\begin{aligned}
\phi_{\min} &= \phi\left(\left(\frac{6b}{\sqrt{a}}\right)^{2/7}\right)\\
&= \sqrt{a\left(\frac{6b}{\sqrt{a}}\right)^{2/7}}+\frac{b}{\left(\frac{6b}{\sqrt{a}}\right)^{6/7}}\\
&= a^{1/2}\cdot 6^{1/7}b^{1/7}a^{-1/14}+b\cdot a^{3/7}6^{-6/7}b^{-6/7}\\
&= 6^{1/7}a^{3/7}b^{1/7}+6^{-6/7}a^{3/7}b^{1/7}\\
&= \left(6^{1/7}+6^{-6/7}\right)a^{3/7}b^{1/7}\\
&= 7\,6^{-6/7}\,a^{3/7}b^{1/7}.
\end{aligned}
$$

Then, choosing $C_\alpha := \left(\frac{6b}{\sqrt{a}}\right)^{2/7}$, and replacing $a$ and $b$ by their values, we obtain

$$C_\alpha = \left(\frac{6b}{\sqrt{a}}\right)^{2/7} = \left(\frac{6\cdot 64\frac{c_{upp}e}{c_{inf}^2}}{\sqrt{12}}\right)^{2/7} = \left(\frac{\sqrt{3}\cdot 64 c_{upp}e}{c_{inf}^2}\right)^{2/7},$$

which in turn implies

$$\sqrt{A_nC_\alpha}+64\frac{c_{upp}e}{c_{inf}^2}{C_\alpha}^{-3} = 7\,6^{-6/7}\,12^{3/7}64^{1/7}\left(\frac{c_{upp}e}{c_{inf}^2}\right)^{1/7}.$$

Finally, we deduce that with the choice

$$\alpha = \left(\frac{\sqrt{3}\cdot 64 c_{upp} e}{c_{inf}^2}\right)^{2/7} \left(\frac{L_2^2 \Delta_f}{L^3 n}\right)^{1/7},$$

we obtain

$$\begin{aligned}
&\mathbb{E}\left[\min_{t\in\{T_1,\dots,T_n\}} \|\nabla f(\overline{x}_t)\| \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] \\
&\leq (1 + C_n + C_n^2/4) 7\, 6^{-6/7}\, 12^{3/7} 64^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2}\right)^{1/7} \left(\frac{c_{upp} e}{c_{inf}^2}\right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} {\Delta_f}^{\frac{4}{7}} n^{3/7} \qquad (6.69) \\
&\approx 9.2(1 + C_n + C_n^2/4) \left(\frac{c_{upp}}{c_{inf}^2}\right)^{1/7} L^{\frac{2}{7}} L_2^{\frac{1}{7}} {\Delta_f}^{\frac{4}{7}} n^{3/7}.
\end{aligned}$$

### 6.C.9 Proof of Lemma 6.5.8

We note

$$\Delta_n := \int_0^{T_n} \left(\int_0^t \alpha e^{\alpha(s-t)}\, dN_s\right)^2 dN_t = \sum_{i=1}^n \left(\sum_{j=1}^i \alpha e^{-\alpha(T_i - T_j)}\right)^2.$$

By Proposition 6.D.1, on the set $\mathcal{A}_n$, we have that for all $j \leq i \leq n$,

$$T_i - T_j \leq c_{upp}(i-j) + c_{upp} K.$$

Hence,

$$\mathbb{1}_{\mathcal{A}_n} \Delta_n \geq \mathbb{1}_{\mathcal{A}_n} \alpha^2 \sum_{i=1}^n \left(\sum_{j=1}^i e^{-c_{upp}\alpha(i-j+K)}\right)^2 = \mathbb{1}_{\mathcal{A}_n} \alpha^2 e^{-2c_{upp}\alpha K} \sum_{i=1}^n \left(\sum_{j=1}^i e^{-c_{upp}\alpha(i-j)}\right)^2.$$

Set $r := e^{-c_{upp}\alpha}$. Then

$$\sum_{j=1}^i e^{-c_{upp}\alpha(i-j)} = \sum_{\ell=0}^{i-1} r^\ell = \frac{1-r^i}{1-r}.$$

Therefore,

$$\mathbb{1}_{\mathcal{A}_n} \Delta_n \geq \mathbb{1}_{\mathcal{A}_n} \frac{\alpha^2}{(1-r)^2} e^{-2c_{upp}\alpha K} \sum_{i=1}^n (1-r^i)^2.$$

Using $1 - e^{-x} \leq x$ for all $x \geq 0$, we get

$$\frac{\alpha^2}{(1-r)^2} = \frac{\alpha^2}{(1-e^{-c_{upp}\alpha})^2} \geq \frac{\alpha^2}{(c_{upp}\alpha)^2} = \frac{1}{c_{upp}^2}.$$

Moreover,

$$e^{-2c_{upp}\alpha K} = e^{-2\frac{c_{upp}}{c_{inf}} C_n}.$$

Hence,

$$\mathbb{1}_{\mathcal{A}_n} \Delta_n \geq \mathbb{1}_{\mathcal{A}_n} \frac{1}{c_{upp}^2} e^{-2\frac{c_{upp}}{c_{inf}} C_n} \sum_{i=1}^n (1-r^i)^2.$$

Now,

$$\sum_{i=1}^n (1-r^i)^2 = n - 2\sum_{i=1}^n r^i + \sum_{i=1}^n r^{2i} \geq n - 2\sum_{i=1}^\infty r^i = n - \frac{2r}{1-r} = n - \frac{2}{e^{c_{upp}\alpha} - 1}.$$

Therefore,

$$\mathbb{1}_{\mathcal{A}_n}\Delta_n \geq \mathbb{1}_{\mathcal{A}_n}\frac{1}{c_{upp}^2}e^{-2\frac{c_{upp}}{c_{inf}}C_n}\left(n-\frac{2}{e^{c_{upp}\alpha}-1}\right).$$

Finally, since $e^x - 1 \geq x$ for all $x \geq 0$,

$$\frac{2}{e^{c_{upp}\alpha}-1} \leq \frac{2}{c_{upp}\alpha},$$

such that

$$\mathbb{1}_{\mathcal{A}_n}\Delta_n \geq \mathbb{1}_{\mathcal{A}_n}\frac{1}{c_{upp}^2}e^{-2\frac{c_{upp}}{c_{inf}}C_n}\left(n-\frac{2}{c_{upp}\alpha}\right) = \mathbb{1}_{\mathcal{A}_n}\frac{n}{c_{upp}^2}e^{-2\frac{c_{upp}}{c_{inf}}C_n}\left(1-\frac{2}{c_{upp}\alpha n}\right).$$

## 6.D Proof of Theorem 6.5.1

We prove Theorem 6.5.1 in this section, which we restate now.

**Theorem 6.5.1.** *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0,1)$, $\alpha \in (0,1)$. Let $K = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right)\right\rceil$, and $c_{upp} \geq 1$ be chosen such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let*

$$\mathcal{C}_n = 32(4 + 4C_n + C_n^2)\frac{c_{upp}e}{c_{inf}^2}, \text{ where } C_n = c_{inf}\alpha K.$$

*There exists a set $\mathcal{A}_n \subset \Omega$ of realizations satisfying $\mathbb{P}(\mathcal{A}_n) \geq 1 - \varepsilon$ such that the three following assertions hold*

$$\begin{aligned}
&(1) \quad \forall\, i \in \{K+1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i - T_j)}\right],\\
&(2) \quad \forall\, i \in \{1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)}\right],\\
&(3) \quad \forall\, i \in \{1, \dots, n\}, \quad \textstyle\sum_{j=1}^{i}(i-j)e^{-\alpha(T_i - T_j)} &&\leq \mathcal{C}_n\, \mathbb{E}\left[\textstyle\sum_{j=1}^{i}(i-j)e^{-\alpha(T_i - T_j)}\right].
\end{aligned}$$

*Sketch of Proof.* Define

$$H_0^i := \sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i - T_j)},\ H_1^i := \sum_{j=1}^{i} e^{-\alpha(T_i - T_j)},\ H_2^i := \sum_{j=1}^{i}(i-j)e^{-\alpha(T_i - T_j)}.$$

The proof follows the same five-step strategy for each of the three sums $H_0^i$, $H_1^i$, $H_2^i$.

**Step 1 — Choice of block size $K$.** Since $T_i - T_j = \sum_{k=j}^{i-1}\Delta_k$ with $\Delta_k \overset{\text{i.i.d.}}{\sim} \mathcal{E}(1)$, the Chernov inequality controls deviations of a sum of *fixed* length $K$ from its expectation, whose value is $K$. However, we need a bound valid for all gap lengths $i-j$ simultaneously. The key device is to fix a block size $K$ (whose value will be determined in Step 2) and decompose each sum into consecutive blocks of size $K$ plus a residual block.

**Step 2 — Concentration of $T_i - T_j$ (Proposition 6.D.1).** Applying Chernov to each block of size $K$, combined with a union bound over the $O(n)$ possible block starting points, gives with probability $1-\varepsilon$, for all $1 \leq j \leq i \leq n$:

$$c_{inf}(i-j) - c_{inf}K \ \leq\ T_i - T_j \ \leq\ c_{upp}(i-j) + c_{upp}K.$$

The value $K = \left\lceil (c_{inf} - 1 - \log c_{inf})^{-1}\log(2n/\varepsilon)\right\rceil$ is chosen to make the failure probability of each block at most $\varepsilon/(2n)$, so that using a union bound implies a probability over all the blocks of $\varepsilon/2$, that does not grows with $n$.

**Step 3 — High-probability upper-bounds on $H_k^i$ (Propositions 6.D.2–6.D.3–6.D.4).** Substituting the bounds of Step 2 into each sum yields upper-bounds with high probability on $\{H_k^i\}_{k\in\{0,1,2\}}$. These upper-bound involve terms that depend on $K$.

**Step 4 — lower-bounds on expectations (Lemmas 6.D.4–6.D.5–6.D.3).** Computing the characteristic function of the Gamma distribution, $\mathbb{E}[H_k^i]$ is computed exactly in closed form, then lower-bounded by quantity whose form matches those of the upper-bound in Step 3.

**Step 5 — Conclusion (Section 6.D.3).** The upper-bounds from Step 3 involve terms that depend on $K$, while the lower-bounds from Step 4 do not. We define $C_n$ such that $K = \mathcal{O}(C_n/\alpha)$, allowing to combine these bounds. We obtain that the sums $H_k^i$ are upper-bounded by their expectations, up to constants involving $C_n$. As long as $\alpha = n^{-\beta}$ for some $\beta > 0$, the constant $C_n$ will be uniformly upper-bounded with $n$, and will verify $C_n \to_{n\to+\infty} 0$.

□

## 6.D.1 Upper-bounding the sums

A key result for our proof is based on a rather classical concentration inequality called the Chernov inequality. We use it in the case of gamma laws, see *e.g.* (Boucheron et al., 2003, Section 2.4).

**Lemma 6.D.1** (Chernov inequality)**.** *For $k \in \mathbb{N}^*$, let $T_k \sim \Gamma(k, 1)$, $0 < c_{inf} \leq 1 \leq c_{upp}$. Then,*

*(i)* $\mathbb{P}(T_k \leq c_{inf} k) \leq e^{-(c_{inf}-1-\log(c_{inf}))k}$.

*(ii)* $\mathbb{P}(T_k \geq c_{upp} k) \leq e^{-(c_{upp}-1-\log(c_{upp}))k}$.

**Proposition 6.D.1.** *Let $\varepsilon, c_{inf} \in (0, 1)$, $n > 0$, and $K := \left\lceil (c_{inf} - 1 - \log c_{inf})^{-1} \log(2n/\varepsilon) \right\rceil$. Let $c_{upp} \geq 1$ be chosen such that $c_{upp}$ such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$. Let the set $\mathcal{A}_n$ such that we have for any $1 \leq j \leq i \leq n$*

$$c_{inf}(i - j) - c_{inf} K \leq T_i - T_j \leq c_{upp}(i - j) + c_{upp} K.$$

*Then, we have $\mathbb{P}(\mathcal{A}_n) \geq 1 - \varepsilon$.*

*Proof.* We note $\Delta_k := T_{k+1} - T_k$ such that

$$T_i - T_j = \sum_{k=j}^{i-1} \Delta_k,$$

where $\Delta_k \overset{i.i.d}{\sim} \mathcal{E}(1)$. Let an integer $K > 0$, to be further fixed.

**Lower-bound.** The idea is to decompose $\sum_{k=j}^{i-1} \Delta_k$ into $\lfloor (i-j)/K \rfloor$ blocks of size $K$ starting at $j, j+K, j+2K, \ldots$, plus a residual block of size $(i-j) \bmod K < K$. Chernov is applied to each complete block; the residual block is non-negative ($\Delta_k \geq 0$ a.s.) and is simply dropped, at the cost of the term $-c_{inf}K$ in the final bound.

We have $\sum_{k=j}^{j-1+K} \Delta_k = T_{j+K} - T_j$ which follows the same law as $T_K \sim \Gamma(K, 1)$. Therefore, for any $c_{inf} \in (0, 1)$, applying Lemma 6.D.1 $(i)$,

$$\sum_{k=j}^{j-1+K} \Delta_k \leq c_{inf} K, \tag{6.70}$$

with probability at most $e^{-K(c_{inf}-1-\log c_{inf})}$. Choosing $K = \left\lceil (c_{inf} - 1 - \log c_{inf})^{-1} \log(2n/\varepsilon) \right\rceil$, this probability is at most $\varepsilon/(2n)$. By a union bound over $j \in \{1, \ldots, n-K\}$, with probability $1 - \varepsilon/2$, (6.70) fails for all such $j$, namely

$$\sum_{k=j}^{j-1+K} \Delta_k > c_{inf} K. \tag{6.71}$$

Now, consider the following decomposition

$$\sum_{k=j}^{i-1} \Delta_k = \underbrace{\sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k}_{\lfloor (i-j)/K \rfloor \text{ complete blocks}} + \underbrace{\sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{i-1} \Delta_k}_{\text{residual block}}.$$

For $s \in \left\{0, \cdots, \left\lfloor \frac{i-j}{K} \right\rfloor - 1\right\}$, we have $j + sK \in \{j, \cdots, i-K\} \subset \{1, \cdots, n-K\}$. So, (6.71) gives that with probability $1 - \varepsilon/2$, for any such $s$ we have $\sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k > c_{inf} K$. The residual block can be dropped, at it is nonnegative. Then,

$$\begin{aligned}
\sum_{k=j}^{i-1} \Delta_k &= \sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{i-1} \Delta_k \\
&\geq \sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} c_{inf} K \\
&= c_{inf} K \left\lfloor \frac{i-j}{K} \right\rfloor \\
&\geq c_{inf} K \left( \frac{i-j}{K} - 1 \right) \qquad (\lfloor x \rfloor \geq x - 1) \\
&= c_{inf}(i-j) - c_{inf} K.
\end{aligned} \tag{6.72}$$

**Upper-bound.** As for the lower-bound, the strategy for the upper-bound exploits the following decomposition

$$\sum_{k=j}^{i-1} \Delta_k = \underbrace{\sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k}_{\lfloor (i-j)/K \rfloor \text{ complete blocks}} + \underbrace{\sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{i-1} \Delta_k}_{\text{residual block}}.$$

However, a difference is that the residual block cannot be dropped by a non-negativity argument, because here we want to obtain an upper-bound. A simple strategy is to bound a residual block by a complete block of size $K$, namely

$$\sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{i-1} \Delta_k \leq \sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{j+\left(\left\lfloor \frac{i-j}{K} \right\rfloor + 1\right)K-1} \Delta_k.$$

However, some terms in the sum of the right-hand side are not defined if $j + \left(\left\lfloor \frac{i-j}{K} \right\rfloor K + 1\right) - 1 \geq i$. So, for the purpose of the analysis, we introduce $\{\Delta_k\}_{n \leq k \leq n+K-1} \overset{\text{i.i.d.}}{\sim} \mathcal{E}(1)$, independent of $\{\Delta_k\}_{k \leq n-1}$, such that we can write

$$\begin{aligned}
\sum_{k=j}^{i-1} \Delta_k &= \sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{i-1} \Delta_k \\
&\leq \sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor - 1} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k + \sum_{k=j+\left\lfloor \frac{i-j}{K} \right\rfloor K}^{j+\left(\left\lfloor \frac{i-j}{K} \right\rfloor + 1\right)K-1} \Delta_k \\
&= \sum_{s=0}^{\left\lfloor \frac{i-j}{K} \right\rfloor} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k
\end{aligned} \tag{6.73}$$

Then, applying Lemma 6.D.1 $(ii)$, we have that for any $c_{upp} > 1$ and for $j \in \{1, \cdots, n-1\}$

$$\sum_{k=j}^{j-1+K} \Delta_k \geq c_{upp} K, \tag{6.74}$$

with probability at most

$$e^{-K(c_{upp} - 1 - \log(c_{upp}))}.$$

Choosing $c_{upp}$ such that $c_{upp} - \log(c_{upp}) = c_{inf} - \log(c_{inf})$, we obtain

$$\left\lceil (c_{upp} - 1 - \log(c_{upp}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil =: K,$$

and it follows that this probability is at most $\varepsilon/(2n)$. By a union bound over $j \in \{1, \dots, n-1\}$, with probability $1 - \frac{\varepsilon}{2}$, we have

$$\sum_{k=j}^{j-1+K} \Delta_k \leq c_{upp} K \tag{6.75}$$

So, from (6.73), we deduce that with probability $1 - \varepsilon/2$, we have

$$\sum_{k=j}^{i-1} \Delta_k \leq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor} \sum_{k=j+sK}^{j+(s+1)K-1} \Delta_k \leq \sum_{s=0}^{\lfloor \frac{i-j}{K} \rfloor} c_{upp} K = c_{upp} K \left( \left\lfloor \frac{i-j}{K} \right\rfloor + 1 \right) \leq c_{upp}(i-j) + c_{upp} K. \tag{6.76}$$

**Conclusion.** The claim follows by combining (6.72) and (6.76), each holding under probability $1 - \frac{\varepsilon}{2}$. □

**Proposition 6.D.2.** *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0,1)$, $\alpha \in (0,1]$ and $K = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil$. With probability $1 - \varepsilon$, we have*

$$\forall i \in \{1, \dots, n\}, \ \sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} \leq \min\left\{ i, \ \frac{2}{\alpha c_{inf}} + K \right\}.$$

*Proof.* We bound

$$H_1^i := \sum_{j=1}^{i} e^{-\alpha(T_i - T_j)}.$$

We note that as long as $i > K$, the $i - K$ first terms of $H_1^i$ can be controlled using the lower-bound on $T_i - T_j$ from Proposition 6.D.1. It remains the last terms, whose amount is at most $K$; these terms can be bounded by 1.

Let $i \leq K$. We have

$$\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} \leq i \leq \min\left\{ i, \ \frac{2}{\alpha c_{inf}} + K \right\}. \tag{6.77}$$

Let $i > K$. With probability $1 - \varepsilon$, we have

$$\begin{aligned}
\sum_{j=1}^{i} e^{-\alpha(T_i - T_j)} &= \sum_{j=1}^{i-K} e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^{i} e^{-\alpha(T_i - T_j)} \\
&\leq \sum_{j=1}^{i-K} e^{-\alpha(T_i - T_j)} + K \\
&\overset{(i)}{\leq} \sum_{j=1}^{i-K} e^{-\alpha(c_{inf}(i-j) - c_{inf} K)} + K \\
&\overset{(ii)}{=} \sum_{\ell=0}^{i-K-1} e^{-\alpha c_{inf} \ell} + K \\
&\overset{(iii)}{\leq} \min\left\{ i - K, \ \frac{1}{1 - e^{-\alpha c_{inf}}} \right\} + K \\
&\overset{(iv)}{\leq} \min\left\{ i, \ \frac{2}{\alpha c_{inf}} + K \right\}.
\end{aligned} \tag{6.78}$$

For step $(i)$, we apply $T_i - T_j \geq c_{inf}(i-j) - c_{inf} K$. This is due to Proposition 6.D.1, valid with probability $1 - \varepsilon$, for any $j \leq i - K$ since $i - j \geq K$. For step $(ii)$ we set the change of variable $\ell = i - K - j$, and simplify $e^{\alpha c_{inf} K} e^{-\alpha c_{inf} K} = 1$. $(iii)$ uses $\sum_{\ell=0}^{i-K-1} e^{-\alpha c_{inf} \ell} = \frac{1 - e^{-\alpha c_{inf}(i-K)}}{1 - e^{-\alpha c_{inf}}} \leq \frac{1}{1 - e^{-\alpha c_{inf}}}$. $(iv)$ uses $e^{-t} \leq 1 - t/2$ for $t = \alpha c_{inf} \in (0,1]$, giving $\frac{1}{1 - e^{-\alpha c_{inf}}} \leq \frac{2}{\alpha c_{inf}}$. □

**Proposition 6.D.3.** *Let $n \in \mathbb{N}^*$, $c_{inf}, \varepsilon \in (0,1)$, $\alpha \in (0,1]$ and $K = \lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \rceil$. With probability $1-\varepsilon$, we have*

$$\forall i \in \{1, \dots, n\},\ \sum_{j=1}^{i} (i-j) e^{-\alpha(T_i - T_j)} \leq \min\left\{ \frac{i^2}{2},\ \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}.$$

*Proof.* We bound

$$H_2^i := \sum_{j=1}^{i} (i-j) e^{-\alpha(T_i - T_j)}.$$

Compared with $H_1^i$, the weight $(i-j)$ adds a linear factor. When treating $H_2^i$, the main difference is that the sum $\sum_\ell \ell\, q^\ell$ is handled via $\sum_{\ell=0}^{\infty} \ell q^\ell = q/(1-q)^2$.

If $i \leq K$,

$$H_2^i \leq \sum_{j=1}^{i} (i-j) = \frac{i(i-1)}{2} \leq \frac{i^2}{2} \leq \min\left\{ \frac{i^2}{2},\ \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}.$$

If $i > K$, with probability $1-\varepsilon$, we have

$$\begin{aligned} H_2^i &= \sum_{j=1}^{i-K} (i-j) e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^{i} (i-j) e^{-\alpha(T_i - T_j)} \\ &\overset{(i)}{\leq} \sum_{j=1}^{i-K} (i-j) e^{-\alpha(c_{inf}(i-j) - c_{inf}K)} + \sum_{j=i-K+1}^{i} (i-j) \\ &\overset{(ii)}{\leq} \sum_{\ell=0}^{i-K-1} \big(\ell + K\big) q^\ell + \frac{K^2}{2}. \end{aligned} \tag{6.79}$$

Step $(i)$ uses $T_i - T_j \geq c_{inf}(i-j) - c_{inf}K$ , holding with probability $1-\varepsilon$, thanks to Proposition 6.D.1. It also uses $e^{-\alpha(T_i - T_j)} \leq 1$. In step $(ii)$ we set $q := e^{-\alpha c_{inf}}$ and the change of variable $\ell = i - j - K$.

Thus,

$$H_2^i \leq \sum_{\ell=0}^{i-K-1} \ell q^\ell + K \sum_{\ell=0}^{i-K-1} q^\ell + \frac{K^2}{2}. \tag{6.80}$$

Now,

$$\sum_{\ell=0}^{m-1} \ell q^\ell \leq \min\left\{ \sum_{\ell=0}^{m-1} \ell,\ \sum_{\ell=0}^{\infty} \ell q^\ell \right\} = \min\left\{ \frac{m(m-1)}{2},\ \frac{q}{(1-q)^2} \right\},$$

and

$$\sum_{\ell=0}^{m-1} q^\ell \leq \min\left\{ m,\ \frac{1}{1-q} \right\}.$$

Applying this with $m = i - K$, we obtain

$$\begin{aligned} H_2^i \leq \min&\left\{ \frac{(i-K)(i-K-1)}{2},\ \frac{q}{(1-q)^2} \right\} \\ &+ K \min\left\{ i-K,\ \frac{1}{1-q} \right\} + \frac{K^2}{2}. \end{aligned} \tag{6.81}$$

Using $\min(a,b) + \min(c,d) \leq \min(a+c, b+d)$, we get

$$\begin{aligned} H_2^i \leq \min&\left\{ \frac{(i-K)(i-K-1)}{2} + K(i-K) + \frac{K^2}{2}, \right. \\ &\left. \frac{q}{(1-q)^2} + \frac{K}{1-q} + \frac{K^2}{2} \right\}. \end{aligned} \tag{6.82}$$

The first term is equal to $\frac{i^2-i+K}{2}$. Since $i > K$, it follows that $\frac{i^2-i+K}{2} \leq \frac{i^2}{2}$. Then,

$$H_2^i \leq \min\left\{ \frac{i^2}{2}, \frac{q}{(1-q)^2} + \frac{K}{1-q} + \frac{K^2}{2} \right\}. \tag{6.83}$$

As moreover $\alpha c_{inf} \leq 1$, then

$$1 - q = 1 - e^{-\alpha c_{inf}} \geq \frac{\alpha c_{inf}}{2},$$

hence

$$\frac{1}{1-q} \leq \frac{2}{\alpha c_{inf}}, \qquad \frac{q}{(1-q)^2} \leq \frac{1}{(1-q)^2} \leq \frac{4}{(\alpha c_{inf})^2}.$$

Therefore,

$$H_2^i \leq \min\left\{ \frac{i^2}{2}, \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + \frac{K^2}{2} \right\}. \tag{6.84}$$

□

The following proposition differs from Propositions 6.D.2 and 6.D.3 for two reasons. First, because the weights $(T_i - T_j)$ are random, we need both the upper bound $(T_i - T_j \leq c_{upp}(i-j) + c_{upp}K)$ and the lower-bound $(T_i - T_j \geq c_{inf}(i-j) - c_{inf}K)$ from Proposition 6.D.1. The upper-bound controls the weight, the lower bound gives exponential decay. Then, a difficulty is that for small $i$, there is no direct upper-bound of $(T_i - T_j)$ that depends on $i$, which will lead to a crude upper-bound in this case that does not depend on $i$. Fortunately, it will be sufficient to prove our main result (Theorem 6.4.1).

**Proposition 6.D.4.** *Let $c_{inf}, \varepsilon \in (0,1)$, $c_{upp}$ and $K$ defined as in Proposition 6.D.1. With probability $1-\varepsilon$, we have*

$$\forall i \in \{1, \dots, K\},\ \sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq c_{upp} K^2,$$

*and*

$$\forall i \in \{K+1, \dots, n\},\ \sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)} \leq \min\left\{ c_{upp} i^2,\ \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp}K^2 \right\}$$

*Proof.* We bound

$$H_0^i := \sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)}.$$

We recall $c_{inf} < 1 < c_{upp}$ are chosen such that

$$K = \left\lceil (c_{upp} - 1 - \log(c_{upp}))^{-1} \log\left(\frac{2n}{\varepsilon}\right) \right\rceil = \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(\frac{2n}{\varepsilon}\right) \right\rceil.$$

If $i \leq K$, with probability $1 - \varepsilon$ we have

$$H_0^i \leq \sum_{j=1}^{i} (T_i - T_j) \overset{(i)}{\leq} \sum_{j=1}^{i} c_{upp} K \leq c_{upp} i K \leq c_{upp} K^2.$$

$(i)$ uses that as $i - j \leq K$, we have $T_i - T_j \leq T_{j+K} - T_j \leq c_{upp}K$, the last inequality being due to the Chernov inequality (see (6.75)).

If $i > K$, we write

$$\sum_{j=1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)} = \sum_{j=1}^{i-K} (T_i - T_j) e^{-\alpha(T_i - T_j)} + \sum_{j=i-K+1}^{i} (T_i - T_j) e^{-\alpha(T_i - T_j)}. \tag{6.85}$$

For the second sum, with probability $1-\varepsilon$,

$$\sum_{j=i-K+1}^{i}(T_i-T_j)e^{-\alpha(T_i-T_j)} \le \sum_{j=i-K+1}^{i}(T_i-T_j) \overset{(i)}{\le} c_{upp}K^2.$$

Again, $(i)$ uses that as $i-j\le K$, we have $T_i-T_j \le T_{j+K}-T_j \le c_{upp}K$.

Then, with probability $1-\varepsilon$, we have by Proposition 6.D.1,

$$T_i - T_j \ge c_{inf}(i-j) - c_{inf}K,$$

and

$$T_i - T_j \le c_{upp}(i-j) + c_{upp}K.$$

We use both inequalities simultaneously, the upper-bound controls the factor $(T_i-T_j)$, the lower-bound gives exponential decay in $e^{-\alpha(T_i-T_j)}$. Let $q := e^{-\alpha c_{inf}}$. Therefore,

$$\begin{aligned} H_0^i &\le \sum_{j=1}^{i-K}\big(c_{upp}(i-j)+c_{upp}K\big)e^{-\alpha(c_{inf}(i-j)-c_{inf}K)} + c_{upp}K^2 \\ &\overset{(i)}{\le} \sum_{\ell=0}^{i-K-1}\big(c_{upp}\ell + 2c_{upp}K\big)q^\ell + c_{upp}K^2. \end{aligned} \tag{6.86}$$

In (i) we set the change $\ell = i-j-K$. Then,

$$H_0^i \le c_{upp}\sum_{\ell=0}^{i-K-1}\ell q^\ell + 2c_{upp}K\sum_{\ell=0}^{i-K-1}q^\ell + c_{upp}K^2. \tag{6.87}$$

Now,

$$\sum_{\ell=0}^{m-1}\ell q^\ell \le \min\left\{\sum_{\ell=0}^{m-1}\ell,\ \sum_{\ell=0}^{\infty}\ell q^\ell\right\} = \min\left\{\frac{m(m-1)}{2},\ \frac{q}{(1-q)^2}\right\},$$

and

$$\sum_{\ell=0}^{m-1}q^\ell \le \min\left\{m,\ \frac{1}{1-q}\right\}.$$

Applying this with $m=i-K$, we obtain

$$\begin{aligned} H_0^i \le c_{upp}\min&\left\{\frac{(i-K)(i-K-1)}{2},\ \frac{q}{(1-q)^2}\right\} \\ &+ 2c_{upp}K\min\left\{i-K,\ \frac{1}{1-q}\right\} + c_{upp}K^2. \end{aligned} \tag{6.88}$$

Using $\min(a,b)+\min(c,d)\le\min(a+c,b+d)$, we get

$$\begin{aligned} H_0^i \le \min\Bigg\{&c_{upp}\frac{(i-K)(i-K-1)}{2} + 2c_{upp}K(i-K) + c_{upp}K^2, \\ &c_{upp}\frac{q}{(1-q)^2} + \frac{2c_{upp}K}{1-q} + c_{upp}K^2\Bigg\}. \end{aligned} \tag{6.89}$$

The first component in the min is equal to $c_{upp}\frac{i^2-i+2iK+K-K^2}{2}$. Then,

$$\begin{aligned} c_{upp}\frac{i^2-i+2iK+K-K^2}{2} &= c_{upp}\frac{i^2-i+(i+1/2)^2-(K-i-1/2)^2}{2} \\ &\le c_{upp}\frac{i^2-i+(i+1/2)^2-(1/2)^2}{2},\ \ as\ K<i \\ &= c_{upp}. \end{aligned}$$

So, we deduce

$$H_0^i \leq \min\left\{c_{upp} i^2,\, c_{upp}\frac{q}{(1-q)^2} + \frac{2c_{upp}K}{1-q} + c_{upp}K^2\right\}. \tag{6.90}$$

As moreover $\alpha c_{inf} \leq 1$, then

$$1 - q = 1 - e^{-\alpha c_{inf}} \geq \frac{\alpha c_{inf}}{2},$$

hence

$$\frac{1}{1-q} \leq \frac{2}{\alpha c_{inf}}, \qquad \frac{q}{(1-q)^2} \leq \frac{1}{(1-q)^2} \leq \frac{4}{(\alpha c_{inf})^2}.$$

Therefore,

$$H_0^i \leq \min\left\{c_{upp} i^2,\, \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp}K^2\right\}. \tag{6.91}$$

□

### 6.D.2 lower-bounding expectations

The three lemmas below compute or lower-bound the terms $\mathbb{E}[H_k^i]$, recalling

$$H_0^i := \sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i - T_j)},\; H_1^i := \sum_{j=1}^{i} e^{-\alpha(T_i-T_j)},\; H_2^i := \sum_{j=1}^{i}(i-j)e^{-\alpha(T_i-T_j)}.$$

Since $T_i - T_j \sim \Gamma(i-j, 1)$, the Laplace transform of the Gamma distribution gives exact closed-form expressions, which are then lower-bounded using elementary inequalities.

**Lemma 6.D.2.** *We have*

- $\mathbb{E}\left[H_0^i\right] = \frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}$
- $\mathbb{E}\left[H_2^i\right] = (1+\alpha)\frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}$

*Proof.* Applying Lemma 6.E.3, we have

$$\mathbb{E}\left[H_2^i\right] = \sum_{j=1}^{i}(i-j)\mathbb{E}\left[e^{-\alpha(T_i-T_j)}\right] = \sum_{j=1}^{i}(i-j)(1+\alpha)^{-(i-j)} = \sum_{k=1}^{i-1} k(1+\alpha)^{-k}.$$

By Lemma 6.E.2 applied to $\gamma = (1+\alpha)^{-1}$, we have

$$\begin{aligned}\mathbb{E}\left[H_2^i\right] &= (1+\alpha)^{-1}\frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{(1-(1+\alpha)^{-1})^2} \\ &= (1+\alpha)\frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}.\end{aligned}$$

By Lemma 6.C.2, since $T_i - T_j \sim \Gamma(i-j, 1)$ and setting $k = i - j$:

$$\begin{aligned}\mathbb{E}\left[H_i^0\right] &= \sum_{j=1}^{i}\mathbb{E}\left[(T_i - T_j)e^{-\alpha(T_i-T_j)}\right] = \sum_{k=1}^{i-1} k(1+\alpha)^{-(k+1)} \\ &= (1+\alpha)^{-1}\sum_{k=1}^{i-1} k(1+\alpha)^{-k} = (1+\alpha)^{-1}\mathbb{E}\left[H_i^2\right] \\ &= \frac{1 - i(1+\alpha)^{-(i-1)} + (i-1)(1+\alpha)^{-i}}{\alpha^2}.\end{aligned}$$

□

**Lemma 6.D.3.**

$$\mathbb{E}\left[H_1^i\right] \geq (1-e^{-1}) \min\left\{i, \frac{1}{\alpha}\right\}.$$

*Proof.* Applying Lemma 6.E.3 and bounding the discrete sum by an integral, we have:

$$\mathbb{E}\left[H_1^i\right] = \sum_{j=0}^{i-1} (1+\alpha)^{-j} \geq \sum_{j=0}^{i-1} e^{-\alpha j} \geq \int_0^i e^{-\alpha t} dt = \frac{1-e^{-\alpha i}}{\alpha}.$$

We also applied the standard inequality $\log(1+x) \leq x$ to $x = \alpha$, which implies $(1+\alpha)^{-1} \geq e^{-\alpha}$, then $(1+\alpha)^{-j} \geq e^{-\alpha j}$. Now, we split the proof between the case $i \leq \alpha^{-1}$ and $i > \alpha^{-1}$.
**Case** $i \leq \alpha^{-1}$**.** Then $\alpha i \in [0,1]$. By concavity of $t \mapsto 1-e^{-t}$ on $[0,1]$, we have $1-e^{-t} \geq (1-e^{-1})t$. Applying it to $t = \alpha i$ we deduce

$$\frac{1-e^{-\alpha i}}{\alpha} \geq (1-e^{-1})i.$$

**Case** $i > \alpha^{-1}$**.** We have $\alpha i > 1$, which implies $1-e^{-\alpha i} > 1-e^{-1}$. Hence

$$\frac{1-e^{-\alpha i}}{\alpha} > \frac{1-e^{-1}}{\alpha}$$

From the two cases, we deduce

$$\mathbb{E}\left[H_1^i\right] \geq (1-e^{-1}) \min\left\{i, \frac{1}{\alpha}\right\}.$$

□

**Lemma 6.D.4.** *We have*

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

*Proof.* Writing $r = (1+\alpha)^{-1}$, Lemma 6.D.2 gives

$$\mathbb{E}[H_0^i] = \frac{1 - i r^{\,i-1} + (i-1) r^i}{\alpha^2} = \frac{1 - r^{\,i-1}(i - r(i-1))}{\alpha^2}.$$

Since

$$i - (i-1)r = 1 + (i-1)(1-r) = 1 + \frac{\alpha(i-1)}{1+\alpha},$$

noting $y := \frac{\alpha(i-1)}{1+\alpha}$, it follows that

$$\mathbb{E}[H_0^i] = \frac{1 - r^{\,i-1}(1+y)}{\alpha^2}.$$

We use $\log(1+x) \geq x/(1+x)$ at $x = \alpha$, which induces

$$r^{\,i-1} = e^{-(i-1)\log(1+\alpha)} \leq e^{-(i-1)\frac{\alpha}{1+\alpha}} = e^{-y},$$

and hence

$$\mathbb{E}[H_0^i] \geq \frac{1-(1+y)e^{-y}}{\alpha^2}.$$

Now, we note the following integration by part property: $\int_0^y te^{-t}\,dt = [-te^{-t}]_0^y + \int_0^y e^{-t}\,dt = -ye^{-y} + (1-e^{-y}) = 1-(1+y)e^{-y}$. We lower-bound this integral form, depending on the value of $y$.

- $y \leq 1$: $\int_0^y te^{-t}\,dt \geq e^{-1}\int_0^y t\,dt = \frac{e^{-1}}{2}y^2$.
- $y \geq 1$: $\int_0^y te^{-t}\,dt \geq \int_0^1 te^{-t}\,dt \geq \frac{e^{-1}}{2}$.

From this we deduce

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{2\alpha^2} \min\{y^2, 1\}.$$

As $0 < \alpha \leq 1$, then $y \geq \alpha(i-1)/2$, so $\min\{y^2, 1\} \geq \min\{\alpha^2(i-1)^2/4, 1\}$. This implies

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{8} \min\left\{(i-1)^2, \frac{1}{\alpha^2}\right\}.$$

For $i \geq 2$, we have $(i-1) \geq i/2$ so $(i-1)^2 \geq i^2/4$. We then conclude

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

□

**Lemma 6.D.5.** *We have*

$$\mathbb{E}[H_2^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}.$$

*Proof.* The claim can be deduced from Lemma 6.D.4. Indeed, recall

$$\mathbb{E}[H_0^i] = \sum_{k=1}^{i-1} \frac{k}{(1+\alpha)^{k+1}} \leq \sum_{k=1}^{i-1} \frac{k}{(1+\alpha)^k} = \mathbb{E}\left[H_2^i\right],$$

where we used $\alpha > 0$. □

### 6.D.3 Combining lower and upper-bounds

We conclude the proof of Theorem 6.5.1 in this section.

We define

$$C_n := \left\lceil (c_{inf} - 1 - \log(c_{inf}))^{-1} \log\left(2\frac{n}{\varepsilon}\right) \right\rceil c_{inf}\alpha = K c_{inf}\alpha. \tag{6.92}$$

In the proof, we will replace $K$ with $C_n/(\alpha c_{inf})$.Therefore, $C_n$ measures the cost of replacing $K$ by $(\alpha c_{inf})^{-1}$ in the upper-bounds of Section 6.D.1. Note that for our convergence result, our final choice of $\alpha$ will be of order $1/n^{1/7}$, such that $C_n$ will be of order $\log(\frac{n}{\varepsilon})n^{-1/7}$. Therefore, we will have $C_n \to 0$ as $n \to +\infty$, and that $C_n$ is uniformly bounded on $n \in \mathbb{N}^*$.

The proof for the 3 sums are similar, it combines the upper-bounds from Section 6.D.1 and the lower-bounds from Section 6.D.2.

**(1st step.)** From Proposition 6.D.4, with probability $1-\varepsilon$, we have

$$\forall i \in \{K+1, \ldots, n\},\ H_0^i \leq \min\left\{c_{upp} i^2,\ \frac{4c_{upp}}{(\alpha c_{inf})^2} + \frac{4c_{upp}K}{\alpha c_{inf}} + c_{upp}K^2\right\}.$$

Using the definition (6.92) of $C_n$, this becomes

$$\forall i \in \{K+1, \ldots, n\},\ H_0^i \leq \min\left\{c_{upp} i^2,\ \left(4 + 4C_n + C_n^2\right)\frac{c_{upp}}{(\alpha c_{inf})^2}\right\}.$$

As $(4 + 4C_n + C_n)/c_{inf}^2 \geq 1$, it implies that with probability $1-\varepsilon$, $\forall i \in \{K+1, \ldots, n\}$, we have

$$H_0^i \leq \frac{c_{upp}}{c_{inf}^2}\left(4 + 4C_n + C_n^2\right) \min\left\{i^2, \frac{1}{\alpha^2}\right\}, \tag{6.93}$$

Now, from Lemma 6.D.4, we have

$$\mathbb{E}[H_0^i] \geq \frac{e^{-1}}{32} \min\left\{i^2, \frac{1}{\alpha^2}\right\}. \tag{6.94}$$

Combining (6.93) and (6.94), we deduce that with probability $1-\varepsilon$, $\forall i \in \{K+1,\dots,n\}$, we have

$$\sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i-T_j)} \leq 32(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2}\mathbb{E}\left[\sum_{j=1}^{i}(T_i - T_j)e^{-\alpha(T_i-T_j)}\right].$$

**(2nd step.)** From Proposition 6.D.2, with probability $1-\varepsilon$

$$\forall i \in \{1,\dots,n\},\ H_1^i \leq \min\left\{i,\ \frac{2}{\alpha c_{inf}} + K\right\}.$$

Using the definition (6.92) of $C_n$, we get that with probability $1-\varepsilon$, $\forall i \in \{1,\dots,n\}$,

$$\forall i \in \{1,\dots,n\},\ H_1^i \leq \min\left\{i,\ \frac{2+C_n}{\alpha c_{inf}}\right\}. \tag{6.95}$$

Now, from Lemma 6.D.3, for any $i \in \{1,\dots,n\}$,

$$(1-e^{-1})^{-1}\mathbb{E}\left[H_1^i\right] \geq \min\left\{i, \frac{1}{\alpha}\right\}. \tag{6.96}$$

Combining (6.95) and (6.96), we deduce that with probability $1-\varepsilon$,

$$\forall i \in \{1,\dots,n\},\ H_1^i \leq \frac{2+C_n}{c_{inf}}(1-e^{-1})^{-1}\mathbb{E}\left[H_1^i\right].$$

**(3d step.)** By Proposition 6.D.3, with probability $1-\varepsilon$,

$$\forall i \in \{1,\dots,n\},\ H_2^i \leq \min\left\{\frac{i^2}{2},\ \frac{4}{(\alpha c_{inf})^2} + \frac{2K}{\alpha c_{inf}} + K^2\right\}.$$

Using the definition (6.92) of $C_n$, we get that with probability $1-\varepsilon$, $\forall i \in \{1,\dots,n\}$,

$$H_2^i \leq \min\left\{\frac{i^2}{2},\ \frac{4+2C_n+C_n^2}{(\alpha c_{inf})^2}\right\}.$$

Because $1/2 < 4 + 2C_n + C_n^2$, we deduce that with probability $1-\varepsilon$, $\forall i \in \{1,\dots,n\}$,

$$H_2^i \leq \frac{(4+2C_n+C_n^2)}{c_{inf}^2}\min\left\{i^2, \frac{1}{\alpha^2}\right\}. \tag{6.97}$$

Now, from Lemma 6.D.5, for any $i \in \{1,\dots,n\}$,

$$\mathbb{E}[H_2^i] \geq \frac{e^{-1}}{32}\min\left\{i^2, \frac{1}{\alpha^2}\right\}. \tag{6.98}$$

Combining (6.97) and (6.98), we deduce that with probability $1-\varepsilon$,

$$\forall i \in \{1,\dots,n\},\ H_2^i \leq 32(4+2C_n+C_n^2)\frac{e}{c_{inf}^2}\mathbb{E}\left[H_2^i\right].$$

**Conclusion.** We conclude by defining

$$\mathcal{C}_n := \max\left\{32(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2}, \frac{2+C_n}{c_{inf}}(1-e^{-1})^{-1}, 32(4+2C_n+C_n^2)\frac{e}{c_{inf}^2}\right\}.$$

It follows that

$$\mathcal{C}_n = 32(4+4C_n+C_n^2)\frac{c_{upp}e}{c_{inf}^2}.$$

## 6.E Additional technical Lemmas used to prove Theorem 6.4.1

In this section, we state and prove lemmas used to prove Theorem 6.4.1, but not explicitly stated in Section 6.5. We start with some elementary results.

**Lemma 6.E.1.** *Let $\alpha > 0$, $i \geq 1$. We have for $c > 0$*

$$\sum_{j=1}^{i}(1+c\alpha)^{-(i-j)} = \frac{1+c\alpha-(1+c\alpha)^{-i+1}}{c\alpha},$$

*Proof.*

$$\begin{aligned}\sum_{j=1}^{i}(1+c\alpha)^{-(i-j)} &= \sum_{k=0}^{i-1}(1+c\alpha)^{-k} \\ &= \frac{1-(1+c\alpha)^{-i}}{1-(1+c\alpha)^{-1}} \\ &= \frac{1+c\alpha-(1+c\alpha)^{-i+1}}{c\alpha}\end{aligned}$$

□

**Lemma 6.E.2.** *Let $\gamma \neq 1$, $n \in \mathbb{N}^*$. We have*

$$\sum_{k=1}^{n} k\gamma^k = \gamma\frac{1-(n+1)\gamma^n+n\gamma^{n+1}}{(1-\gamma)^2}. \tag{6.99}$$

*Proof.* Define $\varphi(\gamma) := \sum_{k=1}^{n}\gamma^k = \frac{\gamma-\gamma^{n+1}}{1-\gamma}$. We have

$$\varphi'(y) = \sum_{k=1}^{n} k\gamma^{k-1} = \frac{(1-(n+1)\gamma^n)(1-\gamma)+\gamma-\gamma^{n+1}}{(1-\gamma)^2} = \frac{1-(n+1)\gamma^n+n\gamma^{n+1}}{(1-\gamma)^2}.$$

Multiply by $\gamma$ allows to get (6.99).

□

**Lemma 6.E.3.** *If $i \leq j$, we have*

$$\mathbb{E}\left[e^{-\alpha(T_j-T_i)}\right] = (1+\alpha)^{-j+i}.$$

*Proof.* It is trivially true if $i = j$. If $i < j$, we have

$$e^{-\alpha(T_j-T_i)} = \prod_{k=0}^{j-i-1} e^{-\alpha(T_{k+i+1}-T_{k+i})}$$

By independence of the $T_{k+1}-T_k$ for all $k$, we have

$$\mathbb{E}\left[\prod_{k=0}^{j-i-1} e^{-\alpha(T_{k+i+1}-T_{k+i})}\right] = \prod_{k=0}^{j-i-1}\mathbb{E}\left[e^{-\alpha(T_{i+j+1}-T_{i+j})}\right].$$

Recall $T_{k+1}-T_k$ for all $k$ follows an exponential law with parameter 1, such that we have $\mathbb{E}\left[e^{-\alpha(T_{i+j+1}-T_{i+j})}\right] = (1+\alpha)^{-1}$. So,

$$\begin{aligned}\prod_{k=0}^{j-i-1}\mathbb{E}\left[e^{-\alpha(T_{i+j+1}-T_{i+j})}\right] &= \prod_{k=0}^{j-i-1}(1+\alpha)^{-1} \\ &= (1+\alpha)^{-j+i}.\end{aligned}$$

□

**Lemma 6.E.4.** *For $n \geq 1$, $\alpha > 0$, we have*

*(i) The following upper-bound*

$$\mathbb{E}\left[\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right] \leq \left(1+\frac{3}{2}\alpha\right)(1+2\alpha)n.$$

*(ii) The following equality*

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right] \\
&= \frac{(2+\alpha)(1+2\alpha)}{2}n + \frac{(1+2\alpha)(1+\frac{3}{2}\alpha)}{2\alpha}(1-(1+2\alpha)^{-n}) \\
&- \frac{2(1+\alpha)(1+2\alpha)}{\alpha}(1-(1+\alpha)^{-n})
\end{aligned}$$

*Proof.* **Point (i)** We have

$$\begin{aligned}
\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t &= \sum_{i=1}^n\left(\sum_{j=1}^i \alpha e^{\alpha(T_j-T_i)}\right)^2 \\
&= \alpha^2\sum_{i=1}^n\sum_{j=1}^i e^{-2\alpha(T_i-T_j)} + 2\alpha^2\sum_{i=1}^n\sum_{1\leq j<l\leq i} e^{-\alpha(T_i-T_j)}e^{-\alpha(T_i-T_l)}. \qquad (6.100)
\end{aligned}$$

We focus on the first term in (6.100). We have by using Lemma 6.E.3 and Lemma 6.E.1 ($c=2$)

$$\begin{aligned}
\mathbb{E}\left[\alpha^2\sum_{i=1}^n\sum_{j=1}^i e^{-2\alpha(T_i-T_j)}\right] &= \alpha^2\sum_{i=1}^n\sum_{j=1}^i \mathbb{E}\left[e^{-2\alpha(T_i-T_j)}\right] \\
&= \alpha^2\sum_{i=1}^n\sum_{j=1}^i (1+2\alpha)^{-(i-j)} \\
&= \alpha^2\sum_{i=1}^n \frac{1+2\alpha-(1+2\alpha)^{-i+1}}{2\alpha}
\end{aligned}$$

Omitting the non positive term, we have

$$\mathbb{E}\left[\alpha^2\sum_{i=1}^n\sum_{j=1}^i e^{-2\alpha(T_i-T_j)}\right] \leq \alpha n\frac{1+2\alpha}{2}.$$

We focus on the second term of (6.100). We have

$$2\alpha^2\sum_{i=1}^n\sum_{1\leq j<l\leq i} e^{-\alpha(T_i-T_j)}e^{-\alpha(T_i-T_l)} = 2\alpha^2\sum_{i=1}^n\sum_{1\leq j<l\leq i} e^{-2\alpha(T_i-T_l)}e^{-\alpha(T_l-T_j)}$$

We take expectation and we use Lemma 6.E.3

$$\begin{aligned}
2\alpha^2\mathbb{E}\left[\sum_i^n \sum_{1\le j<l\le i} e^{-\alpha(T_i-T_j)}e^{-\alpha(T_i-T_l)}\right] &= 2\alpha^2\sum_{i=1}^n \sum_{1\le j<l\le i} \mathbb{E}\left[e^{-2\alpha(T_i-T_l)}e^{-\alpha(T_l-T_j)}\right] \\
&= 2\alpha^2\sum_{i=1}^n \sum_{1\le j<l\le i} \mathbb{E}\left[e^{-2\alpha(T_i-T_l)}\right]\mathbb{E}\left[e^{-\alpha(T_l-T_j)}\right] \\
&= 2\alpha^2\sum_{i=1}^n \sum_{1\le j<l\le i} (1+2\alpha)^{-(i-l)}(1+\alpha)^{-(l-j)} \\
&= 2\alpha^2\sum_{i=1}^n \sum_{l=1}^i \sum_{j=1}^{l-1} (1+2\alpha)^{-(l-j)}(1+\alpha)^{-(i-j)}.
\end{aligned}$$

Now, using twice Lemma 6.E.1 with $c=1$ and $c=2$, and omitting non negative terms

$$\begin{aligned}
2\alpha^2\sum_{i=1}^n \sum_{l=1}^i \sum_{j=1}^{l-1} (1+2\alpha)^{-(i-l)}(1+\alpha)^{-(i-j)} &= 2\alpha^2\sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \sum_{j=1}^{l-1} (1+\alpha)^{-(l-j)} \\
&= 2\alpha^2\sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \left(\frac{1+\alpha-(1+\alpha)^{-i+1}}{\alpha}\right) \\
&\le 2\alpha(1+\alpha)\sum_{i=1}^n \sum_{l=1}^i (1+2\alpha)^{-(i-l)} \\
&= 2\alpha(1+\alpha)\sum_{i=1}^n \left(\frac{1+2\alpha-(1+2\alpha)^{-i+1}}{2\alpha}\right) \\
&\le (1+\alpha)(1+2\alpha)n
\end{aligned}$$

We then deduce that

$$2\alpha^2\mathbb{E}\left[\sum_i^n \sum_{1\le j<l\le i} e^{-\alpha(T_i-T_j)}e^{-\alpha(T_i-T_l)}\right] \le (1+\alpha)(1+2\alpha)n.$$

Finally, we have the bound

$$\begin{aligned}
\mathbb{E}\left[\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)}dN_s\right)^2 dN_t\right] &\le \alpha\frac{1+2\alpha}{2}n + (1+\alpha)(1+2\alpha)n \\
&= \left(1+\frac{3}{2}\alpha\right)(1+2\alpha)n.
\end{aligned}$$

**Point (ii)** Let $Y_k = \left(\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s\right)^2$. We have

$$\begin{aligned}
Y_{k+1} &= \left(\int_{[0,T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})}dN_s\right)^2 \\
&= \left(e^{\alpha(T_k-T_{k+1})}\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s + \int_{]T_k,T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})}dN_s\right)^2.
\end{aligned}$$

We have $\int_{]T_k,T_{k+1}]} \alpha e^{\alpha(s-T_{k+1})}dN_s = \alpha$. So

$$Y_{k+1} = e^{2\alpha(T_k-T_{k+1})}Y_k + 2\alpha e^{\alpha(T_k-T_{k+1})}\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s + \alpha^2.$$

We take the expectation

$$\begin{aligned}
\mathbb{E}\left[Y_{k+1}\right] &= \mathbb{E}\left[e^{2\alpha(T_k - T_{k+1})}Y_k + 2\alpha e^{\alpha(T_k - T_{k+1})}\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s + \alpha^2\right] \\
&= \mathbb{E}\left[e^{2\alpha(T_k - T_{k+1})}Y_k\right] + 2\alpha\mathbb{E}\left[e^{\alpha(T_k - T_{k+1})}\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s\right] + \alpha^2 \\
&= \mathbb{E}\left[e^{2\alpha(T_k - T_{k+1})}\right]\mathbb{E}\left[Y_k\right] + 2\alpha\mathbb{E}\left[e^{\alpha(T_k - T_{k+1})}\right]\mathbb{E}\left[\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)}dN_s\right] + \alpha^2 \\
&= (1+2\alpha)^{-1}\mathbb{E}\left[Y_k\right] + 2\alpha(1+\alpha)^{-1}(1+\alpha-(1+\alpha)^{-k+1}) + \alpha^2 \\
&= (1+2\alpha)^{-1}\mathbb{E}\left[Y_k\right] + 2\alpha(1-(1+\alpha)^{-k}) + \alpha^2 \\
&= (1+2\alpha)^{-1}\mathbb{E}\left[Y_k\right] - 2\alpha(1+\alpha)^{-k} + \alpha(2+\alpha).
\end{aligned}$$

Noting $u_k := \mathbb{E}\left[Y_k\right]$, we get a recursive formula of the form

$$u_{k+1} = au_k + b_k + c,$$

whose solution is given by

$$u_k = a^k \sum_{i=0}^{k-1} \frac{b_i + c}{a^{i+1}} = \sum_{i=0}^{k-1}(b_i + c)a^{k-i-1}. \tag{6.101}$$

We compute (6.101) in two steps. We start with

$$\begin{aligned}
\sum_{i=0}^{k-1} ca^{k-i-1} &= c\sum_{j=0}^{k-1} a^j \\
&= c\frac{1-a^k}{1-a} \\
&= \alpha(2+\alpha)\frac{1-(1+2\alpha)^{-k}}{1-(1+2\alpha)^{-1}} \\
&= \frac{(2+\alpha)(1+2\alpha)}{2} - \frac{2+\alpha}{2}(1+2\alpha)^{-k+1}.
\end{aligned} \tag{6.102}$$

Then, we have

$$\begin{aligned}
a^{k-1}\sum_{i=0}^{k-1}\frac{b_i}{a^i} &= -2\alpha a^{k-1}\sum_{i=0}^{k-1}\left(\frac{1+2\alpha}{1+\alpha}\right)^i \\
&= -2\alpha a^{k-1}\left(1-\left(\frac{1+2\alpha}{1+\alpha}\right)^k\right)\left(1-\left(\frac{1+2\alpha}{1+\alpha}\right)\right)^{-1} \\
&= 2(1+\alpha)\left(a^{k-1} - a^{-1}(1+\alpha)^{-k}\right) \\
&= 2(1+\alpha)(1+2\alpha)((1+2\alpha)^{-k} - (1+\alpha)^{-k}) \\
&= 2(1+\alpha)(1+2\alpha)(1+2\alpha)^{-k} - 2(1+\alpha)(1+2\alpha)(1+\alpha)^{-k}.
\end{aligned} \tag{6.103}$$

Injecting (6.102) and (6.103) into (6.101), we get the following expression for $u_k$

$$\begin{aligned}
u_k &= \frac{(2+\alpha)(1+2\alpha)}{2} - \frac{2+\alpha}{2}(1+2\alpha)^{-k+1} + 2(1+\alpha)(1+2\alpha)(1+2\alpha)^{-k} \\
&\quad - 2(1+\alpha)(1+2\alpha)(1+\alpha)^{-k} \\
&= \frac{(2+\alpha)(1+2\alpha)}{2} + \left(-\frac{(2+\alpha)(1+2\alpha)}{2} + 2(1+\alpha)(1+2\alpha)\right)(1+2\alpha)^{-k} \\
&\quad - 2(1+\alpha)(1+2\alpha)(1+\alpha)^{-k} \\
&= \frac{(2+\alpha)(1+2\alpha)}{2} + (1+2\alpha)(1+\frac{3}{2}\alpha)(1+2\alpha)^{-k} - 2(1+\alpha)(1+2\alpha)(1+\alpha)^{-k}.
\end{aligned}$$

It remains to sum over $k = 0, \dots, n$, to get

$$\sum_{k=1}^{n} u_k = \frac{(2+\alpha)(1+2\alpha)}{2} n + \frac{(1+2\alpha)(1+\frac{3}{2}\alpha)}{2\alpha}(1-(1+2\alpha)^{-n}) - \frac{2(1+\alpha)(1+2\alpha)}{\alpha}(1-(1+\alpha)^{-n}).$$

To conclude, remark that

$$\mathbb{E}\left[\int_0^{T_n}\left(\int_0^t \alpha e^{\alpha(s-t)} dN_s\right)^2 dN_t\right] = \sum_{k=1}^{n} \mathbb{E}\left[\left(\int_{[0,T_k]} \alpha e^{\alpha(s-T_k)} dN_s\right)^2\right] = \sum_{k=1}^{n} \mathbb{E}\left[Y_k\right] = \sum_{k=1}^{n} u_k.$$

□

**Lemma 6.E.5.** *Let $\tilde{\varepsilon} \in (0, 1/2]$, and $c_{upp} \geq 1$ such that it satisfies $1 - \tilde{\varepsilon} - \log(1-\tilde{\varepsilon}) = c_{upp} - \log(c_{upp})$. Then, we have $c_{upp} \leq 1 + 2\tilde{\varepsilon}$.*

*Proof.* We denote $h(t) = t - \log(t)$. By assumption, we have $h(c_{upp}) = h(1-\tilde{\varepsilon})$. We have that $h$ is non-decreasing on $[1, +\infty)$. Then, if we show $h(1 + c_0\tilde{\varepsilon}) \geq h(1-\tilde{\varepsilon})$ for some $c_0 \geq 1$, it will induce $c_{upp} \leq 1 + c_0\tilde{\varepsilon}$. First,

$$h(1+c_0\tilde{\varepsilon}) - 1 = c_0\tilde{\varepsilon} - \log(1+c_0\tilde{\varepsilon}) = \int_0^{c_0\tilde{\varepsilon}} \frac{t}{1+t}\, dt.$$

Since $t \mapsto (1+t)^{-1}$ is decreasing, for $t \in [0, c_0\tilde{\varepsilon}]$, $\frac{1}{1+t} \geq \frac{1}{1+c_0\tilde{\varepsilon}}$, therefore,

$$h(1+c_0\tilde{\varepsilon}) - 1 \geq \frac{1}{1+c_0\tilde{\varepsilon}} \int_0^{c_0\tilde{\varepsilon}} t\, dt = \frac{c_0^2\tilde{\varepsilon}^2}{2(1+c_0\tilde{\varepsilon})}.$$

On the other hand,

$$h(1-\tilde{\varepsilon}) - 1 = -\tilde{\varepsilon} - \log(1-\tilde{\varepsilon}) = \int_0^{\tilde{\varepsilon}} \frac{t}{1-t}\, dt.$$

Since $t \mapsto (1-t)^{-1}$ is increasing on $[0, \tilde{\varepsilon}]$, $\frac{1}{1-t} \leq \frac{1}{1-\tilde{\varepsilon}}$, thus

$$h(1-\tilde{\varepsilon}) - 1 \leq \frac{1}{1-\tilde{\varepsilon}} \int_0^{\tilde{\varepsilon}} t\, dt = \frac{\tilde{\varepsilon}^2}{2(1-\tilde{\varepsilon})}.$$

Consequently, it is enough that $\frac{c_0^2\tilde{\varepsilon}^2}{2(1+c_0\tilde{\varepsilon})} \geq \frac{\tilde{\varepsilon}^2}{2(1-\tilde{\varepsilon})}$, or equivalently $c_0^2(1-\tilde{\varepsilon}) \geq 1 + c_0\tilde{\varepsilon}$. This is equivalent to

$$\tilde{\varepsilon} \leq \frac{c_0^2 - 1}{c_0^2 + c_0} = 1 - \frac{1}{c_0}.$$

Therefore, for any $c_0 > 1$, choosing $a := 1 - \frac{1}{c_0}$ gives, for all $\tilde{\varepsilon} \in (0, a]$, $h(1+c_0\tilde{\varepsilon}) \geq h(1-\tilde{\varepsilon}) = h(c_{upp})$, from which we deduce

$$c_{upp} \leq 1 + c_0\tilde{\varepsilon}.$$

The result is the particular case $c_0 = 2 \Rightarrow \tilde{\varepsilon} \in \left(0, \frac{1}{2}\right]$. □

# Conclusion and Perspectives

This thesis was devoted to a study of Nesterov's momentum for minimizing non-convex functions and/or when using stochastic estimates of the gradient. Specifically, we were interested in settings where acceleration over gradient descent can be proven. Most of our results were not about improving the existing bounds, except in Chapter 4. We rather mostly attempted to tighten the answer to the question: *When can we accelerate with Nesterov's momentum ?*. Our answers were either about having a better understanding of the existing bounds, by exhibiting intuitive conditions that directly impact the possibility to have accelerated bounds, as in Chapters 2 and 5, or about seeing if existing mechanisms that are added to momentum to achieve accelerated bounds can be discarded, motivated by the fact that such "crutches" are not used in many applications, as in Chapters 3 and Chapter 6. We summarize the main messages of our contributions as follows.

1. *In some settings, the possibility of theoretical acceleration can be characterized through geometric correlation conditions, which quantify the relevance of momentum.* In the stochastic finite-sum setting, the degree of agreement between the gradients composing the sum is measured by their average correlation through RACOGA. This quantity captures how well the stochastic estimator aligns with, or deviates from, the full gradient (Chapter 2). In the non-convex PL setting, the correlation between the descent direction and the direction toward a minimizer is captured by the aiming condition. This condition quantifies how momentum contributes to moving toward the solution (Chapter 5). In both cases, these intuitive geometric characterizations directly affect the possibility of acceleration, as they enter the corresponding convergence bounds.

2. *Non-convexity can weaken the classical link between ODEs and algorithms, a difficulty that can be partially addressed by the continuized method.* We highlighted several discrepancies between ODE-based results and algorithmic results, both from the literature and from our own work. These discrepancies suggest limitations of the classical Nesterov momentum algorithm in non-convex settings[1], as discussed in Chapter 3. In Chapter 4, we proposed to use the continuized method from Even et al. (2021) as a potentially general approach to overcome some of these limitations for non-convex acceleration. Perhaps the most striking illustration of its potential is provided in Chapter 6, where this approach allowed us to make significant progress on the open question of whether a vanilla momentum method can achieve the $\mathcal{O}(\varepsilon^{-7/4})$ complexity under Lipschitz gradient and Hessian assumptions[2].

3. *The statement that we obtain accelerated bounds when using Nesterov's momentum when minimizing smooth strongly quasar convex functions is more subtle than what the literature may suggest.* Based on existing results, the question of acceleration under relaxations of convexity might appear essentially settled: acceleration is impossible for smooth PL functions, whereas smooth strongly quasar convex functions exhibit a strongly-convex-like acceleration phenomenon, improving the dependence on the condition number through a square root. In Chapter 5, we showed that the situation is more complex, and that a hidden correlation variable is involved, as discussed in Point 1. More generally, this illustrates the need for caution when interpreting convergence results, especially when drawing analogies between different settings, for instance when comparing quantities of the form $\mu/L$.

[1] We note that the *performance estimation problems* (PEP) framework, a computer-assisted tool to carry worst-case analysis, could provide a definitive answer whether these limitations of Nesterov's momentum algorithm are truly unavoidable.

[2] The recent work of Ushiyama and Sato (2026) casted this problem in the PEP framework, which allowed them to discard the restart mechanism while having a deterministic algorithm. Interestingly, they still require a mechanism that reduces the momentum, depending on the velocity. This may be interpreted as a "soft" restart.

## Limitations and Perspectives

**The gap with the deep-learning practice.** Although we partly motivated our work by the application of training deep neural networks, one should remain cautious about the implications of our findings for this specific problem. Indeed, even though momentum is an important mechanism in widely used modern algorithms, we did not consider the other fundamental aspect: using a *preconditioner*, or a *variable metric*. Doing so amounts to replacing, at each iteration $k$, the descent direction $-\nabla f$ by $-P_k \nabla f$ for some preconditioning matrix $P_k$. Algorithms such as AdaGrad (Duchi et al., 2011) and ADAM (Kingma and Ba, 2015) use a diagonal preconditioner, while recently, non-diagonal preconditioners have become popular, such as in Shampoo (Gupta et al., 2018) and Muon (Jordan et al., 2024). Such preconditioner modify the descent direction, which necessarily heavily impacts the dynamics. Another reason for being cautious is that we carried out our work within the classical $L$-smooth framework. It is both convenient and widely used in the study of first-order optimization algorithms, but its relevance for optimization in deep-learning applications has recently been seriously questioned, most notably through the Edge of Stability phenomenon (Cohen et al., 2021). Recent works consider generalizations of $L$-smoothness that better match the type of smoothness encountered in neural networks, such as $(L_0, L_1)$-smoothness (Zhang et al., 2019a), which assumes that

$$\forall x \in \mathbb{R}^d,\ \left\|\nabla^2 f(x)\right\|_{\text{spec}} \leq L_0 + L_1 \left\|\nabla f(x)\right\|,$$

for $L_0, L_1 \in \mathbb{R}_+$. Such assumptions can be generalized further in several directions, while still allowing Nesterov's momentum to recover accelerated bounds in the deterministic, convex case (Li et al., 2023; Gorbunov et al., 2025). This naturally raises the question of whether the non-convex acceleration results considered in this thesis can be extended to these broader smoothness settings. To the best of our knowledge, in the case of quasar convex functions, only one work (Ablaev et al., 2025) considers such generalized smoothness assumptions, and only for gradient-descent-type algorithms without momentum. Perhaps surprisingly, compared with the two points raised above, we do not view the use of structures such as strong quasar convexity as a major limitation. Of course, loss functions arising in deep learning are certainly unlikely to satisfy such strong structural assumptions globally. However, the neural networks used in practice are also designed to enable efficient training. As discussed several times throughout the manuscript, this is consistent with empirical observations suggesting that their loss landscapes may be "almost convex" around solutions Li et al. (2018), or may satisfy strongly quasar convex-like properties along the optimization path Guille-Escuret et al. (2022, 2024).

**The future of the continuized Nesterov method.** In this thesis, we have shown that the continuized Nesterov momentum method can lead to new accelerated convergence results in non-convex optimization. However, the method has so far received very little attention. This raises the question of whether this is simply because it is recent and somewhat unusual, or because its range of useful applications is intrinsically limited. We see two main motivations for considering this method. The first is to exploit the fact we can perform continuous-time analysis, which may allow one to bypass part of the difficulties of discrete-time Lyapunov analysis. In deterministic optimization, however, obtaining non-deterministic results may be viewed as too high a price to pay for this simplification. The second motivation, which appears more promising to us, is that the continuized method may help overcome limitations of existing methods in achieving accelerated results. It remains unclear how many settings could benefit from the continuized method. We identify at least two.

- **Third order smoothness.** A promising candidate is the class of functions whose first three derivatives are Lipschitz-continuous. In this setting, the $\mathcal{O}(\varepsilon^{-7/4})$ bound obtained under Lipschitz-continuity of the first two derivatives can be improved to $\mathcal{O}(\varepsilon^{-5/3})$ (Carmon et al., 2017). The corresponding algorithm is, however, arguably quite involved, as it combines momentum resets, negative-curvature exploitation, and other mechanisms. The case of functions with Lipschitz-continuous first two derivatives has received significant attention from the viewpoint of simplifying the algorithm (Jin et al., 2018; Li and Lin, 2023; Okamura et al., 2024), with Chapter 6 also contributing to this line of work. The analogous question for functions with Lipschitz-continuous first three derivatives appears to be much less explored. Whether the continuized method can achieve the $\mathcal{O}(\varepsilon^{-5/3})$ bound remains an open and seemingly challenging problem. A first step could be to develop an ODE-based analysis in the spirit of (Okamura et al., 2024). This is already nontrivial, since it is unclear whether the tools developed in that work can be adapted to this higher-order setting.

- **Extending (strongly) quasar convex results.** In quasar convex optimization, the accelerated result based on the subspace-search procedure of (Hinder et al., 2020) has already been extended to non-Euclidean optimization (Lezane et al., 2024) and constrained optimization (Martínez-Rubio, 2025). However, this search procedure can face limitations in some settings. For example, since it relies on first-order information, it is unclear whether it can be used for zeroth-order optimization, where only function values are accessible. In the recent work Gbaguidi and Hermant (2026), we have observed that the continuized framework can be used to design a zeroth-order Nesterov momentum algorithm that achieves an accelerated rate similar to the one obtained in zeroth-order convex optimization (Nesterov and Spokoiny, 2017; Gorbunov et al., 2022). More generally, it is an interesting challenge wether one can fit classical optimization tools into the continuized method, such as proximal updates in the spirit of FISTA (Beck and Teboulle, 2009; Chambolle and Dossal, 2015), mirror steps or adaptive procedures for choosing parameters, and derive new accelerated (strongly) quasar convex results.

**A broad line of research: the scope of (strongly) quasar convex optimization** The following discussion also applies to quasar convex functions, which correspond to strongly quasar convex functions with $\mu = 0$. We identify two directions for future work.

- **New examples.** Some examples of problems that are (strongly) quasar convex were presented in Sections 1.1.2 and 1.4.1, but their number remains rather limited. Although quasar convex functions have received increasing attention in recent years, they are still a relatively new class of functions, see our discussion in Section 1.4.1. We believe that several further applications may still fit within this framework, making this an interesting direction for future research. Some such results may already exist in the literature, especially in the strongly quasar convex case, but under different terminology. Indeed, some classes of functions are essentially equivalent. For instance, strongly quasar convex functions also verify the restricted secant inequality (Zhang and Yin, 2013), also known as one-point convexity, and conversely under $L$-smoothness (Hermant et al., 2026, Lemma 4). Some results may therefore have already been established for the latter class, without being explicitly connected to strong quasar convexity. This is precisely how one can deduce that some generalized linear models are strongly quasar convex (Wang and Wibisono, 2023, Section 3).

- **Extensions of convex optimization.** We have seen that, despite relaxing convexity, (strongly) quasar convex functions seem to retain many algorithmic properties of (strongly) convex functions. A broad and natural research direction is therefore to determine to what extent results from the rich literature about convex optimization can be extended to quasar convex optimization. In other words, if taking a book about convex optimization, a question is what remains of the results if replacing convexity with quasar convexity in each statement. We believe this is case-dependent, and that when it comes to accelerated results, the continuized method may have a role to play. Whenever such an extension is possible at no additional cost, it would suggest that quasar convexity may provide a more appropriate viewpoint than convexity for the results under consideration.

# Bibliography


SS Ablaev, FS Stonyakin, MN Fedotov, MS Alkousa, OS Savchuk, and AV Gasnikov. A gradient method with inexact gradient information: A study on some classes of (l0, l1)-smooth nonconvex problems. *Automation and Remote Control*, 86(9), 2025.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, page 1195–1199, 2017.

David Aldous and James Allen Fill. Reversible markov chains and random walks on graphs, 2002. unfinished monograph, recompiled 2014, 2002.

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective. *arXiv preprint arXiv:2510.03164*, 2025.

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pages 3:1–3:22, 2017.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025a.

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization: Jm altschuler, pa parrilo. *Mathematical Programming*, 213(1):1105–1118, 2025b.

Felipe Alvarez, Hedy Attouch, Jérôme Bolte, and Patrick Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Vassilis Apidopoulos, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 187 (1):151–193, 2021.

Vassilis Apidopoulos, Nicolo Ginatta, and Silvia Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under polyak–łojasiewicz condition. *Journal of Global Optimization*, 84(3):563–589, 2022.

Vassilis Apidopoulos, Vasiliki Mavrogeorgou, and Theodoros G Tsironis. Heavy-ball dynamics with hessian-driven damping for non-convex optimization under the łojasiewicz condition. *arXiv preprint arXiv:2506.11705*, 2025.

Kenneth J Arrow and Alain C Enthoven. Quasi-concave programming. *Econometrica: Journal of the Econometric Society*, pages 779–800, 1961.

Mahmoud Assran and Michael Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Hedy Attouch and Juan Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than 1/k^2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2022.

Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. Convergence of iterates for first-order optimization algorithms with inertia and hessian driven damping. *Optimization*, 72(5):1199–1238, 2023. doi: 10.1080/02331934.2021.2009828. URL https://doi.org/10.1080/02331934.2021.2009828.

Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization. *SIAM Journal on Optimization*, 32(3):1817–1842, 2022.

Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of the Heavy-Ball method under the łojasiewicz property. *Mathematical Programming*, 198(1):195–254, 2023a.

Jean-François Aujol, Charles Dossal, and Aude Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. *Mathematical Programming*, pages 1–43, 2023b.

Jean-François Aujol, Luca Calatroni, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Parameter-free fista by adaptive restart and backtracking. *SIAM Journal on Optimization*, 34(4):3259–3285, 2024a.

Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Strong convergence of fista iterates under hölderian and quadratic growth conditions. *arXiv preprint arXiv:2407.17063*, 2024b.

Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Heavy Ball momentum for non-strongly convex optimization. *arXiv preprint arXiv:2403.06930*, 2024c.

Jean-François Aujol and Charles Dossal. Stability of over-relaxations for the forward-backward algorithm, application to FISTA. *SIAM Journal on Optimization*, 25(4):2408–2433, 2015.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Rishab Balasubramanian, Rupashree Dey, and Kunal Rathore. Contrastive learning for ood in object detection, 2022. URL https://arxiv.org/abs/2208.06083.

Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on learning theory*, pages 361–382. PMLR, 2016.

Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. On global convergence of ResNets: From finite to infinite width using linear parameterization. *Advances in Neural Information Processing Systems*, 35:16385–16397, 2022.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974997.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 2006.

Yoshua Bengio, Nicolas Le Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

J Frédéric Bonnans, Roberto Cominetti, and Alexander Shapiro. Sensitivity analysis of optimization problems under second order regular constraints. *Mathematics of Operations Research*, 23(4):806–831, 1998.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.

Nicolas Boumal, Christopher Criscitiello, and Quentin Rebjock. Smooth, globally polyak-{\L} ojasiewicz functions are nonlinear least-squares. *arXiv preprint arXiv:2604.07972*, 2026.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jingjing Bu and Mehran Mesbahi. A note on nesterov's accelerated method in nonconvex optimization: a weak estimate sequence approach. *arXiv preprint arXiv:2006.08548*, 2020.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Herbert Busemann. Note on a theorem on convex sets. *Matematisk Tidsskrift. B*, pages 32–34, 1947.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021a.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021b.

Coralia Cartis, Nicholas IM Gould, and Ph L Toint. On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6):2833–2852, 2010.

Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31, 2021.

Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

Volkan Cevher and Bang Cong Vu. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019.

Antonin Chambolle and Charles Dossal. On the convergence of the iterates of" fista". *Journal of Optimization Theory and Applications*, 166(3):25, 2015.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Patrick Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212, 2011.

Yaim Cooper. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670. PMLR, 2023.

Aris Daniilidis and Nicolas Hadjisavvas. On the subdifferentials of quasiconvex and pseudoconvex functions and cyclic monotonicity. *Journal of mathematical analysis and applications*, 237(1):30–42, 1999.

Marina Danilova, Anastasiia Kulakova, and Boris Polyak. Non-monotone behavior of the heavy ball method. In *International Conference on Difference Equations and Applications*, pages 213–230. Springer, 2018.

Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pages 79–163. Springer, 2022.

Aaron Defazio. On the curved geometry of accelerated optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends in Optimization*, 5(1-2):1–245, 2021.

Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendrikx, Pierre Gaillard, Laurent Massoulié, and Adrien Taylor. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. *Advances in Neural Information Processing Systems*, 34:28054–28066, 2021.

Zixu Feng and Hao Yuan. Optimal local linear convergence of nesterov's accelerated gradient method for $\hat{c}2$ functions under the polyak–{\L} ojasiewicz inequality. *arXiv preprint arXiv:2603.21516*, 2026.

Simon Fitzpatrick and Robert R Phelps. Differentiability of the metric projection in hilbert space. *Transactions of the American Mathematical Society*, 270(2):483–501, 1982.

Qiang Fu, Dongchu Xu, and Ashia Camage Wilson. Accelerated stochastic optimization methods under quasar-convexity. In *International Conference on Machine Learning*, pages 10431–10460. PMLR, 2023.

Yuan Gao, Anton Rodomanov, and Sebastian U Stich. Non-convex stochastic composite optimization with polyak momentum. *arXiv preprint arXiv:2403.02967*, 2024.

Guillaume Garrigos. Square distance functions are Polyak-Łojasiewicz and vice-versa. *arXiv preprint arXiv:2301.10332*, 2023.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Eméric Gbaguidi and Julien Hermant. Accelerated stochastic zeroth-order quasar-convex optimization. *arXiv preprint arXiv:2607.19965*, 2026.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International conference on machine learning*, pages 1233–1242. PMLR, 2017.

Benjamin Gess and Sebastian Kassing. Exponential convergence rates for momentum stochastic gradient descent in the overparametrized setting. *arXiv e-prints*, pages arXiv–2302, 2023.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *SIAM Journal on Optimization*, 32(2):1210–1238, 2022.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex $(l_0, l_1)$-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, volume 2025, pages 97261–97305, 2025.

Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method: B. goujaud et al. *Mathematical Programming*, pages 1–59, 2025.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

Sorin-Mihai Grad, Felipe Lara, and Raúl T Marcavillaca. Strongly quasiconvex functions: what we know (so far). *Journal of Optimization Theory and Applications*, 205(2):38, 2025.

Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

Charles Guille-Escuret, Adam Ibrahim, Baptiste Goujaud, and Ioannis Mitliagkas. Gradient descent is optimal under lower restricted secant inequality and upper error bound. *Advances in Neural Information Processing Systems*, 35:24893–24904, 2022.

Charles Guille-Escuret, Hiroki Naganuma, Kilian Fatras, and Ioannis Mitliagkas. No wrong turns: The simple geometry of neural networks optimization paths. In *International Conference on Machine Learning*, pages 16751–16772. PMLR, 2024.

Sergey Guminov, Alexander Gasnikov, and Ilya Kuruzov. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science*, 20(1):36, 2023.

Huiyuan Guo, Juan José Maulén, and Juan Peypouquet. A speed restart scheme for a dynamical system with hessian-driven damping and three constant coefficients. *arXiv preprint arXiv:2412.06691*, 2024.

Kanan Gupta and Stephan Wojtowytsch. Nesterov acceleration in benignly non-convex landscapes. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kanan Gupta, Jonathan W Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. doi: 10.1016/0095-0696(78)90006-2.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846. URL `https://books.google.fr/books?id=eBSgoAEACAAJ`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Julien Hermant, Jean-François Aujol, Charles Dossal, Lorick Huang, and Aude Rondepierre. Continuized nesterov acceleration for non-convex optimization. *arXiv preprint arXiv:2512.16533*, 2025a.

Julien Hermant, Marien Renaud, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Gradient correlation is a key ingredient to accelerate sgd with momentum. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Julien Hermant, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Study of the behaviour of nesterov accelerated gradient in a non convex setting: the strongly quasar convex case. *Optimization*, 0 (0):1–26, 2026. doi: 10.1080/02331934.2026.2680423. URL `https://doi.org/10.1080/02331934.2026.2680423`.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, pages 1894–1938. PMLR, 2020.

Geoffrey Hinton. Neural networks for machine learning. Coursera Lecture 6e, 2012. Available at `https://www.coursera.org/learn/neural-networks`.

Richard B Holmes. Smoothness of certain metric projections on hilbert space. *Transactions of the American Mathematical Society*, 184:87–100, 1973.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR, 2022.

Børge Jessen. To saetninger om konvekse punktmaengder. *Matematisk tidsskrift. B*, pages 66–70, 1940.

Ruichen Jiang, Aryan Mokhtari, and Francisco Patitucci. Improved complexity for smooth nonconvex optimization: a two-level online learning approach with quasi-newton methods. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2225–2236, 2025.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Sebastian Kassing and Simon Weissmann. Polyak's heavy ball method achieves accelerated local rate of convergence under polyak-lojasiewicz inequality. *arXiv preprint arXiv:2410.16849*, 2024.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Abbas Khademi. The convexity zoo: a taxonomy of function classes in optimization. *Optimization*, 0(0):1–52, 2026. doi: 10.1080/02331934.2026.2672525. URL `https://doi.org/10.1080/02331934.2026.2672525`.

Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159:81–107, 2016.

Jungbin Kim and Insoon Yang. Unifying Nesterov's accelerated gradient methods for convex and strongly convex objective functions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 16897–16954. PMLR, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

N Kritikos. Sur quelques propriétés des ensembles convexes. *Bulletin mathématique de la Société Roumaine des Sciences*, 40(1/2):87–92, 1938.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jasper CH Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614. IEEE, 2016.

Jongmin Lee, Chanwoo Park, and Ernest Ryu. A geometric structure of acceleration and its role in making gradients small fast. *Advances in Neural Information Processing Systems*, 34:11999–12012, 2021.

Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065. IEEE, 2015.

Clement Lezane, Sophie Langer, and Wouter M Koolen. Accelerated mirror descent for non-euclidean star-convex functions. *arXiv preprint arXiv:2405.18976*, 2024.

Bowen Li, Bin Shi, and Ya-Xiang Yuan. Linear convergence of forward-backward accelerated algorithms without knowledge of the modulus of strong convexity. *SIAM Journal on Optimization*, 34(2):2150–2168, 2024.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023.

Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the in the o (epsilonˆ(-7/4)) complexity. *Journal of Machine Learning Research*, 24(157):1–37, 2023.

Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. In *International Conference on Learning Representations*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023.

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

Stanisław Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier. Université de Grenoble*, 43(5):1575–1595, 1993. ISSN 0373-0956.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020.

Olvi L Mangasarian. Pseudo-convex functions. In *Stochastic optimization models in finance*, pages 23–32. Elsevier, 1975.

David Martínez-Rubio. Smooth quasar-convex optimization with constraints. *arXiv preprint arXiv:2510.01943*, 2025.

Naoki Marumo and Akiko Takeda. Parameter-free accelerated gradient descent for nonconvex minimization. *SIAM J. Optim.*, 34:2093–2120, 2022. URL `https://api.semanticscholar.org/CorpusID:254591346`.

Naoki Marumo and Akiko Takeda. Universal heavy-ball method for nonconvex optimization under hölder continuous hessians. *Mathematical Programming*, 212(1):147–175, 2025.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. In *Conference on learning theory*, pages 1476–1515. PMLR, 2017.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer Science & Business Media, 1986.

P. T. Mocanu, I. Şerb, and G. Toader. Real star-convex functions. *Studia Universitatis Babeş-Bolyai, Mathematica*, 42(3):65–80, 1997.

Adel Nabli and Edouard Oyallon. Decentralized asynchronous optimization with dadao allows decoupling and acceleration. *Journal of Machine Learning Research*, 26(207):1–48, 2025. URL `http://jmlr.org/papers/v26/23-0748.html`.

Adel Nabli, Eugene Belilovsky, and Edouard Oyallon. $A^2CiD^2$: Accelerating Asynchronous Communication in Decentralized Deep Learning. *Advances in Neural Information Processing Systems*, 36:47451–47474, 2023.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii 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.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4): 773–810, 2021.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Kaito Okamura, Naoki Marumo, and Akiko Takeda. Heavy-ball differential equation achieves $o(\varepsilon^{-7/4})$ convergence for nonconvex functions. *arXiv preprint arXiv:2406.06100*, 2024.

Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Michael O'Neill and Stephen J Wright. Behavior of accelerated gradient methods near critical points of nonconvex functions. *Mathematical Programming*, 176(1):403–427, 2019.

RR Phelps. Convex sets and nearest points. *Proceedings of the American Mathematical Society*, 8(4):790–797, 1957.

René Poliquin, R Rockafellar, and Lionel Thibault. Local differentiability of distance functions. *Transactions of the American mathematical Society*, 352(11):5231–5249, 2000.

B. T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. doi: 10.1016/0041-5553(63)90382-3.

Boris Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL `https://www.sciencedirect.com/science/article/pii/0041555364901375`.

Boris Polyak and Pavel Shcherbakov. Lyapunov functions: An optimization theory perspective. *IFAC-PapersOnLine*, 50(1):7456–7461, 2017.

Boris T Polyak. Introduction to optimization. 1987.

Philip E Protter. Stochastic differential equations. In *Stochastic integration and differential equations*, pages 249–361. Springer, 2012.

Marien Renaud, Julien Hermant, Deliang Wei, and Yu Sun. Provably accelerated imaging with restarted inertia and score-based image priors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=8pQsiFyTQi`.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971. ISBN 978-0-12-604550-5.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM journal on imaging sciences*, 10(4):1804–1844, 2017.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. *Learning representations by back-propagating errors*, volume 323. Nature Publishing Group, 1986.

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *International conference on machine learning*, pages 8469–8479. PMLR, 2020.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

Alexander Shapiro. Differentiability properties of metric projections onto convex sets. *Journal of Optimization Theory and Applications*, 169(3):953–964, 2016.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2022.

Jonathan W Siegel. Accelerated first-order methods: Differential equations and lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

Sidak Pal Singh, Bobby He, Thomas Hofmann, and Bernhard Schölkopf. The directionality of optimization trajectories in neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mikhail V Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. pmlr, 2013.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1):557–594, 2023.

GH Toader. On the hierarchy of convexity of functions. *Mathematica-Revue d'analyse numérique et de théorie de l'approximation. L'analyse numérique et la théorie de l'approximation*, 15(2):167–172, 1986.

Gheorghe Toader. Some generalizations of the convexity. In *Proceedings of the Colloquium on Approximation and Optimization*, pages 329–338, Cluj-Napoca, Romania, 1984.

Silvia Toader. The order of a star-convex function. *Bulletin of Applied and Computational Mathematics*, 85-B:347–350, 1998. BAM-1473.

Paul Tseng. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

Kansei Ushiyama and Shun Sato. A restart-free accelerated algorithm for non-convex minimization: Continuous and discrete analysis. *arXiv preprint arXiv:2606.30050*, 2026.

Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996.

Aditya Varre and Nicolas Flammarion. Accelerated sgd for non-strongly-convex least squares. In *Conference on Learning Theory*, pages 2062–2126. PMLR, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Jun-Kun Wang and Andre Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *Advances in neural information processing systems*, 31, 2018.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1998–2007. PMLR, 2018.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing Polyak-Łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2948–2968. PMLR, 2023.

Eduardo H. Zarantonello. Projections on convex sets in hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In Eduardo H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, pages 237–424. Academic Press, 1971.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019a.

Jingzhao Zhang, Suvrit Sra, and Ali Jadbabaie. Acceleration in first order quasi-strongly convex optimization by ode discretization. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1501–1506. IEEE, 2019b.

Dongruo Zhou. Sharp first-order lower bounds for higher-order smooth nonconvex optimization. *arXiv preprint arXiv:2606.05438*, 2026.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BylIciRcYQ`.