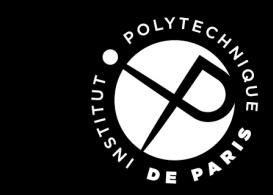
Heavy-tailed nature of stochastic gradient descent in deep learning : theoretical and empirical analysis

Thanh Huy Nguyen



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT003

Thèse de doctorat

Heavy-tailed Nature of Stochastic Gradient Descent in Deep Learning: Theoretical and Empirical Analysis

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique, Données et Intelligence Artificielle

Thèse présentée et soutenue à Palaiseau, le 21/01/2021, par

THANH HUY NGUYEN

Composition du Jury :

Olivier Cappé DR CNRS	Président
Zaid Harchaoui Associate Prof., Univ. Washington	Rapporteur
Lenka Zdeborova Associate Prof. EPFL, HDR	Rapporteure
Alain Durmus MdC ENS-Saclay	Examinateur
Ali Taylan Cemgil Prof. Bogazici University, Istanbul, Turkey	Examinateur
Gaël Richard Prof. Télécom Paris	Directeur de thèse
Umut Şimşekli CR INRIA	Co-directeur de thèse

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisors Assoc. Prof. Umut Şimşekli and Prof. Gaël Richard for the continuous support and help during my PhD, and for their patience, motivation, and immense knowledge. Without their guidance, I would not have been able to complete this research and the writing of this thesis.

Besides my advisors, I would like to thank the rest of my thesis committee members: Dr. Olivier Cappé, Assoc. Prof. Zaid Harchaoui, Assoc. Prof. Lenka Zdeborova, Assoc. Prof. Alain Durmus and Prof. Ali Taylan Cemgil for their insightful comments, valuable suggestions and encouragement.

My sincere thanks also go to all the members of the LTCI lab at Télécom Paris, who provided me an opportunity to join a big team with great researchers, and who gave access to a welcoming and professional research environment.

I would like to offer my special thanks to my family for their unwavering support and belief in me. Without their precious support, I would not have made it through my PhD journey.

Finally, I would like to thank the French National Research Agency (ANR) and the industrial chair Data Science & Artificial Intelligence from Télécom Paris. This thesis is supported by the ANR as a part of the FBIMATRIX (ANR-16-CE23-0014) project, and by the industrial chair Data Science & Artificial Intelligence from Télécom Paris.

Abstract

In this thesis, we are concerned with the stochastic gradient descent (SGD) algorithm, which has been widely used in machine learning due to its computational efficiency and favorable generalization properties. Specifically, we perform theoretical and empirical analysis of the behavior of the stochastic gradient noise (GN) in deep neural networks, which is defined as the difference between the true gradient and the stochastic gradient. Based on these results, we bring an alternative perspective to the existing approaches for investigating SGD. The GN in SGD is often considered to be Gaussian for mathematical convenience. This assumption enables SGD to be studied as a stochastic differential equation (SDE) driven by a Brownian motion. We argue that the Gaussianity assumption might fail to hold in deep learning settings and hence render the Brownian motion-based analyses inappropriate. Inspired by non-Gaussian natural phenomena, we consider the GN in a more general context that suggests that the GN is better approximated by a *heavy-tailed* α-stable random vector, where *tail-index* α decides the heavy-tailedness of the distribution. Accordingly, we propose to analyze SGD as a discretization of an SDE driven by a Lévy motion. *Firstly*, to justify the α-stable assumption, we conduct experiments on common deep learning scenarios and show that in all settings, the GN is highly non-Gaussian and exhibits heavy-tails. We investigate the tail behavior in various network architectures and sizes, loss functions, and datasets. *Secondly*, under the heavy-tailed GN assumption, we provide a non-asymptotic analysis for the discrete-time dynamics to converge to the global minimum in terms of suboptimality. This finite-time guarantee is then extended to application in posterior sampling. *Finally*, we investigate the metastability nature of the SDE driven by Lévy motion that can then be exploited for clarifying the behavior of SGD, especially in terms of 'preferring wide minima'. More precisely, we provide formal theoretical analysis where we derive explicit conditions for the step-size such that the metastability behavior of SGD, viewed as a discrete-time SDE, is similar to its continuous-time limit. We show that the behaviors of the two systems are indeed similar for small step-sizes and we describe how the error depends on the algorithm and problem parameters. We illustrate our metastability results with simulations on a synthetic model and neural networks. Our results open up a different perspective and shed more light on the view that SGD prefers wide minima.

Résumé

Dans cette thèse, nous nous intéressons à l'algorithme du gradient stochastique (SGD), qui est largement utilisé en apprentissage automatique en raison de son efficacité de calcul et de ses propriétés de généralisation. SGD est applicable à un large ensemble de problèmes d'optimisation convexe et non convexe survenant dans l'apprentissage automatique, y compris l'apprentissage profond où ils sont particulièrement réussis. Dans cette étude, nous effectuons une analyse théorique et empirique du comportement du bruit de gradient stochastique (GN), qui est défini comme la différence entre le gradient réel et le gradient stochastique, dans les réseaux de neurones profonds. Sur la base de ces résultats, nous apportons une perspective alternative aux approches existantes pour étudier SGD. Le bruit de gradient stochastique dans SGD est souvent considéré comme gaussien dans le régime des données volumineuses en supposant que le théorème limite central classique (CLT) entre en jeu. Cette hypothèse est souvent faite pour des raisons de commodité mathématique, car elle permet d'étudier SGD comme une équation différentielle stochastique (SDE) pilotée par un mouvement brownien. Nous soutenons que l'hypothèse de la gaussianité pourrait ne pas tenir dans les contextes d'apprentissage profond et donc rendre inappropriées les analyses basées sur le mouvement brownien. Inspiré de phénomènes naturels non gaussiens, nous considérons le bruit de gradient stochastique dans un contexte plus général et invoquons le théorème limite central généralisé, qui suggère que le bruit de gradient stochastique est mieux approché par un vecteur aléatoire *à queue lourde* α-stable, où l'indice de queue α décide de la lourdeur de la distribution. En conséquence, nous proposons d'analyser SGD comme une discrétisation d'une équation différentielle stochastique pilotée par un mouvement Lévy. Cette équation différentielle stochastique peut subir des "sauts", qui obligent l'équation différentielle stochastique et son transition de discrétisation de minima étroits vers des minima plus larges, comme le prouve la théorie existante de la métastabilité.

Premièrement, pour justifier l'hypothèse α-stable, nous menons des expériences sur des scénarios communs d'apprentissage en profondeur et montrons que dans tous les contextes, le bruit de gradient stochastique est hautement non gaussien et présente des queues lourdes. Nous étudions le comportement de la queue dans diverses architectures et tailles de réseau, fonctions de perte et ensembles de données.

Deuxièmement, sous l'hypothèse du bruit de gradient stochastique à queue lourde, nous fournissons une analyse non asymptotique pour que la dynamique en temps discret SGD converge vers le minimum global en termes de sous-optimalité. Nos résultats montrent que l'erreur faible sous notre hypothèse augmente plus rapidement que sous l'hypothèse du bruit gaussien, ce qui suggère d'utiliser des tailles de pas plus petits dans notre cas. Cette garantie à temps fini est ensuite étendue à l'application en postérieur échantillonnage.

Enfin, nous étudions la nature de métastabilité de l'équation différentielle stochastique pilotée par le mouvement de Lévy qui peut ensuite être exploitée pour clarifier le

comportement de SGD, notamment en termes de "préférence de larges minima". Bien que notre approche apporte une nouvelle perspective pour l'analyse de SGD, elle est limitée en ce sens qu'en raison de la discrétisation temporelle, SGD peut admettre un comportement sensiblement différent de sa limite de temps continu. Dans ce sujet, nous fournissons une analyse théorique formelle où nous dérivons des conditions explicites pour la taille de pas de sorte que le comportement de métastabilité de SGD, considéré comme une équation différentielle stochastique en temps discret, est similaire à sa limite de temps continu. Nous montrons que les comportements des deux systèmes sont en effet similaires pour les petits tailles de pas et nous décrivons comment l'erreur dépend de l'algorithme et des paramètres du problème. Nous illustrons nos résultats de métastabilité par des simulations sur un modèle synthétique et des réseaux de neurones. Nos résultats ouvrent une perspective différente et éclairent davantage l'idée selon laquelle SGD préfère les minima larges.

Contents

Chapter 1

Introduction

Machine learning is the study of computer algorithms that instruct a system how to improve from experience. The primary goal is to allow the system to learn automatically to perform predictions, decisions or classifications without being explicitly programmed. In order to carry out these assigned tasks, machine learning algorithms construct mathematical models using sample data. The more complex the tasks are, the more data is needed for a good performance of the learning process. With the accessibility of large amounts of data (images, speech, video) in recent years, the interest in machine learning has been increasing rapidly. Another even more important reason that makes machine learning successful nowadays is the availability of modern technologies such as parallel-processing power, high-performance graphics processing units, which can be combined in clusters to significantly reduce the training time for a learning model.

Machine learning systems are present all around us. They are exploited in a broad variety of applications, for example: computer vision Khan and Al-Habsi [2020], audio recognition Purwins et al. [2019], Cunningham et al. [2020], speech recognition Padmanabhan and Johnson Premkumar [2015], natural language processing Shetty [2018], machine translation Popel et al. [2020], bioinformatics Larranaga et al. [2006], material inspection Sacco et al. [2020], medical image analysis Lundervold and Lundervold [2019], drug design Vamathevan et al. [2019] or board game programs Xenou et al. [2018]. In some cases, they have achieved outcomes as good as and in some circumstances surpassing human expert performance.

Deep learning is a machine learning technique that employs neural networks with a large number of layers for training, using huge amounts of data. Deep neural networks have revolutionized machine learning and have ubiquitous use in many application domains [LeCun et al., 2015, Krizhevsky et al., 2012, Hinton et al., 2012a]. There are different types of neural networks that are suitable for different tasks: convolutional neural networks (CNNs) are appropriate for image recognition, while recurrent neural networks (RNNs) are suited for language processing and speech recognition due to their

better capabilities to model time series. The architecture of neural networks is also developing through time: researchers design a more efficient type of RNN model called long short-term memory (LSTM), making it run fast enough to be applied in on-demand applications like Google Translate.

In full generality, many key tasks in deep learning reduce to solving the following optimization problem:

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \Big\{ f(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^{(i)}(\mathbf{w}) \Big\} \tag{1.1}$$

where $\mathbf{w} \in \mathbb{R}^d$ denotes the weights of the neural network, $f : \mathbb{R}^d \to \mathbb{R}$ denotes the loss function that is typically non-convex in $\mathbf{w}$, each $f^{(i)}$ denotes the (instantaneous) loss function that is contributed by the *data point* $i \in \{1, \ldots, n\}$, and n denotes the total number of data points. Stochastic gradient descent (SGD) is one the most popular approaches for attacking this problem in practice and is based on the following iterative updates:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta \nabla \tilde{f}_k(\mathbf{w}^k) \tag{1.2}$$

where $k \in \{1, \ldots, K\}$ denotes the iteration number, η is the step-size (or the learning rate), and $\nabla \tilde{f}_k$ denotes the stochastic gradient at iteration k, that is defined as follows:

$$\nabla \tilde{f}_k(\mathbf{w}) \triangleq \nabla \tilde{f}_{\Omega_k}(\mathbf{w}) \triangleq \frac{1}{b}\sum_{i\in\Omega_k} \nabla f^{(i)}(\mathbf{w}). \tag{1.3}$$

Here, $\Omega_k \subset \{1, \ldots, n\}$ is a random subset that is drawn with or without replacement at iteration k, and $b = |\Omega_k|$ denotes the number of elements in Ω_k.

SGD is widely used in deep learning with a great success in its computational efficiency [Bottou, 2010, Bottou and Bousquet, 2008, Daneshmand et al., 2018]. Beyond efficiency, understanding how SGD performs better than its full batch counterpart in terms of test accuracy remains a major challenge. Even though SGD seems to find perfect training performance at (near-) zero loss solutions on the training landscape (at least in certain regimes [Zhang et al., 2017a, Sagun et al., 2015, Keskar et al., 2016, Geiger et al., 2018]), it appears that the algorithm finds solutions with different properties depending on how it is tuned [Sutskever et al., 2013, Keskar et al., 2016, Jastrzebski et al., 2017, Hoffer et al., 2017, Masters and Luschi, 2018, Smith et al., 2017]. Despite the fact that the impact of SGD on generalization has been studied [Advani and Saxe, 2017, Wu et al., 2018, Neyshabur et al., 2017], a satisfactory theory that can explain its success in a way that encompasses such peculiar empirical properties is still lacking.

A popular approach for investigating the behavior of SGD is based on considering SGD as a discretization of a continuous-time process [Mandt et al., 2016, Jastrzebski et al., 2017, Li et al., 2017, Hu et al., 2017, Zhu et al., 2018, Chaudhari and Soatto, 2018]. This approach models the stochastic gradient noise, i.e., $\nabla \tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})$ as a Gaussian

distribution. From this perspective, the SGD recursion can be seen as a first-order Euler-Maruyama discretization of the Langevin dynamics [Li et al., 2017, Jastrzebski et al., 2017, Hu et al., 2017], which is often referred to as the Unadjusted Langevin Algorithm (ULA) [Roberts and Stramer, 2002, Lamberton and Pages, 2003, Durmus and Moulines, 2015, Durmus et al., 2016]. Based on this observation, Jastrzebski et al. [2017] focus on the relation between this invariant measure and the algorithm parameters, namely the step-size η and minibatch size. They conclude that the ratio of step-size divided by the batch size is the control parameter that determines the width of the minima found by SGD. Furthermore, they revisit the famous wide minima folklore [Hochreiter and Schmidhuber, 1997]: Among the minima found by SGD, the wider it is, the better it performs on the test set.

However, we will identify several fundamental issues with this approach Şimşekli et al. [2019], which show that the gradient noise in SGD can be highly non-Gaussian. Inspired by non-Gaussian natural phenomena, our main approach will be to model the gradient noise by using a more general family of heavy-tailed distributions, which contains the Gaussian distribution as a special case.

This thesis aims to investigate the behavior of SGD by inheriting the idea of stochastic differential equations (SDE). Our general approach can be summarized by the following points:

- By assuming the gradient noise to be Gaussian, SGD is often studied as an approximation of an SDE driven by a Brownian motion. However, by providing empirical evidence on deep learning settings, we show that the gradient noise is highly non-Gaussian. Also according to these experiments, we come up with a better-suited hypothesis for the gradient noise - the α-stable noise. This noise assumption will enable us to use a Lévy-driven SDE as a proxy to SGD.
- Next, we investigate the global convergence of SGD by using our new hypothesis. In particular, by using heavy-tailed assumption, we establish a non-asymptotic guarantee for SGD to converge to the global optimum.
- We conclude the thesis by studying the metastable behavior of SGD, which is often considered as an aspect for explaining its generalization property in deep learning settings. Especially, we derive explicit conditions for the step-size such that the metastability of discrete-time SGD is similar to its continuous-time limit.

1.1 Summary of the thesis

The main contributions of this thesis are twofold: (i) we perform theoretical and empirical analysis of the tail-index of the stochastic gradient noise in deep neural networks and (ii) based on these results, we bring an alternative perspective to the existing approaches for analyzing SGD and shed more light on the folklore that SGD prefers wide minima by establishing a bridge between SGD and the related theoretical results from

statistical physics and stochastic analysis.

The thesis is organized as follows. In Chapter 2 we provide the technical background for Brownian motions, the α-stable distributions as well as the Lévy motions. We briefly introduce the notion of stochastic differential equations (SDEs) and the basic tools such as Wasserstein distance and total variation, which are needed for the theoretical analysis of Fractional Langevin Monte Carlo methods and the metastability analysis.

In Chapter 3, we present the related work that concerns some important aspects in deep learning and diffusion-based Markov chain Monte Carlo.

In Chapter 4 we first formalize the framework in which we analyze SGD by using such SDEs as a proxy. We then describe in Section 4.4 the metastability and first exit time properties of such SDEs and their discretization, and discuss their connection with the wide minima phenomenon. In Section 4.5 we describe our experimental methodology. In Section 4.6 we provide our empirical results which validate our theory. We conduct experiments on the most common deep learning architectures. In particular, we investigate the tail behavior under fully-connected and convolutional models using negative log likelihood (NLL) and linear hinge loss functions on MNIST, CIFAR10, and CIFAR100 datasets. For each configuration, we scale the size of the network and batch size used in SGD and monitor the effect of each of these settings on the tail index α. In particular, we present results on stability tests (Section 4.6.1), finer-grained layer-wise tail-index estimation (sections 4.6.2 and 4.6.3), and an investigation of the relation between the tail-index and the generalization properties of the network (4.6.4).

In Chapter 5, we investigate the global convergence property of SGD for non-convex optimization via a stochastic process, which can be seen as a perturbed version of the gradient descent algorithm with heavy-tailed α-stable noise. In Section 5.2, we state the assumptions that imply the main result presented in Theorem 7, which provides a finite-time guarantee for the discrete-time dynamics of SGD, in terms of suboptimality with respect to the global minimum, as a function of the step-size and the scale parameter. We then describe in Section 5.3 the proof strategy of Theorem 7. Finally in Section 5.5, we extend our results to the case where exact gradients are replaced by stochastic gradients and show that similar results hold in this setting as well. All the proofs of the results in the chapter are given in appendix.

In Chapter 6, we derive explicit conditions for the step-size such that the discrete-time SGD (6.3) can inherit the metastability behavior of its continuous-time limit (6.2). In sections 6.2 and 6.3, we in turn present the main theorem of the chapter and describe the proof overview of the theorem. Our theoretical result is illustrated by numerical experiments on a synthetic model and neural networks in Section 6.4.

Our experiments in Chapter 4 reveal several remarkable results:

- In all our configurations, the stochastic gradient noise turns out to be highly non-Gaussian and possesses a heavy-tailed behavior.
- There is a strong interaction between the network architecture, network size, dataset,

and the tail-index, which ultimately determine the dynamics of SGD on the training surface. This observation supports the view that, the geometry of the problem and the dynamics induced by the algorithm cannot be separated from each other.

- In almost all configurations, we observe two distinct phases of SGD throughout iterations. During the first phase, the tail-index rapidly decreases and SGD possesses a clear jump when the tail-index is at its lowest value and causes a sudden jump in the accuracy. This behavior strengthens the view that SGD crosses barriers at the very initial phase.

Our approach also opens up several interesting future directions and open questions, as we discuss in Chapter 7.

1.2 List of publications

In this section, we present the list of works that have been realized during this PhD thesis. First, we specify the publications associated directly to the central topic of this document:

- Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., & Sagun, L. *On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks.* Submitted to Journal of Machine Learning Research, 2019. Under revision.
  This work is mainly presented in Chapter 4 of this document.
- Nguyen, T. H., Şimşekli, U., & Richard, G. *Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization.* International Conference on Machine Learning, 2019.
  This work is mainly presented in Chapter 5 of this document.
- Nguyen, T. H., Şimşekli, U., Gürbüzbalaban, M., & Richard, G. *First Exit Time Analysis of Stochastic Gradient Descent Under Heavy-Tailed Gradient Noise.* Neural Information Processing Systems, 2019.
  This work is mainly presented in Chapter 6 of this document.

Some other problems have also been dealt with within the time frame of the PhD however they are not included in this document in order to maintain a coherent manuscript. These works are listed below.

- Nguyen, T. H., Şimşekli, U., Richard, G., & Cemgil, A. T. *Efficient Bayesian Model Selection in PARAFAC via Stochastic Thermodynamic Integration.* IEEE Signal Processing Letters, 2018.
  In this work, we develop a novel parallel and distributed Bayesian model selection technique for rank estimation in a large-scale tensor factorization model, called the Parallel factor analysis (PARAFAC). The proposed approach integrates ideas from stochastic gradient MCMC, statistical physics, and distributed stochastic optimization. Our method has a clear mathematical interpretation, and has significantly lower computational requirements, thanks to data sub-sampling and parallelization.

- Şimşekli, U., Yıldız, C., Nguyen, T. H., Richard, G., & Cemgil, A. T. *Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization.* International Conference on Machine Learning, 2018.

  In this work, we develop an asynchronous-parallel stochastic L-BFGS algorithm for non-convex optimization. The proposed algorithm is suitable for both distributed and shared-memory settings. We provide formal theoretical analysis and show that the proposed algorithm provides a significant speedup over the recently proposed synchronous distributed L-BFGS algorithm.

Chapter 2

Elements of notations and definitions

In this chapter, we introduce basic notations and the definitions of the Wasserstein distance and the total variation. Then, we provide some technical backgrounds for stochastic differential equations driven by Brownian motions as well as by the Lévy motions.

2.1 Basic notations

For $z > 0$, we denote $\Gamma(z)$ as the gamma function at z:

$$\Gamma(z) \triangleq \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x.$$

For any Borel probability measures μ and ν with domain Ω, their total variation (TV) distance is defined as follows:

$$\|\mu - \nu\|_{TV} \triangleq 2 \sup_{A \in \mathcal{B}(\Omega)} |\mu(A) - \nu(A)|,$$

where $\mathcal{B}(\Omega)$ denotes the Borel subsets of Ω.

We use $< \cdot, \cdot >$ to denote the inner product between two vectors, $\| \cdot \|$ denotes the Euclidean norm, $\mathbb{E}_\omega[\cdot]$ denotes the expectation with respect to the random variable ω, and $\mathbb{E}[\cdot]$ denotes the expectation with respect to all the random sources. We will use the Wasserstein metric to quantify the distance between two probability measures.

Definition 1 (Wasserstein distance)**.** *Let μ and ν be two probability measures. For $\lambda \geq 1$, we define the λ-Wasserstein distance between μ and ν as follows:*

$$\mathcal{W}_\lambda(\mu, \nu) \triangleq (\inf\{\mathbb{E}\|V - W\|^\lambda : V \sim \mu, W \sim \nu\})^{1/\lambda},$$

where the infimum is taken over all the couplings of μ and ν (i.e. the joint probability distributions whose marginal distributions are μ and ν).

Notational convenience: In this thesis, depending on certain circumstances, we use $f(t)$ (variable t is in parentheses) as well as f_t (variable t is a subscript) to denote a function of t. Accordingly, we use $Z(t)$ as well as Z_t to denote a stochastic process.

2.2 Stochastic differential equations

Stochastic differential equations (SDEs) play an important role in modeling various phenomena in physics, biology and finance, such as thermal fluctuations, fluid flow or stock prices. An SDE is a differential equation in which one or more coefficients are stochastic processes or stochastic functions of stochastic processes.

2.2.1 SDEs driven by Brownian motions

Consider a (real) differential equation of the form:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = a(t)x(t), \qquad x(0) = x_0. \tag{2.1}$$

Suppose that $a(t)$ is a stochastic function and is given as:

$$a(t) = f(t) + h(t)Z(t),$$

where $f(t)$ and $h(t)$ are some deterministic functions and $Z(t)$ denotes some stochastic process. Then, equation (2.1) can be written as follows:

$$\mathrm{d}x(t) = f(t)x(t)\mathrm{d}t + h(t)x(t)Z(t)\mathrm{d}t.$$

If we further assume that $Z(t)\mathrm{d}t$ is in fact the differential form of a Brownian motion, then we obtain the following stochastic differential equation:

$$\mathrm{d}x(t) = f(t)x(t)\mathrm{d}t + h(t)x(t)\mathrm{dB}(t),$$

where $\mathrm{B}(t)$ denotes the Brownian motion (or Wiener process), which is defined as:

Definition 2 (Brownian motion)**.** *A Brownian motion $B(t)$ is a stochastic process satisfying the following properties:*

(i) $B(0) = 0$.

(ii) *Independent increments: for every $t > 0, s \geq 0, u \geq 0$ with $s \leq t$, the random variables $B(t+u) - B(t)$, $B(s)$ are independent.*

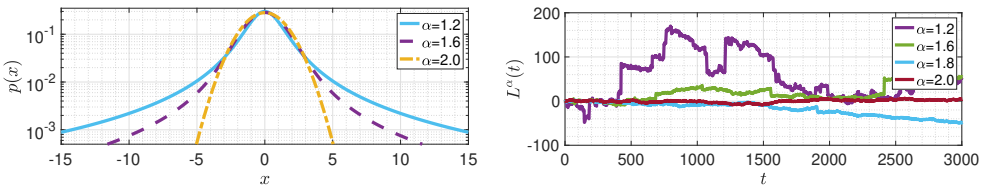
(iii) *Gaussian increments: for all $0 \leq u \leq t$, the random variable $B(t+u) - B(t)$ has the same distribution as Gaussian random variable $\mathcal{N}(0, u)$.*

(iv) *Continuous paths: $B(t)$ is continuous in t.*

Notational convenience: In this thesis, we also use B_t to denote the Brownian motion.

More generally, an SDE is given as:

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + h(t, X(t))\mathrm{dB}(t). \tag{2.2}$$



(a) (b)

Figure 2.1: Illustration of $\mathcal{S}\alpha\mathcal{S}$ (a), L_t^α (b). As α gets smaller, $\mathcal{S}\alpha\mathcal{S}$ becomes heavier-tailed and consequently, $L^\alpha(t)$ incurs larger jumps.

In the following, we consider examples of SDEs driven by Brownian motions.

Examples:

1. Ornstein–Uhlenbeck process. It is defined by the following SDE:

$$\mathrm{d}X(t) = -\theta X(t)\mathrm{d}t + \sigma \mathrm{dB}(t),$$

where $\theta, \sigma > 0$. The solution of this SDE can be written as follows.

$$X(t) = X(0)\exp(-\theta t) + \int_0^t \sigma \exp(-\theta(t-s))\mathrm{dB}(s).$$

2. Langevin equation. The Langevin diffusion is described by the following SDE:

$$\mathrm{d}X(t) = -\nabla f(X(t))\mathrm{d}t + \sqrt{2}\mathrm{dB}(t),$$

where f is a smooth function. The above Langevin diffusion possesses an invariant measure π, whose density is proportional to $\exp(-f(x))$.

In order to illustrate the random behaviors of different phenomena, it is possible to use other types of stochastic processes, such as jump processes like Lévy process, instead of Brownian motion. In the next section, we introduce the notion of stable distributions and the SDEs driven by Lévy motions.

2.2.2 Stable distributions and Lévy motions

The central limit theorem (CLT) states that the sum of i.i.d. random variables with a finite second moment converges to a normal distribution if the number of summands grows. However, if the variables have heavy-tails, the second moment may not exist. For instance, if their density $p(x)$ has a power-law tail decreasing as $1/|x|^{\alpha+1}$ (Figure 2.1) where $0 < \alpha < 2$; only r-th moment exists with $r < \alpha$. In this case, the generalized central limit theorem (GCLT) says that the sum of such variables will converge to a distribution called the *α-stable* distribution instead as the number of summands grows (see e.g. [Fischer, 2010]). In this work, we focus on the centered *symmetric α-stable* ($\mathcal{S}\alpha\mathcal{S}$) distribution, which is a special case of α-stable distributions that are symmetric around the origin.

We can view the $\mathcal{S}\alpha\mathcal{S}$ distribution as a heavy-tailed generalization of a centered Gaussian distribution. The $\mathcal{S}\alpha\mathcal{S}$ distributions are defined through their characteristic

function:

Definition 3 (Symmetric α-stable random variables)**.** *The α-stable distribution appears as the limiting distribution in the generalized CLT Samorodnitsky and Taqqu [1994]. A scalar random variable $X \in \mathbb{R}$ is called symmetric α-stable if its characteristic function has the following form:*

$$\mathbb{E}[e^{i\omega X}] = \exp(-\sigma|\omega|^{\alpha})$$

where $\alpha \in (0, 2]$ and $\sigma > 0$. We denote $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$.

Even though their probability density function does not admit a closed-form formula in general except in special cases, their density decays with a power law tail like $1/|x|^{\alpha+1}$ where $\alpha \in (0, 2]$ is called the *tail-index* which determines the behavior of the distribution: as α gets smaller, the distribution has a heavier tail. In fact, the parameter α also determines the moments: when $\alpha < 2$, $\mathbb{E}[|X|^r] < \infty$ if and only if $r < \alpha$; implying X has infinite variance when $\alpha \neq 2$. The parameter $\sigma \in \mathbb{R}_+$ is the *scale* parameter and controls the spread of X around 0. We recover the Gaussian distribution $\mathcal{N}(0, 2\sigma^2)$ as a special case when $\alpha = 2$ and the Cauchy distribution when $\alpha = 1$.

For the scalar case, α-stable Lévy motion is defined as follows for $\alpha \in (0, 2]$ (Duan [2015]):

Definition 4 (Symmetric α-stable Lévy motion)**.** *A scalar symmetric α-stable Lévy motion $L^{\alpha}(t)$, with $0 < \alpha \leq 2$, is a stochastic process satisfying the following properties:*

(i) $L^{\alpha}(0) = 0$*, almost surely.*

(ii) Independent increments: for $0 \leq t_1 < \ldots < t_n$, the random variables $L^{\alpha}(t_2) - L^{\alpha}(t_1)$,..., $L^{\alpha}(t_n) - L^{\alpha}(t_{n-1})$ are independent.

(iii) Stationary increments: for all $0 \leq s < t$, the random variables $L^{\alpha}(t) - L^{\alpha}(s)$ and $L^{\alpha}(t-s)$ have the same distribution as $\mathcal{S}\alpha\mathcal{S}((t-s)^{1/\alpha})$.

(iv) Continuity in probability: for any $\delta > 0$ and $s \geq 0$, $\mathbb{P}(|L^{\alpha}(s) - L^{\alpha}(t)| > \delta) \to 0$, as $t \to s$.

Notational convenience: In this thesis, we also use L_t to denote the Lévy motion.

By using this definition, one can define the Lévy-driven SDEs by replacing the Brownian motion in SDE (2.2) by a Lévy motion:

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + h(t, X(t))\mathrm{dL}(t). \tag{2.3}$$

To conclude this chapter, let us define the notion of invariant measure for equation (2.3).

Definition 5. *Let $q(X, t)$ be the probability density function of $X(t)$. Then $q(X, t)$ is called invariant measure for* (2.3) *if $\partial_t q(X, t) = 0$.*

In the next chapter, we will give an example of Lévy-driven SDE, which is called the Fractional Langevin Monte Carlo, along with its invariant measure.

Chapter 3

Related work

In this chapter, we present some lines of works and contributions that have connections with our work: Important aspects of deep learning theory and diffusion-based Markov Chain Monte Carlo algorithms.

3.1 Deep learning theory

Despite the success of deep neural networks in many important machine learning applications Carleo et al. [2019], Barbier et al. [2019], Decelle et al. [2011], a satisfactory theory of deep learning is still lacking. The theoretical understanding of deep learning includes the following areas Poggio et al. [2019]: 1) representation power of deep networks, 2) optimization of the empirical risk, 3) generalization properties of gradient descent techniques.

3.1.1 Representation power of deep networks

We start with the **representation power of deep networks**. Even though both deep and shallow networks have a universal property Poggio et al. [2019], that is, they can approximate any continuous function of finite variables on a compact domain, using deep networks for approximation can achieve much better performance than using shallow networks. In contrast to shallow networks, deep networks can avoid the curse of dimensionality on certain classes of problems. For instance, with the same degree of approximation, one can represent compositional functions with much smaller number of parameters for the deep networks than for the shallow networks Poggio et al. [2019]. That is to say, the expressive power of deep neural networks can be considered from the view of the depth of a network. Recent works Cohen et al. [2016], Telgarsky [2016] show the depth-efficiency of deep neural networks by proving that there exist classes of deep networks that cannot be approximated by any shallow network whose width is smaller than or equal to an exponential bound. In Lu et al. [2017], the authors deal with the

dual problem on the width-efficiency of neural networks. By using width-bounded ReLU (rectified linear unit) networks, they show that there exist classes of wide networks that cannot be represented by any narrow network whose depth is smaller than or equal to an polynomial bound. These results suggest that increasing the depth may be more effective than the width for increasing expressive power of neural networks.

Recently, several new approaches are proposed to understand the representation power of deep neural networks. In Raghu et al. [2017], the authors introduce a set of measures of expressivity, which determines how the output of a network changes as the input moves along an one-dimensional trajectory. By using these measures, they prove that the complexity of the computed function grows exponentially with depth. Furthermore, they find that the weights of trained networks are not equal in the sense that the networks tend to be more sensitive to the weights in initial layers. This result suggests that optimizing these weights is particularly important. Another approach to investigate the representation problem is based on the mathematical theory of quiver representation Armenta and Jodoin [2020], which is used for exploring the combinatorial and algebraic nature of neural networks. In Armenta and Jodoin [2020], the authors establish an explicit connection between neural networks and quiver representations and show that quiver representations are able to adapt common concepts of neural networks such as fully-connected layers, convolution operations, residual connections, batch normalization and pooling operations. The quiver representations also help understand how neural networks create representations from the data. The representation power alters when dealing with different types of neural networks. In Dehmamy et al. [2019], the authors study the expressiveness of graph convolutional networks (GCNs), which are capable of distinguishing graphs from different graph generation models. They conclude that GCNs with different propagation rules could improve the representation power significantly.

3.1.2 Optimization techniques

Deep neural networks have proven successful in a wide variety of applications. However, as neural networks grow deeper and datasets become bigger, training these networks becomes more difficult. Therefore, a large number of **optimization techniques** have been developed over the past years to improve the training performance of these networks.

Although traditional gradient-based methods using a full-gradient approach may be effective for small-scale learning problems, stochastic-gradient-based methods (SG), first proposed by Robbins and Monro [1951], are the dominant techniques for large-scale learning problems. Stochastic gradient methods gain a great success in the study of perceptual tasks such as speech or image recognition Bottou et al. [2018], in which, due to the use of deep neural networks, highly nonlinear and non-convex problems are involved. Over the years, a large variety of stochastic-gradient-based algorithms have

been proposed. These algorithms can be classified into two categories: noise reduction methods and second-order methods.

In optimization problems, methods with noise reduction capabilities have been developed to improve the accuracy of the outcomes as well as the convergence performance. The first type of noise reduction is dynamic sampling Bottou et al. [2018], in which one attempts to reduce the stochastic gradient noise by gradually increasing the mini-batch size, thus improving the accuracy of the gradient estimates as the algorithm proceeds. On the other hand, gradient aggregation methods, such as stochastic variance reduced gradient (SVRG), stochastic average gradient (SAG) and SAGA (Defazio et al. [2014]), achieve the noise reduction by improving the quality of the search directions using the information from computed gradient estimates in previous iterations Bottou et al. [2018]. These noise reduction techniques allow the optimization process to attain a linear convergence rate to the optimal value using constant step-size Bottou et al. [2018].

Another important family of stochastic-gradient-based methods, known as second-order methods, consists of algorithms that aim at addressing the negative effects of high non-linearity and ill-conditioning of the objective function with the help of second-order information Bottou et al. [2018] such as second derivative, Hessian matrix, second-order Taylor series, Fisher information matrix, etc. Popular representatives of these methods are inexact Newton and quasi-Newton methods, (generalized) Gauss-Newton methods Bertsekas [1996], Schraudolph [2002], the natural gradient method Amari [1998], and scaled gradient iterations Duchi et al. [2011], Tieleman and Hinton [2017].

Some other well-known adaptive SG methods are not well classified within the two categories: noise reduction and second-order methods, yet show great potential in theoretical and/or practical problems. We start with SG methods with momentum, which aim at improving learning performance by solving the problem of poor conditioning on the Hessian matrix and the variance in stochastic gradient Soydaner [2020]. The idea is to take a moving average between the stochastic gradient of the current and previous iterations according to a momentum parameter Alpaydin [2020]. A variant of the standard momentum is Nesterov momentum Sutskever et al. [2013], in which the gradient of the objective function is measured slightly ahead of the current update, in the direction of the momentum. Next, we move to AdaGrad Duchi et al. [2011], which is an optimization algorithm that adapts the learning rates of the model parameters, namely, at each iteration the parameters with larger (smaller) partial derivative are assigned with larger (smaller) learning rates. The learning rates are computed using all the squared values of the gradient from the previous iteration. One drawback of AdaGrad is that the continual decay of learning rates may cause the algorithm to stop too early when training neural networks, as the learning rates may become infinitesimally small. Therefore, AdaDelta Zeiler [2012] is proposed to address this issue. By using the window of fixed size w of past gradients instead of the full-gradients, AdaDelta is able to continue running even after many iterations have been progressed. Another modified

version of AdaGrad is root mean square propagation (RMSProp) Hinton et al. [2012b], which is designed to perform better in non-convex setting. In RMSProp, the gradient accumulation is replaced by an exponentially weighted moving average, which removes the excessive decay of learning rates of AdaGrad and makes the algorithm converge quickly after finding a convex bowl Goodfellow et al. [2016]. One of the most popular optimization algorithm in deep learning is Adam, which incorporate the advantages of AdaGrad and RMSProp Kingma and Ba [2014]. The algorithm adapts learning rates for model parameters from the information of the first and second moments of the gradients Soydaner [2020]. Besides computational efficiency and small-memory requirement, Adam is known to be suitable for non-stationary objectives and systems with highly noisy and sparse gradients Kingma and Ba [2014]. Some well-known modified/improved algorithm based on Adam are AdaMax, Nadam Dozat [2016] and AMSGrad Reddi et al. [2019], which are proposed to improve the speed of convergence while maintaining the benefits of Adam. While having great advantages in terms of computation efficiency and convergence speed, Adam and its modified algorithms have been found to generalize poorly compared to SGD Keskar and Socher [2017], that suggests using a hybrid strategy that combines an adaptive method and SGD Keskar and Socher [2017].

3.1.3 Generalization and SGD

Besides representation and optimization problems of neural networks, understanding **generalization** in deep learning becomes more and more important. Despite training a complex, non-convex objective function, simple methods such as stochastic gradient descent (SGD) are capable of finding solutions that have good generalization property, even when the model is over-parameterized Neyshabur et al. [2014], Zhang et al. [2017a]. In such non-convex setting, the objective function may have multiple local minima, however not all of them are able to generalize well: bad local minima can lead to poor generalization performance. Different algorithms such as SGD, Adam, and different parameter settings such as initialization, learning rate, batch-size, may lead to local minima with different generalization performance Chaudhari et al. [2016], Keskar et al. [2016], Neyshabur et al. [2015].

In recent years, a considerable number of approaches have been suggested to explain why deep learning can generalize well and how to improve generalization behavior. Even though over-parameterized neural networks can perfectly fit data labels without generalizing Zhang et al. [2017a], the improvement in generalization error as the number of hidden units increases cannot be explained in terms of number of parameters Neyshabur et al. [2014]. On the other hand, generalization behavior can be controlled by different norms of network parameters Bartlett et al. [2017], which are independent of number of parameters. In terms of various norms such as square norm and spectral norm, Neyshabur et al. [2017] provide the bounds for the number of data samples required to ensure generalization. In another line of work, Keskar et al. [2016] introduce

the notion of 'sharpness' of local minima to investigate the generalization behavior, based on a well-known hypothesis Keskar et al. [2016], Chaudhari et al. [2016] that flat minima tend to generalize better than sharp minima. Later, Neyshabur et al. [2017] improve this approach by viewing the notion of sharpness in the context of PAC-Bayesian framework. In different aspects, Xiao et al. provide necessary conditions for generalization of neural networks by analyzing the spectrum of neural tangent kernel and Li et al. [2020] advance the understanding of the relations between the compressibility and generalization of neural networks by using tensor analysis. Remarkably, Şimşekli et al. [2020] study generalization properties of SGD via stochastic differential equations (SDEs). The authors show that the generalization error can be estimated by a peculiar characteristic of SDEs under heavy-tailed gradient noise - the Hausdorff dimension.

3.2 Diffusion-based Markov Chain Monte Carlo

Diffusion-based Markov Chain Monte Carlo (MCMC) algorithms aim at generating samples from a distribution that is only accessible by its unnormalized density function. Recently, they have become increasingly popular due to their nice scalability properties and theoretical guarantees Ma et al. [2015], Chen et al. [2015], Şimşekli et al. [2016], Durmus et al. [2016]. In addition to their success in Bayesian machine learning, they have also been used for analyzing large-scale non-convex optimization algorithms Raginsky et al. [2017], Xu et al. [2018], Şimşekli et al. [2018], Birdal et al. [2018], Birdal and Şimşekli [2019] and understanding the behavior of stochastic gradient descent in deep learning settings Jastrzebski et al. [2017], Şimşekli et al. [2019].

One of the most popular approaches in this field is based on the so-called Langevin diffusion, which is described by the following stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \sqrt{2/\beta}\,\mathrm{dB}_t, \;\; t \geq 0, \tag{3.1}$$

where $\mathbf{w}_t \in \mathbb{R}^d$, f is a smooth function which is often non-convex, $\beta \in \mathbb{R}_+$ is called the 'inverse temperature' parameter, and B_t is the standard Brownian motion in $\mathbb{R}^d$.

Under some regularity conditions on f, one can show that the Markov process $(\mathbf{w}_t)_{t\geq 0}$, i.e. the solution of the SDE (3.1), is ergodic with its unique invariant measure π, whose density is proportional to $\exp(-\beta f(\mathbf{w}))$ Roberts and Stramer [2002]. An important feature of this measure is that, when β goes to infinity, its density concentrates around the global minimum $\mathbf{w}^\star \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$ Hwang [1980], Gelfand and Mitter [1991]. This property implies that, if we could simulate (3.1) for large enough β and t, the simulated state $\mathbf{x}_t$ would be close to $\mathbf{w}^\star$.

This connection between diffusions and optimization, motivates simulating (3.1) in discrete-time in order to obtain 'almost global optimizers'. If we use a first-order Euler-Maruyama discretization, we obtain a 'tempered' version of the well-known Unadjusted

Langevin Algorithm (ULA) Roberts and Stramer [2002]:

$$\mathbf{w}_{\mathrm{ULA}}^{k+1} = \mathbf{w}_{\mathrm{ULA}}^{k} - \eta \nabla f(\mathbf{w}_{\mathrm{ULA}}^{k}) + \sqrt{\frac{2\eta}{\beta}} Z_{k+1}, \tag{3.2}$$

where $k \in \mathbb{N}_+$ denotes the iterations, η denotes the step-size, and $(Z_n)_n$ is a sequence of independent and identically-distributed (i.i.d.) standard Gaussian random variables. When $\beta = 1$, we obtain the classical ULA, which is mainly used for Bayesian posterior sampling. Theoretical properties of the classical ULA have been extensively studied Roberts and Stramer [2002], Lamberton and Pages [2003], Durmus and Moulines [2015, 2016], Dalalyan [2017b].

When $\beta \gg 1$, the algorithm is called tempered and becomes more suitable for optimization. Indeed, one can observe that the noise term Z_k in (3.2) becomes less dominant, and the overall algorithm can be seen as a 'perturbed' version of the gradient descent (GD) algorithm. The connection between ULA and GD has been recently established in Dalalyan [2017a] for strongly convex f. Moreover, Raginsky et al. [2017] and Xu et al. [2018] proved non-asymptotic guarantees for this perturbed scheme[1]. Their results showed that, even in non-convex settings, the algorithm is guaranteed to escape from local minima and converge near the global minimizer. These results were extended in Zhang et al. [2017b] and Tzen et al. [2018], which showed that the iterates converge near a local minimum in polynomial time and stay there for an exponential time. Recently, the guarantees for ULA were further extended to second-order Langevin dynamics Gao et al. [2018b,a].

Fractional Langevin Monte Carlo

Another line of research has extended Langevin Monte Carlo by replacing the Brownian motion with a motion which can incur 'jumps' (i.e. discontinuities), such as the α-*stable Lévy Motion* (see Figure 2.1) Şimşekli [2017], Ye and Zhu [2018]. Coined under the name of Fractional Langevin Monte Carlo (FLMC) methods, these approaches are motivated by the statistical physics origins of the Langevin equation (3.1). In such a context, the Langevin equation aims to model the position of a small particle that is under the influence of a force, which has a deterministic and a stochastic part. If we assume that the stochastic part of this force is a sum of many i.i.d. random variables with finite variance, then by the central limit theorem (CLT), we can assume that their sum follows a Gaussian distribution, which justifies the Brownian motion in (3.1).

The main idea in FLMC is to relax the finite variance assumption and allow the random pulses to have infinite variance. In such a case, the classical CLT will not hold; however, the *extended* CLT Lévy [1937] will still be valid: the law of the sum of the pulses converges to an α-*stable* distribution, a family of 'heavy-tailed' distributions that

[1]The results given in Raginsky et al. [2017] are more general in the sense that they are proved for the Stochastic Gradient Langevin Dynamics (SGLD) algorithm Welling and Teh [2011], which is obtained by replacing the gradients in (3.2) with stochastic gradients.

contains the Gaussian distribution as a special case. Then, by using a similar argument to the previous case, we can replace the Brownian motion with the α-stable Lévy Motion Yanovsky et al. [2000], whose increments are α-stable distributed.

Based on an SDE driven by an α-stable Lévy Motion, Şimşekli [2017] proposed the following iterative scheme that is referred to as Fractional Langevin Algorithm (FLA):

$$\mathbf{w}_{\mathrm{FLA}}^{k+1} = \mathbf{w}_{\mathrm{FLA}}^{k} - \eta c_\alpha \nabla f(\mathbf{w}_{\mathrm{FLA}}^{k}) + \left(\frac{\eta}{\beta}\right)^{\frac{1}{\alpha}} S_{k+1}, \tag{3.3}$$

where $\alpha \in (1, 2]$ is called the characteristic index, c_α is a known constant, and $\{S_k\}_{k\in\mathbb{N}_+}$ is a sequence of α-stable distributed random variables. Recently, Ye and Zhu [2018] extended FLA to Hamiltonian dynamics. The experimental results in Şimşekli [2017] and Ye and Zhu [2018] showed that the use of the heavy-tailed increments can provide advantages in multi-modal settings, robustness to algorithm parameters. Ye and Zhu [2018] further illustrated that in an optimization context their algorithm achieves better generalization in deep neural networks.

The FLA algorithm is based on a Lévy-driven SDE, that is defined as follows:

$$\mathrm{d}\mathbf{w}_t = \Psi(\mathbf{w}_{t-}, \alpha)\mathrm{d}t + (1/\beta)^{1/\alpha}\mathrm{dL}_t^\alpha \tag{3.4}$$

where $\mathbf{w}_{t-}$ denotes the *left limit* of the process at time t and L_t^α denotes the d-dimensional Lévy motion whose components are independent α-stable Lévy motions in $\mathbb{R}$.

We have the following result for FLMC:

Theorem 1 (Şimşekli [2017])**.** *Consider the SDE* (3.4) *in the case* $d = 1$, $\beta = 1$, *and* $\alpha \in (1, 2]$, *where the drift* Ψ *is defined as follows:*

$$\Psi(x, \alpha) \triangleq -\frac{\mathcal{D}^{\alpha-2}\left(\phi(x)\frac{\partial f(x)}{\partial x}\right)}{\phi(x)}. \tag{3.5}$$

where $\phi(x) \triangleq \exp(-\beta f(x))$ *and* $\mathcal{D}$ *denotes the fractional Riesz derivative and is defined as follows for a function* u*:*

$$\mathcal{D}^\gamma u(x) \triangleq \mathcal{F}^{-1}\{|\omega|^\gamma \hat{u}(\omega)\}.$$

Here, $\mathcal{F}$ *denotes the Fourier transform and* $\hat{u} \triangleq \mathcal{F}(u)$.

Let π *be a random measure whose density is* $\phi(x)$ *(up to a multiplicative factor). Then,* π *is an invariant measure of the Markov process* $(\mathbf{w}_t)_{t\geq 0}$ *that is a solution of the SDE given by* (3.4).

This theorem states that if the drift (3.5) can be computed, then the sample paths of (3.4) can be considered as samples drawn from π. However, computing (3.5) is in general not tractable, therefore one needs to approximate it for computational purposes.

If we use the alternative definition of the Riesz derivative given by Ortigueira [2006], we can approximate the drift as follows Şimşekli [2017], Ye and Zhu [2018]:

$$-\frac{\mathcal{D}^{\alpha-2}\left(\phi(x)\frac{\partial f(x)}{\partial x}\right)}{\phi(x)} \approx -c_\alpha \frac{\partial f(x)}{\partial x},$$

where $c_\alpha \triangleq \Gamma(\alpha-1)/\Gamma(\alpha/2)^2$ and Γ denotes the Gamma function. With this choice of approximation, in the d-dimensional case we obtain FLA, as given in (3.3). We can observe that, when $\alpha = 2$, (3.4) becomes the Langevin equation (3.1) and FLA becomes ULA.

Chapter 4

Heavy-tailed behavior in stochastic gradient descent

The gradient noise in SGD is often considered to be Gaussian for mathematical convenience. This assumption enables SGD to be studied as a stochastic differential equation (SDE) driven by a Brownian motion. However, the Gaussianity assumption might fail to hold in deep learning settings and hence render the Brownian motion-based analyses inappropriate. Inspired by non-Gaussian natural phenomena, we consider a better-suited hypothesis for the stochastic gradient noise that has more pertinent theoretical implications for the phenomena associated with SGD. Accordingly, we propose to analyze SGD as a discretization of an SDE driven by a Lévy motion.

This chapter is based on the article [Şimşekli et al., 2019].

4.1 Overview

A popular approach for investigating the behavior of SGD is based on considering SGD as a discretization of a continuous-time process [Mandt et al., 2016, Jastrzebski et al., 2017, Li et al., 2017, Hu et al., 2017, Zhu et al., 2018, Chaudhari and Soatto, 2018]. This approach models the stochastic gradient noise as a Gaussian distribution, i.e. $U_k(\mathbf{w}) \triangleq \nabla \tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})$ satisfies

$$U_k(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4.1}$$

where $\mathcal{N}$ denotes the multivariate (Gaussian) normal distribution and $\mathbf{I}$ denotes the identity matrix of appropriate size.[1] The rationale behind this assumption is that, if the size of the minibatch b is large enough, then we can invoke the Central Limit Theorem

[1] We note that more sophisticated assumptions than (4.1) have been made in terms of the covariance matrix of the Gaussian distribution (e.g. state dependent, anisotropic). However, in all these cases, the resulting distribution is still a Gaussian, therefore the same criticism holds.

(CLT) and assume that the distribution of U_k is approximately Gaussian. Then, under this assumption, SGD (1.2) can be written as follows:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta \nabla f(\mathbf{w}^k) + \sqrt{\eta}\sqrt{\eta\sigma^2} Z_k, \tag{4.2}$$

where Z_k denotes a standard normal random vector in $\mathbb{R}^d$. If we further assume that η is small enough, then the continuous-time analogue of the discrete-time process (4.2) is the following stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \sqrt{\eta\sigma^2}\mathrm{dB}_t, \tag{4.3}$$

where B_t denotes the standard Brownian motion. This SDE is a variant of the well-known *Langevin diffusion* and under mild regularity assumptions on f, one can show that the Markov process $(\mathbf{w}_t)_{t\geq 0}$ is ergodic with its unique invariant measure, whose density is proportional to $\exp(-f(\mathbf{w})/(\eta\sigma^2))$ for any $\eta > 0$ [Roberts and Stramer, 2002]. From this perspective, the SGD recursion in (4.2) can be seen as a first-order Euler-Maruyama discretization of the Langevin dynamics (see also [Li et al., 2017, Jastrzebski et al., 2017, Hu et al., 2017]), which is often referred to as the Unadjusted Langevin Algorithm (ULA) [Roberts and Stramer, 2002, Lamberton and Pages, 2003, Durmus and Moulines, 2015, Durmus et al., 2016].

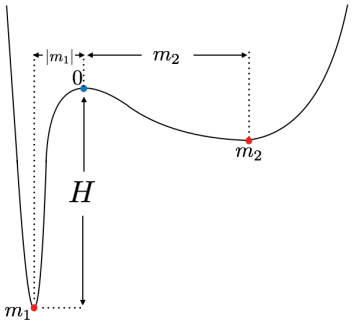
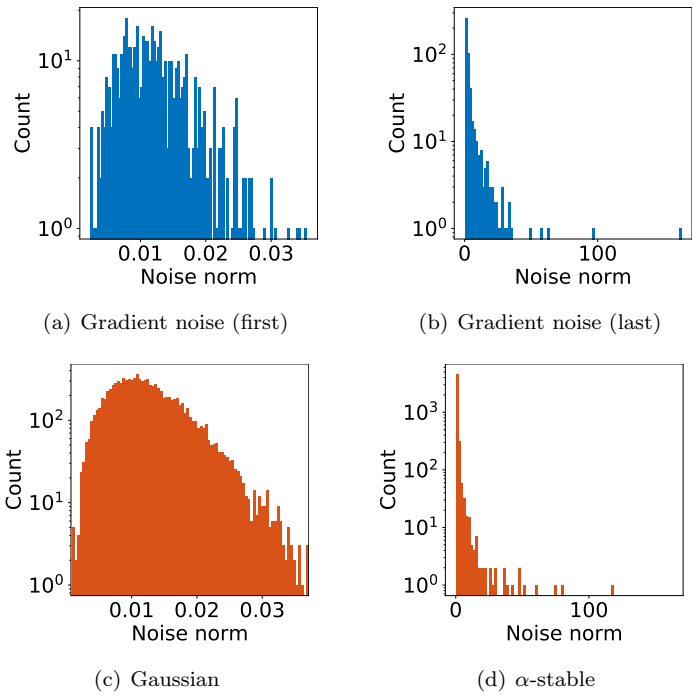


Figure 4.1: A function with a narrow minimum of width m_1 and a wide minimum of width m_2.

Based on this observation, Jastrzebski et al. [2017] focused on the relation between this invariant measure and the algorithm parameters, namely the step-size η and mini-batch size, as a function of σ^2. They concluded that the ratio of step-size divided by the batch size is the control parameter that determines the width of the minima found by SGD. Furthermore, they revisit the famous wide minima folklore [Hochreiter and Schmidhuber, 1997]: Among the minima found by SGD, the wider it is, the better it performs on the test set. This is visualized in Figure 4.1, where the local minimum on the right lies on a wider valley with width m_2 compared to the local minimum on the left with width m_1 lying in a sharp valley of depth H. However, there are several fundamental issues with this approach, which we will explain below.



(a) Gradient noise (first) (b) Gradient noise (last)

(c) Gaussian (d) α-stable

Figure 4.2: A mismatch between the Gaussianity assumption and the empirical behavior of the stochastic gradient noise. (a) and (b) The histogram of the norms of the gradient noises after the first and the last iterations, respectively (computed with AlexNet on CIFAR10). (c) and (d) the histograms of the norms of (scaled) Gaussian and α-stable random vectors, respectively.

We first illustrate a typical mismatch between the Gaussianity assumption and the empirical behavior of the stochastic gradient noise in terms of the long term behavior. In Figure 4.2, we plot the histogram of the norms of the stochastic gradient noise at the first and the last iterations that are computed using a convolutional neural network (AlexNet) in an image classification problem on the CIFAR10 dataset and compare it to the histogram of the norms of Gaussian random vectors. It can be clearly observed that, even though the shape of the histogram corresponding to gradients resembles the one of the Gaussian vectors at the first iteration, throughout training, it drifts apart from the Gaussian and exhibits a *heavy-tailed* behavior.

In addition to the empirical observations, the Gaussianity assumption also yields some theoretical issues. The first issue with this assumption is that the current SDE analyses of SGD are based on the *invariant measure* of the SDE, which implicitly assumes that sufficiently many iterations have been taken to converge to that measure. Recent results on ULA [Raginsky et al., 2017, Xu et al., 2018] have shown that, the required number of iterations to achieve the invariant measure often grows exponentially with the dimension d. This result contradicts with the current practice: considering the large size of the neural networks and limited computational budget, only a limited

number of iterations – which is much smaller than $\exp(\mathcal{O}(d))$ – can be taken. This conflict becomes clearer in the light of the recent works that studied the *local* behavior of ULA [Tzen et al., 2018, Zhang et al., 2017b]. These studies showed that ULA will get close to the nearest local optimum in polynomial time; however, the required amount of time for escaping from that local optimum increases exponentially with the dimension. Therefore, the phenomenon that SGD prefers wide minima within a considerably small number of iterations cannot be explained using the asymptotic distribution of the SDE given in (4.3).

The second issue is related to the local behavior of the process and becomes clear when we consider the *metastability* analysis of Brownian motion-driven SDEs. These studies [Freidlin and Wentzell, 1998, Bovier et al., 2004, Imkeller et al., 2010b] consider the case where $\mathbf{w}_0$ is initialized in a quadratic basin and then analyze the minimum time t such that $\mathbf{w}_t$ is outside that basin. They show that this so-called *first exit time* depends *exponentially* on the height of the basin; however, this dependency is only *polynomial* with the width of the basin. These theoretical results directly contradict with the wide minima phenomenon: even if the height of a basin is slightly larger, the exit-time from this basin will be dominated by its height, which implies that the process would stay longer in (or in other words, 'prefer') deeper minima as opposed to wider minima. The reason why the exit-time is dominated by the height is due to the *continuity* of the Brownian motion, which is in fact a direct consequence of the Gaussian noise assumption.

A final remark on the issues of this approach is the observation that landscape is flat at the bottom regardless of the batch size used in SGD [Sagun et al., 2017]. In particular, the spectrum of the Hessian at a near critical point with close to zero loss value has many near zero eigenvalues. Therefore, local curvature measures that are used as a proxy for measuring the width of a basin can be misleading. Such measures usually correlate with the magnitudes of large eigenvalues of the Hessian which are few [Keskar et al., 2016, Jastrzebski et al., 2017]. Besides, during the dynamics of SGD it has been observed that the algorithm does not cross barriers except perhaps at the very initial phase [Xing et al., 2018, Baity-Jesi et al., 2018]. Such dependence of width on an essentially-flat landscape combined with the lack of explicit barrier crossing during the SGD descent forces us to rethink the analysis of basin hopping under a noisy dynamics.

In this study, we aim at addressing these contradictions and come up with an arguably better-suited hypothesis for the stochastic gradient noise that has more pertinent theoretical implications for the phenomena associated with SGD.

4.2 Proposed framework

We go back to (1.3) and (4.1) and reconsider the application of CLT. This *classical* CLT assumes that U_k is a sum of many independent and identically distributed (i.i.d.)

random vectors, whose covariance matrix exists and is invertible, and then it states that the law of U_k converges to a Gaussian distribution, which then paves the way for (4.2). Even though the finite-variance assumption seems natural and intuitive at the first sight, it turns out that in many domains, such as turbulent motions (Weeks et al. [1995]), oceanic fluid flows (Woyczyński [2001]), finance (Mandelbrot [2013]), biological evolution (Jourdain et al. [2012]), audio signals (Liutkus and Badeau [2015], Şimşekli et al. [2015], Leglaive et al. [2017], Şimşekli et al. [2018]), brain signals (Jas et al. [2017]), the assumption might fail to hold (see [Duan, 2015] for more examples). In such cases, the classical CLT along with the Gaussian approximation will no longer hold. While this might seem daunting, fortunately, one can prove a *generalized* CLT and show that the law of the sum of these i.i.d. variables with infinite variance still converges to a family of *heavy-tailed* distributions that is called the α-stable distribution [Lévy, 1937]. As we detailed in Section 2.2.2, these distributions are parametrized by their *tail-index* $\alpha \in (0, 2]$ and they coincide with the Gaussian distribution when $\alpha = 2$.

In this study, we relax the finite-variance assumption on the stochastic gradient noise and by invoking the generalized CLT: we assume that U_k follows an α-stable distribution, as hinted in Figure 4.2(d). By following a similar rationale to (4.2) and (4.3), we reformulate SGD with this new assumption and consider its continuous-time limit for small step-sizes. Since the noise might not be Gaussian anymore (i.e. when $\alpha \neq 2$), the use of the Brownian motion would not be appropriate in this case and we need to replace it with the α-stable Lévy motion, whose increments have an α-stable distribution (Yanovsky et al. [2000]). Due to the heavy-tailed nature of α-stable distribution, the Lévy motion might incur large discontinuous jumps and therefore exhibits a fundamentally different behavior than the Brownian motion, whose paths are on the contrary almost surely continuous. The discontinuities also reflect in the metastability properties of Lévy-driven SDEs, which indicate that, as soon as $\alpha < 2$, the first exit time from a basin does *not* depend on its height; on the contrary, it directly depends on its width and the tail-index α Imkeller and Pavlyukevich [2006], Imkeller et al. [2010b,a]. Informally, this implies that the process will *escape* from narrow minima – no matter how deep they are – and stay longer in wide minima. Besides, as α gets smaller, the probability for the dynamic to jump into a wide basin will increase. Therefore, if the α-stable assumption on the stochastic gradient noise holds, then the existing metastability results automatically provide strong theoretical insights for illuminating the behavior of SGD.

4.3 SGD as a Lévy-driven SDEs

Following the above argument, a more general assumption on the stochastic gradient noise (cf. (4.1)) can be given by:

$$[U_k(\mathbf{w})]_i \sim \mathcal{S}\alpha\mathcal{S}_i(\sigma_i(\mathbf{w})), \quad \forall i = 1, \dots, n \tag{4.4}$$

where $[v]_i$ denotes the i'th component of a vector v, and $\mathcal{S}\alpha\mathcal{S}_i$ distributed with $\alpha_i(w)$. Clearly, this assumption is way too general to offer reasonable theoretical treatment. We will resort to several simplifications: (1) We assume that each coordinate of U_k is $\mathcal{S}\alpha\mathcal{S}$ distributed with the same σ which depends on the state $\mathbf{w}$. Here, this dependency is not crucial since we are mainly interested in the tail-index α, which can be estimated *independently* from the scale parameter (Section 4.5.1). Therefore, we will simply denote $\sigma(\mathbf{w})$ as σ for clarity. (2) We further assume that each coordinate of U_k is $\mathcal{S}\alpha\mathcal{S}$ distributed with the same α independent of the state $\mathbf{w}$. We will demonstrate the state independence at later stages of SGD experimentally in Section 4.6.3, however, imposing the coordinate dependence is a much harder challenge which will be addressed in the section devoted for open problems (Chapter 7).

By using the assumption (4.4), we can rewrite the SGD recursion as follows [Şimşekli, 2017]:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta \nabla f(\mathbf{w}^k) + \eta^{1/\alpha} \Big(\eta^{\frac{\alpha-1}{\alpha}} \sigma \Big) S_k, \tag{4.5}$$

where $S_k \in \mathbb{R}^d$ is a random vector such that $[S_k]_i \sim \mathcal{S}\alpha\mathcal{S}(1)$. If the step-size η is small enough, then we can consider the continuous-time limit of this discrete-time process, which is expressed in the following SDE driven by an α-stable Lévy process:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \eta^{(\alpha-1)/\alpha} \sigma \, \mathrm{dL}_t^\alpha, \tag{4.6}$$

where L_t^α denotes the d-dimensional α-stable Lévy motion with *independent components*. In other words, each component of L_t^α is an independent α-stable Lévy motion in $\mathbb{R}$.

It is easy to check that the noise term in (4.5) is obtained by integrating dL_t^α from $k\eta$ to $(k+1)\eta$. When $\alpha = 2$, L_t^α coincides with a scaled version of Brownian motion, $\sqrt{2}\mathrm{B}_t$. $\mathcal{S}\alpha\mathcal{S}$ and L_t^α are illustrated in Figure 2.1.

Stochastic processes based on Lévy motion such as (4.6) and their mathematical properties have also been studied in the literature, we refer the reader to Tankov [2003], Øksendal and Sulem [2005] for details.

4.4 First exit time and metastability properties

Consider a basin in which a particle is initialized and undergoes fluctuations continually, the particle persists in the basin for a long time before exiting it by the influence of fluctuations. This relative instability phenomenon is described by the term 'metastability'. More formally, the metastability studies consider the case where $\mathbf{w}_0$ is initialized in a basin and analyze the minimum time t such that $\mathbf{w}_t$ *exits* that basin.

We start by reviewing known metastability properties of the α-stable Lévy process (4.6) from the literature. We will also focus on the *first exit time* which is, roughly speaking, the average time it takes for the process to exit a neighborhood of a local minima (a quantity we define formally later in (4.12)).

For clarity of the presentation and notational simplicity we focus on the scalar case and consider the SDE (4.6) in $\mathbb{R}$ (i.e. $d = 1$). Multidimensional generalizations of the metastability results presented in this work can be found in [Imkeller et al., 2010a] and will be summarized at the end of this section. We rewrite (4.6) as follows:

$$\mathrm{d}w_t^\varepsilon = -f'(w_t^\varepsilon)\mathrm{d}t + \varepsilon \mathrm{d}\mathrm{L}_t^\alpha \tag{4.7}$$

for $t \geq 0$, started from the initial point $w_0 \in \mathbb{R}$, where L_t^α is the α-stable Lévy process, $\varepsilon \geq 0$ is the noise level and f is a non-convex objective with $r \geq 2$ local minima. We denote the derivative of f by f'. When $\varepsilon = 0$, we recover the gradient descent dynamics in continuous time: $\mathrm{d}w_t^0 = -f'(w_t^0)\mathrm{d}t$, where the local minima are the stable points of this differential equation. However, as soon as $\varepsilon > 0$, these states become 'metastable', meaning that there is a positive probability for w_t^ε to transition from one basin to another. However, the time required for transitioning to another basin strongly depends on the characteristics of the injected noise. The two most important cases are $\alpha = 2$ and $\alpha < 2$. When $\alpha = 2$, (i.e. the Gaussianity assumption) the process $(w_t^\varepsilon)_{t\geq 0}$ is continuous, which requires it to 'climb' the basin all the way up, in order to be able to transition to another basin. This fact makes the transition-time depend on the height of the basin. On the contrary, when $\alpha < 2$, the process can incur discontinuities and does not need to cross the boundaries of the basin in order to transition to another one, since it can directly jump. This property is called the 'transition phenomenon' [Duan, 2015] and makes the transition-time mostly depend on the *width* of the basin. In the rest of the section, we will formalize these explanations.

Gradient-like flows driven by Brownian motion and weak error for their discretization are well studied from a theoretical standpoint (see e.g. [Li et al., 2017, Mertikopoulos and Staudigl, 2018]), however their Lévy-driven analogue (4.7) and the discrete-time versions [Burghoff and Pavlyukevich, 2015] are relatively less studied. Under some assumptions on the objective f, it is known that the process (4.7) admits a stationary density [Samorodnitsky and Grigoriu, 2003]. For a general f, an explicit formula for the equilibrium distribution is not known, however when the noise level ε is small enough, finer characterizations of the structure of the equilibrium density in dimension one is known. We next summarize known results in this area, which show that Lévy-driven dynamics spend more time in 'wide valleys' in the sense of [Chaudhari et al., 2016] when ε goes to zero.

Assume that f is smooth with r local minima $\{m_i\}_{i=1}^r$ separated by $r-1$ local maxima $\{s_i\}_{i=1}^{r-1}$, i.e.

$$-\infty := s_0 < m_1 < s_1 < \cdots < s_{r-1} < m_r < s_r := \infty.$$

Furthermore, assume that the local minima and maxima are not degenerate, i.e. $f''(m_i) > 0$ and $f''(s_i) < 0$ for every i. We also assume the objective gradient has a growth condition $f'(w) > |w|^{1+c}$ for some constant $c > 0$ and when $|w|$ is large enough. Each local

minima m_i lies in the (interval) valley $S_i = (s_{i-1}, s_i)$ of (width) length $L_i = |s_i - s_{i-1}|$. Consider also a δ-neighborhood $B_i := \{|x - m_i| \leq \delta\}$ around the local minimum with $\delta > 0$ small enough so that the neighborhood is contained in the valley S_i for every i. We are interested in the first exit time from B_i starting from a point $w_0 \in B_i$ and the transition time $T^i_{w_0}(\varepsilon) := \inf\{t \geq 0 : w^\varepsilon_t \in \cup_{j \neq i} B_j\}$ to a neighborhood of another local minimum, we will remove the dependency to w_0 of the transition time in our discussions as it is clear from the context. The following result shows that the transition times are asymptotically exponentially distributed in the limit of small noise and scales like $1/\varepsilon^\alpha$ with ε.

Theorem 2 (Pavlyukevich [2007])**.** *For an initial point $w_0 \in B_i$, in the limit $\varepsilon \to 0$, the following statements hold regarding the transition time:*

$$\begin{aligned} \mathbb{P}_{w_0}(w^\varepsilon_{T^i(\varepsilon)} \in B_j) &\to q_{ij} q_i^{-1} \quad \text{if} \quad i \neq j, \\ \varepsilon^\alpha T^i(\varepsilon) &\xrightarrow{d} \mathcal{E}(q_i), \\ \mathbb{E}[\varepsilon^\alpha T^i(\varepsilon)] &\to q_i^{-1}. \end{aligned}$$

where $\mathcal{E}$ denotes the exponential distribution, $\xrightarrow{d}$ denotes convergence in distribution and

$$q_{ij} = \frac{1}{\alpha} \left| \frac{1}{|s_{j-1} - m_i|^\alpha} - \frac{1}{|s_j - m_i|^\alpha} \right|, i \neq j, \tag{4.8}$$

$$q_i = \sum_{j \neq i} q_{ij}. \tag{4.9}$$

If the SDE (4.7) would be driven by the Brownian motion instead, then an analogous theorem to Theorem 2 holds saying that the transition times are still exponentially distributed but the scaling ε^α needs to be replaced by e^{2H/ε^2} where H is the maximal depth of the basins to be traversed between the two local minima [Day, 1983, Bovier et al., 2005]. This means that in the small noise limit, Brownian-motion driven gradient descent dynamics need exponential time to transit to another minimum whereas Lévy-driven gradient descent dynamics need only polynomial time. We also note from Theorem 2 that the mean transition time between valleys for Lévy SDE does not depend on the depth H of the valleys they reside in which is an advantage over Brownian motion driven SDE in the existence of deep valleys. Informally, this difference is due to the fact that Brownian motion driven SDE has to typically climb up a valley to exit it, whereas Lévy-driven SDE could jump out.

The following theorem says that as $\varepsilon \to 0$, up to a normalization in time, the process w^ε_t behaves like a finite state-space Markov process that has support over the set of local minima $\{m_i\}_{i=1}^r$ admitting a stationary density $\pi = (\pi_i)_{i=1}^r$ with an infinitesimal generator Q. The process jumps between the valleys S_i, spending time proportional to probability π_i amount of time in each valley in the equilibrium where the probabilities $\pi = (\pi_i)_{i=1}^r$ are given by the solution to the linear system $Q\pi = 0$.

Theorem 3 (Pavlyukevich [2007]). *Let $w_0 \in S_i$, for some $1 \le i \le r$. For $t \ge 0$, $w^{\varepsilon}_{t\varepsilon^{-\alpha}} \to Y_{m_i}(t)$, as $\varepsilon \to 0$, in the sense of finite-dimensional distributions, where $Y = (Y_y(t))_{t\ge 0}$ is a continuous-time Markov chain on a state space $\{m_1, m_2, \dots, m_r\}$ with the infinitesimal generator $Q = (q_{ij})_{i,j=1}^r$ with*

$$q_{ij} = \frac{1}{\alpha}\left|\frac{1}{|s_{j-1} - m_i|^\alpha} - \frac{1}{|s_j - m_i|^\alpha}\right|, \tag{4.10}$$

$$q_{ii} = -\sum_{j\neq i} q_{ij}. \tag{4.11}$$

This process admits a density π satisfying $Q^T\pi = 0$.

A consequence of this theorem is that equilibrium probabilities π_i are typically larger for "wide valleys". To see this consider the special case illustrated in Figure 4.1 with $r = 2$ local minima $m_1 < s_1 = 0 < m_2$ separated by a local maximum at $s_1 = 0$. For this example, $m_2 > |m_1|$, and the second local minimum lies in a wider valley. A simple computation reveals

$$\pi_1 = \frac{|m_1|^\alpha}{|m_1|^\alpha + m_2^\alpha}, \quad \pi_2 = \frac{|m_2|^\alpha}{|m_1|^\alpha + |m_2|^\alpha}.$$

We see that $\pi_2 > \pi_1$, that is in the equilibrium the process spends more time on the wider valley. In particular, the ratio $\frac{\pi_2}{\pi_1} = \left(\frac{m_2}{|m_1|}\right)^\alpha$ grows with an exponent α when the ratio $\frac{m_2}{|m_1|}$ of the width of the valleys grows. Consequently, if the gradient noise is indeed α-stable distributed, these results directly provide theoretical evidence for the wide-minima behavior of SGD assuming the loss landscape is not degenerate.

In addition to the transition time between the basins of attraction of two local minima, understanding how long it takes for the continuous-time process w_t given by (4.6) to exit a neighborhood of a local minimum $\bar{w}$ (given that it is started in that neighborhood) is also relevant. We formally define the *first exit time* of the stochastic process (4.6) as follows:

$$\tau_a(\varepsilon) \triangleq \inf\{t \ge 0 : |w_t - \bar{w}| \notin [0, a]\}. \tag{4.12}$$

The following result characterizes the first exit time in dimension one.

Theorem 4 (Imkeller and Pavlyukevich [2006]). *Consider the SDE (4.6) in dimension $d = 1$ and assume that it has a unique strong solution. Assume further that the objective f has a global minimum at zero, satisfying the conditions $f'(x)x \ge 0$ for every $x \in \mathbb{R}$, $f(0) = 0$, $f'(x) = 0$ if and only if $x = 0$, and $f''(0) > 0$. Then, there exist positive constants ε_0, γ, δ, and $C > 0$ such that for $0 < \varepsilon \le \varepsilon_0$, the following holds:*

$$\mathrm{e}^{-u\varepsilon^\alpha \frac{\theta}{\alpha}(1+C\varepsilon^\delta)}(1 - C\varepsilon^\delta) \le \mathbb{P}(\tau_a(\varepsilon) > u) \le \mathrm{e}^{-u\varepsilon^\alpha \frac{\theta}{\alpha}(1-C\varepsilon^\delta)}(1 + C\varepsilon^\delta) \tag{4.13}$$

for all initialization $w_0 \in [-a + \varepsilon^\gamma, a - \varepsilon^\gamma]$ and $u \ge 0$, where $\theta = \frac{2}{a^\alpha}$. Consequently,

$$\mathbb{E}[\tau_a(\varepsilon)] = \frac{\alpha}{2}\frac{a^\alpha}{\varepsilon^\alpha}(1 + \mathcal{O}(\varepsilon^\delta)), \quad \text{for all } w_0 \in [-a + \varepsilon^\gamma, a - \varepsilon^\gamma]. \tag{4.14}$$

Extension of Theorem 4 to $\mathbb{R}^d$. The exit behavior of the SDE (4.6) from an arbitrary domain in $\mathbb{R}^d$ has also been studied in the literature. Imkeller et al. [2010a] generalizes Theorem 4 from dimension $d = 1$ to arbitrary dimensions and shows that in the small noise limit the exit time from a domain is exponentially distributed with a parameter that depends on the tail-index α. In case the components of the Lévy motion in (4.6) is replaced by a process that consists of the sum of finitely many one-dimensional Lévy processes with different tail-indices α_i, it is also shown that the first exit time from a domain is determined by the smallest α_i when the noise level ε is small enough.

4.5 Experimental methodology

In this section, we describe our experimental methodology regarding how we estimate the heavy-tailedness of the stochastic gradients. First, we discuss how we can compute the tail-index α based on a recent estimator proposed in Mohammadi et al. [2015]. Second, we describe the procedure proposed in [Brcich et al., 2005] for testing whether stochastic gradients follow a symmetric α-stable distribution. Our experimental results will be presented in Section 4.6.

4.5.1 Tail index estimation

Estimating the tail-index of an extreme-value distribution is a long-standing topic. Some of the well-known estimators for this task are [Hill, 1975, Pickands, 1975, Dekkers et al., 1989, De Haan and Peng, 1998]. Despite their popularity, these methods are not specifically developed for α-stable distributions and it has been shown that they might fail for estimating the tail-index for α-stable distributions [Mittnik and Rachev, 1996, Paulauskas and Vaičiulis, 2011].

In this section, we use a relatively recent estimator proposed in [Mohammadi et al., 2015] for α-stable distributions. It is given in the following theorem.

Theorem 5 (Mohammadi et al. [2015])**.** *Let $\{X_i\}_{i=1}^{K}$ be a collection of random variables with $X_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ and $K = K_1 \times K_2$. Define $Y_i \triangleq \sum_{j=1}^{K_1} X_{j+(i-1)K_1}$ for $i \in [\![1, K_2]\!]$. Then, the estimator*

$$\widehat{\frac{1}{\alpha}} \triangleq \frac{1}{\log K_1}\Big(\frac{1}{K_2}\sum_{i=1}^{K_2}\log|Y_i| - \frac{1}{K}\sum_{i=1}^{K}\log|X_i|\Big). \tag{4.15}$$

converges to $1/\alpha$ almost surely, as $K_2 \to \infty$.

As shown in Theorem 2.3 of [Mohammadi et al., 2015], this estimator admits a faster convergence rate and smaller asymptotic variance than all the aforementioned methods.

In order to verify the accuracy of this estimator, we conduct a preliminary experiment, where we first generate $K = K_1 \times K_2$ many $\mathcal{S}\alpha\mathcal{S}(1)$ distributed random variables with $K_1 = 100$, $K_2 = 1000$ for 100 different values of α. Then, we estimate α by using

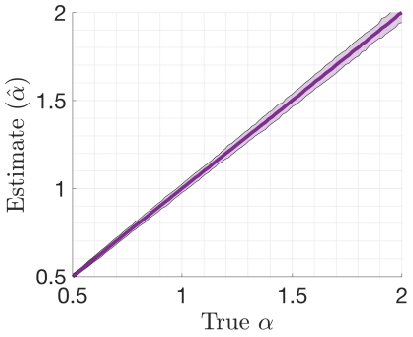


Figure 4.3: Illustration of the tail-index estimator $\hat{\alpha}$.

$\hat{\alpha} \triangleq (\widehat{\frac{1}{\alpha}})^{-1}$. We repeat this experiment 100 times for each α. As shown in Figure 4.3, the estimator is very accurate for a large range of α. Due to its favorable theoretical properties such as independence of the scale parameter σ, combined with its empirical stability, we choose this estimator in our experiments.

In order to estimate the tail-index α at iteration k, we first partition the set of data points $\mathcal{D} \triangleq \{1, \ldots, n\}$ into many disjoint sets $\Omega_k^i \subset \mathcal{D}$ of size b, such that the union of these subsets gives all the data points. Formally, for all $i, j = 1, \ldots, n/b$, $|\Omega_k^i| = b$, $\cup_i \Omega_k^i = \mathcal{D}$, and $\Omega_k^i \cap \Omega_k^j = \emptyset$ for $i \neq j$. This approach is similar to sampling without replacement. We then compute the full gradient $\nabla f(\mathbf{w}_k)$ and the stochastic gradients $\nabla \tilde{f}_{\Omega_k^i}(\mathbf{w}_k)$ for each minibatch Ω_k^i. We finally compute the stochastic gradient noises $U_k^i(\mathbf{w}_k) = \nabla \tilde{f}_{\Omega_k^i}(\mathbf{w}_k) - \nabla f(\mathbf{w}_k)$, vectorize each $U_k^i(\mathbf{w}_k)$ and concatenate them to obtain a single vector, and compute the reciprocal of (4.15). In this case, we have $K = dn/b$ and we set K_1 to the divisor of K that is the closest to $\sqrt{K}$.

4.5.2 Stability test

Besides estimating the tail-index of a random process, it is also important to verify whether the process is symmetric α-stable. In this section, we describe a procedure (Brcich et al. [2005]) for obtaining a confidence level for the stability of a random process, based on the following property:

Theorem 6 (Brcich et al. [2005])**.** *A necessary and sufficient condition for a random variable X to have an $\mathcal{S}\alpha\mathcal{S}$ distribution is*

$$X_1 + X_2 \sim C_1 X \tag{4.16}$$

$$X_1 + X_2 + X_3 \sim C_2 X \tag{4.17}$$

where $C_1, C_2 > 0$ and X_1, X_2 and X_3 are independent copies of X.

Here we adopt the stability test presented in [Brcich et al., 2005]. To obtain a statistical test from (4.16), we first separate the observations into three equal-size subsets X, X_1 and X_2, which are considered as independent copies of the observations. We then

assign the first subset X to the right side of (4.16) and estimate the tail index α_X of this subset using the idea of the previous section. For the left side of (4.16), we sum X_1 and X_2 term by term, and estimate α_{12} of the resulting sum. Similarly, by separating the observations into four equal-size subsets X', X'_1, X'_2 and X'_3, then repeating these above steps, we get $\alpha_{X'}$ from X' and α_{123} from $X'_1 + X'_2 + X'_3$, for a statistical test of (4.17). In the end, the process is considered to be α-stable if the tail indices estimated from the left and the right sides of (4.16) (as well as of (4.17)) are relatively close to each other, i.e. if the 'condition number' $c_{st} \triangleq \max\{|\alpha_X - \alpha_{12}|, |\alpha_{X'} - \alpha_{123}|\}$ is smaller than some threshold.

4.6 Numerical results

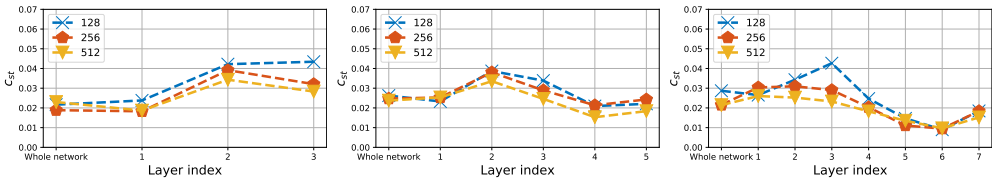
We investigate the tail behavior of the stochastic gradient noise in a variety of scenarios. We first consider a fully-connected network (FCN) on the MNIST and CIFAR10 datasets. For this model, we vary the depth (i.e. the total number of layers) in the set $\{2, 3, \ldots, 10\}$, the width (i.e. the number of neurons per hidden layer) in the set $\{2, 4, 8, \ldots, 1024\}$, and the minibatch size ranging from 1 to full batch.

We then consider a convolutional neural network (CNN) architecture (AlexNet) on the CIFAR10 and CIFAR100 datasets. We scale the number of filters in each convolutional layer in range $\{2, 4, \ldots, 512\}$. We use the existing random split of the MNIST dataset into train and test parts of sizes 60K and 10K, and CIFAR10 and CIFAR100 datasets into train and test parts of sizes 50K and 10K, respectively. The order of the total number of parameters d range from several thousands to tens of millions.

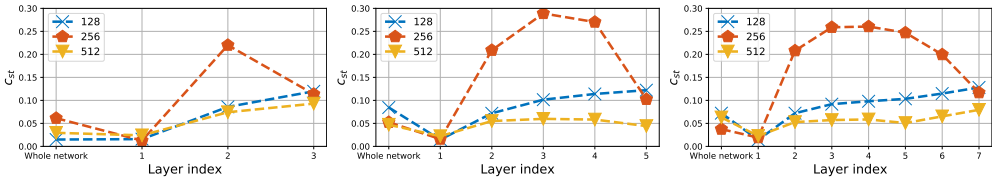
For both FCN and CNN, we run each configuration with the negative-log-likelihood (i.e. cross entropy) and with the linear hinge loss, and we repeat each experiment with three different random seeds (see [Geiger et al., 2018] for details on the choice of the hinge loss). The training algorithm is SGD with no explicit modification such as momentum or weight decay. The training runs for a fixed number of iterations unless it hits 100% training accuracy first. At every 100th iteration, we log the full training and test accuracies, and the tail estimate of the gradients that are sampled using the corresponding minibatch size. The codebase is implemented in python using pytorch [2].

Below, we present the most relevant and representative results. We have observed that, in all configurations, the three different initializations yielded no significant difference. Therefore, the effects of the randomness in initialization (under a given scheme) do not appear to affect the gradient noise. Similarly, the choice of the loss function do not yield different behaviors in terms of the tail index. Even though the heavy tailed nature remains the same, the choice of the loss function results in a different way of dependence to the hyperparameters of the system, which we discuss in Section 4.6.4 and leave the investigation to a further study.

[2]The codebase can be found at `https://github.com/umutsimsekli/sgd_tail_index`.



(a) MNIST



(b) CIFAR10

Figure 4.4: Stability confidence results for each layer as well as for whole network (indicated by layer index at 0). From left to right: depth 3, depth 5, depth 7.

4.6.1 Stability test results

We first start by investigating the stability of the stochastic gradient noises under the datasets that we use for our estimation experiments. We will first focus on the later iterations of SGD, where the tail-index becomes stationary. Using an FCN on the MNIST and CIFAR10 datasets, we estimate the condition number c_{st} (as described in Section 4.5.2) at every 50th iteration of the training stage, then take its average over the last 10K iterations to get the final result. Here, we consider $c_{st} \approx 0.05$ to be an acceptable level for the test since it is a quite small number with respect to estimated α in our experiments.

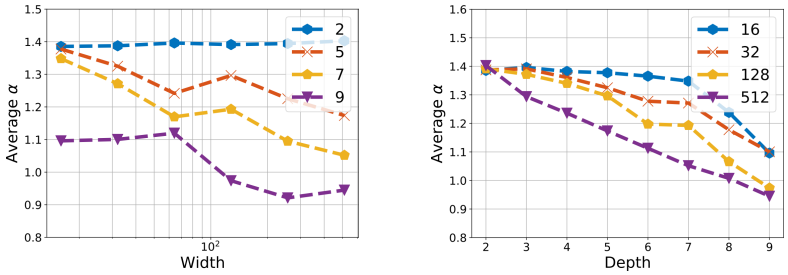
The results using the MNIST dataset are illustrated by Figure 4.4(a), in which layer index at 0 corresponds to the whole network while the indices $1, 2, \ldots, 7$ represent the hidden layers of the network. Our experiments show that the condition number c_{st} for the whole network are always smaller than the threshold 0.05, which means the gradient noise of the network satisfies our required stability criterion, even when we change the number of layers (depths) and the number of neurons per layer (widths). The same conclusion on the stability test is true when we investigate each of the hidden layers of the network.

Figure 4.4(b) shows the results of the stability test for CIFAR10. As can be seen from the figure, the condition c_{st} of the network fails to be smaller than our required criterion in some cases. However, the gap from this number to the criterion is quite small that we can consider that it does not violate the α-stable assumption on the gradient noise of the network. Unlike the MNIST dataset, we observe that for the networks with 256 neurons per layer, even though the overall gradient noise strongly exhibits an α-stable behavior, some of the hidden layers are very far from being α-stable, suggesting that the characteristics of the first layer dominate the overall structure. In contrast, the gradient

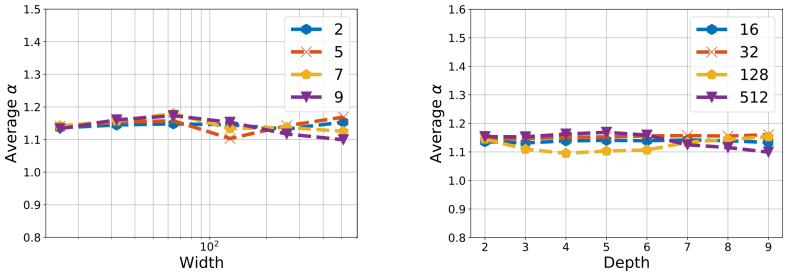
noise with respect to the parameters of the hidden layers becomes more α-stable with a very high number (512) of neurons per layer.

By these experiments, we observe that the structure of the dataset has a strong impact on the statistical properties of the gradient noise, especially for the layers with smaller number of parameters. When this number of parameters is large (which is usually the case in practice), the gradient noise corresponding to these parameters becomes more α-stable. In short, this means increasing the size of the network (the number of the network parameters) tends to make the gradient noise behave similarly to an α-stable noise.

4.6.2 Effect of varying network size



(a) MNIST



(b) CIFAR10

Figure 4.5: Estimation of α for varying widths and depths in FCN. The curves in the left figures correspond to different depths, and the ones on the right figures correspond to widths.

We measure the tail-index for varying the widths and depths for the FCN, and varying widths (i.e. the number of filters) for the CNN. For very small sizes, the networks perform poorly; therefore, we only illustrate sufficiently large network sizes, which yield similar accuracies. For these experiments, we compute the average of the tail-index measurements for the last 10K iterations (i.e. when $\hat{\alpha}$ becomes stationary) to focus on the late stage dynamics.

Figure 4.5 shows the results for the FCN. The first striking observation is that in all the cases, the estimated tail-index is far from 2, meaning that the distribution of

the gradient noise is highly non-Gaussian. For the MNIST dataset, we observe that α systematically decreases for increasing network size, where this behavior becomes more prominent with the depth. This result shows that, for MNIST, increasing the dimension of the network results in a gradient noise with heavier tails and therefore increases the probability of ending up in a wider basin. For the CIFAR10 dataset, we still observe that α is far from 2; however, in this case, increasing the network size does not have a clear effect on α. In all cases, we observe that α is in the range 1.1–1.2.

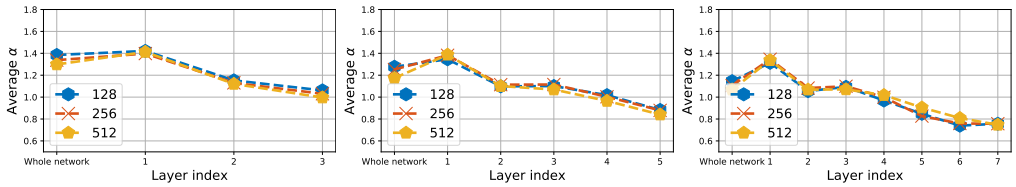
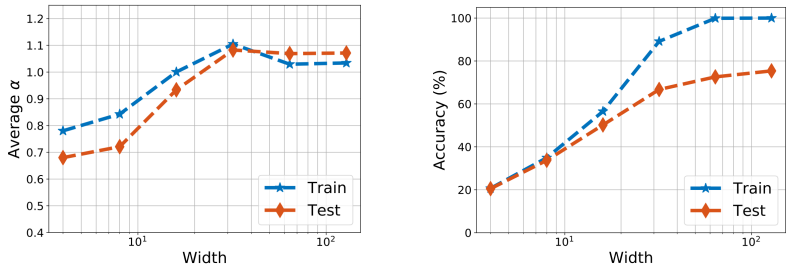
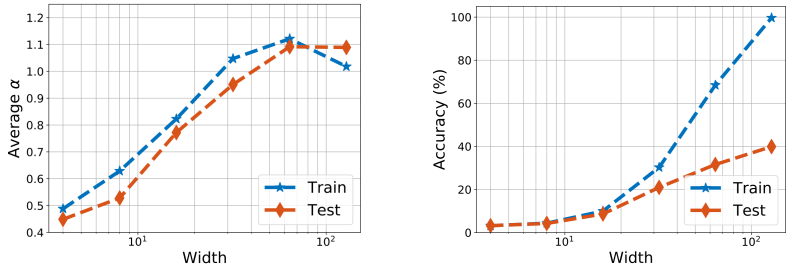


Figure 4.6: Estimation of α for varying widths and depths in FCN, dataset MNIST. From left to right: depth 3, depth 5, depth 7. Different lines correspond to different widths.

In Figure 4.6, we plot estimated α for each layer of FCNs, using MNIST dataset where the minibatch is of size 100. The resulting α is obtained by averaging α over the last 10K iterations. The layer index '0' corresponds to the estimated α of the whole network. In this experiment, we observe that α becomes smaller (heavier-tailed) for the deeper layers. In addition, the value of the tail-index for the whole network has a strong connection with the first layers: the α for the whole network is closer to that of the first layers than of the last layers.



(a) CIFAR10



(b) CIFAR100

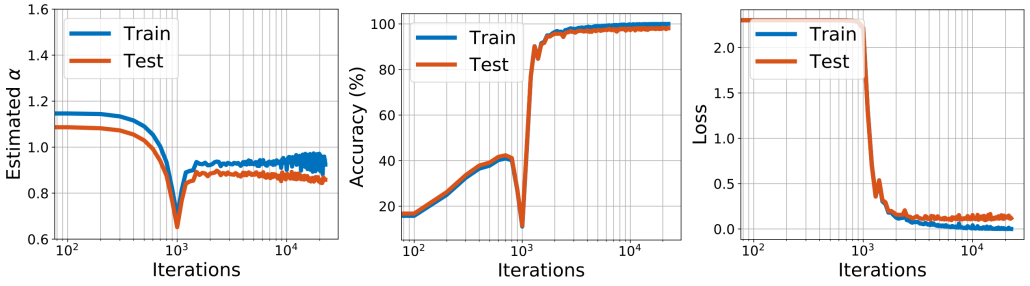
Figure 4.7: The accuracy and $\hat{\alpha}$ of the CNN for varying widths.

Figure 4.7 shows the results for the CNN. In this figure, we also depict the train and test accuracy, as well as the tail-index that is estimated on the test set. These results show that, for both CIFAR10 and CIFAR100, the tail-index is extremely low for the under-parametrized regime (e.g. the case when the width is 2, 4, or 8 for CIFAR10). As we increase the size of the network the value of α increases until the network performs reasonably well and stabilizes in the range 1.0–1.1. We also observe that α behaves similarly for both train and test sets[3].

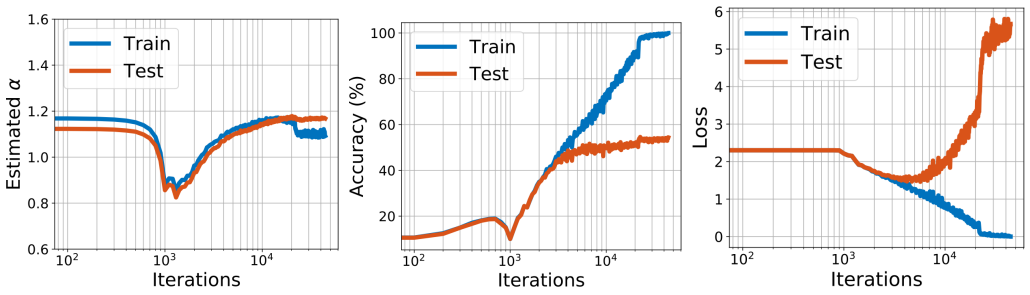
These results show that there is strong interplay between the network architecture, dataset, and the algorithm dynamics: (i) we see that the size of the network can strongly influence α, (ii) for the exact same network architecture, the choice of the dataset has a significant impact on not only the landscape of the problem, but also the noise characteristics, hence on the algorithm dynamics.

[3]We observed a similar behavior in under-parametrized FCN; however, did not plot those results to avoid clutter.

4.6.3 Tail behavior throughout the iterations



(a) MNIST



(b) CIFAR10

Figure 4.8: The iteration-wise behavior of of α for the FCN.

So far, we have focused on the late stages of SGD, where α is in a rather stationary regime. In this set of experiments, we shift our focus on the first iterations and report an intriguing behavior that we observed in almost all our experiments. As a representative, in Figure 4.8, we show the temporal evolution of SGD for the FCN with 9 layers and 512 neurons/layer.

The results clearly show that there are two distinct phases of SGD (in this configuration before and after iteration 1000). In the first phase, the loss decreases very slowly, the accuracy slightly increases, and more interestingly α rapidly decreases. When α reaches its lowest level, the process possesses a jump, which causes a sudden decrease in the accuracy. After this point the process recovers again and we see a stationary behavior in α and an increasing behavior in the accuracy.

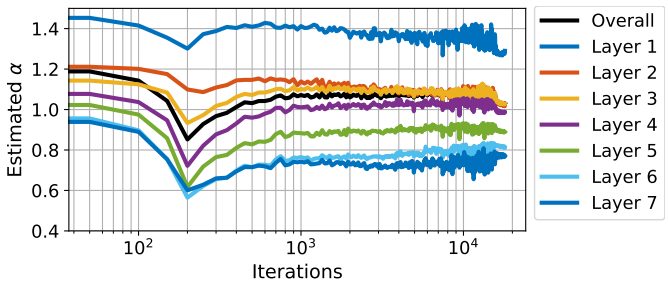


Figure 4.9: Estimation of α with an FCN on MNIST.

We also investigate this behavior for each layer of an FCN with depth 7 and width 512 in Figure 4.9. The estimated tail-index for each layer has a clear phase change at earlier iterations, where we observe that this jump is more prominent in the deeper layers where the tail-index is smaller. On the other hand, unlike the whole network, the tail-index of each layer undergoes a fluctuation period before becoming stationary at the last 2000 iterations. However, this observation might be due to the measurement error since the size of the sample that is used in the estimator (4.15) gets smaller when we make layer-wise measurements.

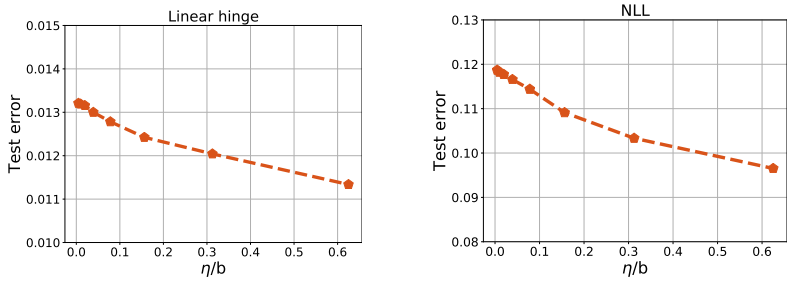
The fact that the process has a jump when α is at its smallest value provides a strong support to our assumptions and the metastability theory that we discussed in the previous section. Furthermore, these results also strengthen the view that SGD crosses barriers at the very initial phase and continues searching until it reaches a "wide and flat enough" region of a local optimum. On the other hand, our current analysis is not able to determine whether the process jumps in a different basin or a 'better' part of the same basin and we leave it as a future work.

4.6.4 A note on generalization

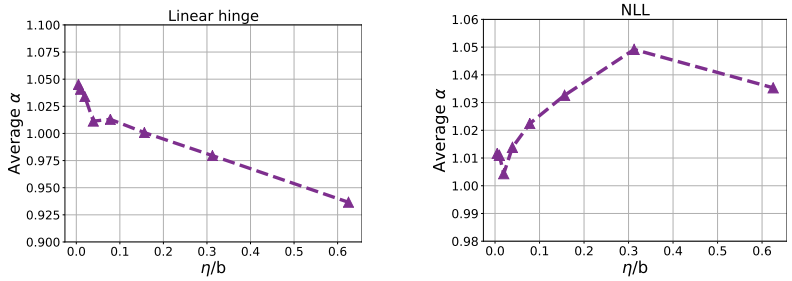
In this section, we investigate the connection between the tail-index and the generalization performance. In particular, we consider the relation between α and the ratio of the step-size to the batch size η/b which is proportional to the *noise scale* of SGD when there is no momentum [Park et al., 2019]. It has been empirically demonstrated that this ratio correlates with performance of the model [Jastrzebski et al., 2017], hence the higher the noise scale, the better the generalization performance until a certain level. Clearly, when the noise is too high, training may diverge, however, proper level of noise leads to better solutions.

In this section, we will investigate how the tail index of the gradient noise is affected for different noise scales. We reproduce and follow the initialization convention and the hyper-parameter scale that is studied in [Park et al., 2019, Appendix G]: A fully connected model with 3 hidden layers, each hidden layer has 512 nodes. Weights are initialized $\sim \mathcal{N}(0, 1)$, bias terms are set to zero at the initial point. Each layer is then

passed through ReLU non-linearity, and multiplied by the inverse of the width of the previous layer. As usual, the network is trained with SGD without momentum; the dataset is the standard MNIST. Minibatch size ranges in the set $[24, 48, 96, 192]$ and step-size ranged from the set $[0.9375, 1.875, 3.75, 7.5, 15]$[4].



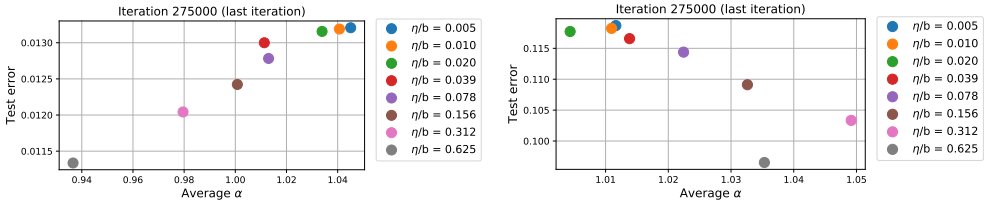
(a) Test error vs η/b



(b) Tail-index vs η/b

Figure 4.10: Test error and tail-index in accordance with the change of η/b ratio.

Figure 4.10(a) and Figure 4.10(b) visualize the results. The estimated α and the test error are averaged among the candidates with the same η/b, at the last iteration of the training process. We ignore the particular values of test error but rather focus on the way certain choices affect the trends in the system and the behavior of SGD dynamics.



(a) Linear hinge (b) NLL

Figure 4.11: Test error, estimated α in accordance with the change of η/b ratio.

In both choices of loss functions, hinge and NLL, the behavior of the test error with respect to the noise scale is consistent with previous observations. Similarly, in both

[4]Note that this particular scaling is introduced in Jacot-Guillarmod et al. [2018] and it admits slightly larger values of learning rates compared to standard initialization schemes

cases, the estimated α remains within a narrow band of 1, indicating the heavy tail behavior. However, the trends in estimated α are different depending on the choice of the loss. Therefore, we cannot attribute the improvement in performance to lower α when increasing the noise scale. To better emphasize this point, we plot the correlation of estimated α and test error in Figure 4.11 where the positive and negative correlations are clearly visible depending on the choice of the loss function. This contrasting behavior is another hint that there exists a connection between α and the test performance (since they are correlated in both cases) and suggests us to examine this connection in order to understand when exactly the dynamics falls into basins with better performance.

Chapter 5

Global convergence analysis

In this chapter, we investigate the global convergence property of stochastic gradient descent (SGD) for non-convex optimization by using a stochastic process driven by α-stable distribution. Our results show that the weak-error by SGD analyzed by this method increases faster than when analyzed using a stochastic process driven by Gaussian distribution, which suggests using smaller step-sizes.

This chapter is based on the article [Nguyen et al., 2019b].

5.1 Summary of the main result

So far, we have illustrated that stochastic processes driven by α-stable noise can be used as a proxy for understanding the dynamics of stochastic gradient descent in deep learning. Therefore, to understand the behaviors of SGD, we will provide an analysis of the non-asymptotic behavior of the following algorithm for non-convex optimization:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta c_\alpha \nabla f(\mathbf{w}^k) + \left(\frac{\eta}{\beta}\right)^{\frac{1}{\alpha}} S_{k+1}, \tag{5.1}$$

where $\alpha \in (1, 2]$, c_α is a known constant, $\beta \in \mathbb{R}_+$ is called the 'inverse temperature' parameter and $\{S_k\}_{k\in\mathbb{N}_+}$ is a sequence of α-stable distributed random variables.

Even though asymptotic convergence properties of (5.1) were established for decreasing step-sizes in Şimşekli [2017], Panloup [2008], these results do not explain the behavior of SGD algorithm for finite number of iterations. Besides, in practice, using a constant step-size often yields better performance Baker et al. [2017], a situation which cannot be handled by the existing theory.

In particular, we analyze the expected suboptimality $\mathbb{E}[f(\mathbf{w}^k) - f^\star]$, where $f^\star \triangleq f(\mathbf{w}^\star) = \min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. As we will describe in detail in Section 5.3, we decompose this suboptimality into four different terms, and we bound each of those terms one by one. Due to the choice of the α-stable Lévy motion, the standard tools for analyzing SDEs driven by a Brownian motion are not available for our use, and therefore, we cannot use

the proof strategies developed for ULA (3.2) as they are (such as Raginsky et al. [2017], Xu et al. [2018], Erdogdu et al. [2018]). Instead, we follow an alternative path, where we first relate the expected discrepancies to Wasserstein distance of fractional orders, and then, inspired by Gairing et al. [2018], we prove a result that expresses the Wasserstein distance (Definition 1) between the laws of two SDEs (driven by α-stable Lévy motion) in terms of their drift functions.

Informally, we show that the expected suboptimality $\mathbb{E}[f(\mathbf{w}^k) - f^\star]$ is bounded by a sum of four terms, summarized as follows:

$$\mathbb{E}[f(\mathbf{w}^k) - f^\star] \leq \mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4,$$

where

$$\begin{aligned}
\mathcal{A}_1 &= \mathcal{O}\Big(k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}}\Big),\\
\mathcal{A}_2 &= \mathcal{O}\Big(\frac{k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}d}{\beta^{\frac{(q-1)\gamma}{\alpha q}}}\Big),\\
\mathcal{A}_3 &= \mathcal{O}\Big(\beta + d\Big)\exp\Big(-\frac{\lambda_* k\eta}{\beta}\Big),\\
\mathcal{A}_4 &= \mathcal{O}\Big(\frac{1}{\beta^{\gamma+1}} + \frac{d}{\beta}\log(\beta+1)\Big).
\end{aligned}$$

Here $\gamma \in (0,1)$ is the Hölder exponent of the gradients of f, and $q \in (1,\alpha)$, $\lambda_* > 0$ are some constants. This result has the following implications. For any $\varepsilon > 0$,

1. If $\frac{1}{q} > \gamma + \frac{\gamma}{q}$ and $k \simeq \varepsilon^{-1}$ and $\eta < \varepsilon^{2q+1}$, then $\mathcal{A}_1$ scales as $C\varepsilon$ and $\mathcal{A}_2$ scales as $\varepsilon\text{Poly}(\beta, d)$.
2. If $\frac{1}{q} \leq \gamma + \frac{\gamma}{q}$ and $k \simeq \varepsilon^{-1}$ and $\eta < \varepsilon^{2q+\gamma+\gamma q}$, then $\mathcal{A}_1$ scales as $C\varepsilon$ and $\mathcal{A}_2$ scales as $\varepsilon\text{Poly}(\beta, d)$.
3. If we choose $k\eta > \frac{\beta}{\lambda_*}\log\Big(\frac{1}{\varepsilon}\Big)$, then $\mathcal{A}_3$ scales as $\varepsilon\text{Poly}(\beta, d)$.

where Poly(...) denotes a formal polynomial, i.e., an expression containing the real-ordered exponents of the variables, coefficients, and only the operations of addition, subtraction, and multiplication.

In Section 5.5, we extend our results in two directions: (i) obtaining guarantees for Bayesian posterior sampling and (ii) non-convex optimization where exact gradients are replaced with stochastic gradients. Our results imply that, in the context of global optimization, the error induced by (5.1) has a worse dependency on k and η, as compared to ULA. This suggests that one should use smaller step-sizes in (5.1).

5.2 Assumptions and the main result

We start by defining three different stochastic processes $\mathbf{x}_1(t)$, $\mathbf{x}_2(t)$, and $\mathbf{x}_3(t)$, which will be the main constructs in our analysis. We first informally define these processes as follows: $\mathbf{x}_2$ is a continuous-time process that interpolates $\mathbf{w}^k$ in time and it will let

us avoid dealing with the discrete-time process $\mathbf{w}^k$ directly. $\mathbf{x}_1$ is the limiting process of $\mathbf{w}^k$ when the step-size goes to zero. Finally, $\mathbf{x}_3$ is a process whose law converges to the Gibbs measure π, whose density is $\exp(-\beta f(\mathbf{w}))$ (up to a multiplicative factor).

In our approach, we will first relate $\mathbf{x}_2$ to its limiting process $\mathbf{x}_1$. Since it is more challenging to relate $\mathbf{x}_1$ to $\mathbf{w}^\star$, we will then relate $\mathbf{x}_1$ to $\mathbf{x}_3$, and $\mathbf{x}_3$ to π. By following a similar approach to Raginsky et al. [2017], we will finally relate π to $f^\star$. Formally, we decompose the expected suboptimality in the following manner:

$$\begin{aligned}\mathbb{E}f(\mathbf{w}^k) - f^* = &\Big(\mathbb{E}f(\mathbf{x}_2(k\eta)) - \mathbb{E}f(\mathbf{x}_1(k\eta))\Big) + \Big(\mathbb{E}f(\mathbf{x}_1(k\eta)) - \mathbb{E}f(\mathbf{x}_3(k\eta))\Big) \\ &+ \Big(\mathbb{E}f(\mathbf{x}_3(k\eta)) - \mathbb{E}f(\hat{\mathbf{w}})\Big) + \Big(\mathbb{E}f(\hat{\mathbf{w}}) - f^*\Big),\end{aligned} \tag{5.2}$$

where $\mathbf{x}_i(k\eta)$ with $i = 1, 2, 3$ denotes the state reached by the three stochastic processes at time $k\eta$, and $\hat{\mathbf{w}}$ is a random variable drawn from π. We will now formally define the processes $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$.

The first SDE is the continuous-time limit of the algorithm given in (5.1) and defined as follows for $t \geq 0$:

$$\mathrm{d}\mathbf{x}_1(t) = b_1(\mathbf{x}_1(t-), \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{dL}_t^\alpha, \tag{5.3}$$

where the drift function has the following form:

$$b_1(\mathbf{x}, \alpha) \triangleq -c_\alpha \nabla f(\mathbf{x}).$$

The second SDE is a *linearly interpolated* version of the discrete-time process $\{\mathbf{w}^k\}_{k\in\mathbb{N}_+}$, defined as follows:

$$\mathrm{d}\mathbf{x}_2(t) = b_2(\mathbf{x}_2, \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{dL}_t^\alpha, \tag{5.4}$$

where $\mathbf{x}_2 \equiv \{\mathbf{x}_2(t)\}_{t\geq 0}$ denotes the whole process and the drift function is chosen as follows:

$$b_2(\mathbf{x}_2, \alpha) \triangleq -c_\alpha \sum_{k=0}^{\infty} \nabla f(\mathbf{x}_2(k\eta))\mathbb{I}_{[k\eta,(k+1)\eta[}(t).$$

Here, $\mathbb{I}$ denotes the indicator function, i.e. for any set $A \subset \mathbb{R}^d$, $\mathbb{I}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{I}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$. It is easy to verify that $\mathbf{x}_2(k\eta) = \mathbf{w}^k$ for all $k \in \mathbb{N}_+$ Dalalyan [2017b], Raginsky et al. [2017].

The last SDE is designed in such a way that its solution has the Gibbs distribution as the invariant distribution and is defined as follows:

$$\mathrm{d}\mathbf{x}_3(t) = b(\mathbf{x}_3(t-), \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{dL}_t^\alpha, \tag{5.5}$$

where the drift is a d-dimensional vector whose i-th component, $i = 1, \ldots, d$, has the following form:

$$(b(\mathbf{x}, \alpha))_i \triangleq -\frac{\mathcal{D}_i^{\alpha-2}\Big(\phi(\mathbf{x})\partial_i f(\mathbf{x})\Big)}{\phi(\mathbf{x})}. \tag{5.6}$$

Here, $\mathcal{D}_i$ denotes the Riesz derivative (defined in Theorem 1) along the direction i Ortigueira et al. [2014] and ∂_i denotes the derivative with respect to the i-th component. With this definition for the drift, we have the following result for the invariant measure of $\mathbf{x}_3$, which is an extension of Theorem 1 to general d and β.

Lemma 1. *The SDE* (5.5) *with drift b defined by* (5.6) *admits π as an invariant distribution of its solution $(\mathbf{x}_3(t))_{t\geq 0}$.*

The process $\{\mathbf{x}_3(t)\}_t$ will play an important role in our analysis, since it will enable us to relate $\mathbf{x}^k$ to the Gibbs measure π, whose samples will be close to the global optimum $\mathbf{w}^\star$ with high probability Pavlyukevich [2007].

We now state our assumptions that will be used to imply the main result of the chapter.

A 1. *There exists a constant $B \geq 0$ such that*

$$c_\alpha \|\nabla f(0)\| \leq B.$$

A 2. *The gradient of f is Hölder continuous with constants $M > 0$, $0 \leq \gamma < 1$:*

$$c_\alpha \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|^\gamma, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

A 3. *For some $m > 0$ and $b \geq 0$, f is (m, b, γ)-dissipative:*

$$c_\alpha \langle \mathbf{x}, \nabla f(\mathbf{x})\rangle \geq m\|\mathbf{x}\|^{1+\gamma} - b, \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

The assumptions **A**1-3 are mild and when $\gamma = 1$, they become the standard Lipschitz and dissipativity conditions that are often considered in diffusion-based non-convex optimization algorithms Raginsky et al. [2017], Xu et al. [2018], Erdogdu et al. [2018]. However, due to the choice of the α-stable Lévy motion with $\alpha \in (1, 2)$, we need to consider a 'fractional' version of those assumptions and exclude the case where $\gamma = 1$, which makes **A**3 weaker and **A**2 more restrictive than the case where $\gamma = 1$. Nevertheless, **A**2 can be weakened to local Hölder continuity by using the localization techniques given in the proof of Proposition 4.2.2 of Kunze [2012]. This approach requires rewriting all the expressions which employ **A**2 in our proofs, by using stopping-times in such a way that we can adopt the local Hölder continuity in the same manner of Kunze [2012].

In our analysis, we will make a repeated use of the Hölder and Minkowski inequalities, which require the following condition to hold:

A 4. *There exist positive real numbers p, q, p_1, q_1 such that*

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{p_1} + \frac{1}{q_1} = 1, \text{ and}$$
$$q < \alpha, \quad \gamma p < 1, \quad \gamma q_1 < 1, \quad (q-1)p_1 < 1.$$

Even though this assumption looks rather technical, when combined with **A**2-3, it will in fact impose smoothness constraints on f and restrict γ to be less than 1. We will discuss this observation in more detail in Section 5.4.

Next, we require the drift b (presented in equation (5.5)) to be dissipative for *large distances* and we assume a bounded moment condition, which will be used for establishing the ergodicity of $\mathbf{x}_3$.

A 5. *1) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and for some constants $\bar{\gamma} \in [0, 1]$, $l_0 \geq 0$, $K_1 > 0$ and $K_2 > 0$, the following holds:*

$$\frac{\langle b(\mathbf{x}) - b(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle}{\|\mathbf{x} - \mathbf{y}\|} \leq \begin{cases} K_1 \|\mathbf{x} - \mathbf{y}\|^{\bar{\gamma}}, & \|\mathbf{x} - \mathbf{y}\| < l_0, \\ -K_2 \|\mathbf{x} - \boldsymbol{y}\|, & \|\mathbf{x} - \mathbf{y}\| \geq l_0. \end{cases}$$

2) For any $t > 0$, $\hat{\gamma} \in (0, \alpha)$, and for any coupling P_t of $\mathbf{x}_3(t)$ and $\hat{\mathbf{w}} \sim \pi$, we have:

$$\int \|\mathbf{x}_3(t) - \hat{\mathbf{w}}\|^{\hat{\gamma}} \mathrm{d}P_t < C_*,$$

for some constant $C_* > 0$.

Remark The first part of assumption **A**5 can be satisfied under a set of (rather non-trivial) assumptions (see appendix for more details). □

Proposition 1. *Under assumptions **A**1-3 and **A**5, the distribution of $\mathbf{x}_3(t)$ exponentially converges to its unique invariant distribution π in the Wasserstein metric, i.e., for any $\lambda \geq 0$ such that $\lambda < \alpha$, there exist constants $C > 0$ and $C_1 > 0$ such that*

$$\mathcal{W}_\lambda(\mu_{3t}, \pi) \leq C e^{-C_1 t}, \tag{5.7}$$

where μ_{3t} denotes the probability measure of $\mathbf{x}_3(t)$.

In the rest of the paper, we will assume that the constants C and C_1 behave similarly to the case of the unadjusted Langevin algorithm ($\alpha = 2$). In particular, we assume that C is proportional to β and C_1 is proportional to β^{-1}, so that we can rewrite (5.7) as follows:

$$\mathcal{W}_\lambda(\mu_{3t}, \pi) \leq C\beta e^{-\lambda_* t/\beta}.$$

In the unadjusted Langevin algorithm, the constant λ_* turns out to be the *uniform spectral gap* associated with the Gibbs measure π and it has shown to scale exponentially with respect to the dimension d in the worst case Raginsky et al. [2017]. We believe that a similar property holds in our case as well.

Our next assumption is on the approximation quality of the function b by b_1.

A 6. *There exists a constant $L > 0$ such that $L < m$ and*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \|c_\alpha \nabla f(\mathbf{x}) + b(\mathbf{x}, \alpha)\| \leq L,$$

where the function b is defined in (5.6).

In Corollary 2 of Şimşekli [2017], it has been shown that **A**6 holds if the tails of π vanish sufficiently quickly (cf. assumption **H**4 in Şimşekli [2017]). On the other hand, the gap between functions b and b_1 can be diminished even more if we consider a more sophisticated numerical approximation scheme, such as the one given in Çelik and Duman [2012] (cf. Theorem 2 of Şimşekli [2017]).

In our final condition, we assume that the fractional moments of π is uniformly bounded.

A 7. *There exists a constant $C > 0$ such that*

$$\int_{\mathbb{R}^d} \|\mathbf{x}\|^r \pi(\mathrm{d}\mathbf{x}) \leq C \frac{b + d/\beta}{m}$$

for all $0 \leq r \leq 2$.

Now, we are ready to state the main result of this chapter.

Theorem 7. *Under conditions **A**1-7 and for $0 < \eta < \frac{m}{M^2}$, there exists a positive constant C independent of k and η such that the following bound holds:*

$$\mathbb{E}[f(\mathbf{w}^k)] - f^* \leq C\left\{k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}} + \frac{k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}d}{\beta^{\frac{(q-1)\gamma}{\alpha q}}} + \frac{\beta b + d}{m}\exp(-\frac{\lambda_* k\eta}{\beta})\right\}$$
$$+ \frac{M c_\alpha^{-1}}{\beta^{\gamma+1}(1+\gamma)} + \frac{1}{\beta}\log\frac{(2e(b+\frac{d}{\beta}))^{\frac{d}{2}}\Gamma(\frac{d}{2}+1)\beta^d}{(dm)^{\frac{d}{2}}}.$$

More details on constant C can be found in appendix.

Theorem 7 provides finite-time guarantees for discrete-time dynamics (4.5) in terms of suboptimality with respect to the global minimum as a function of the step-size and the scale parameter. In particular, it is shown that the heavy-tailed system has a worse dependency on both K and η as compared to the Gaussian case. Besides, it is known that if the scale parameter σ gets smaller, the dynamics admits a stationary distribution that will concentrate more and more on the global minimizer although reaching out to stationary would require an exponential number of steps in the dimension in the worst case. Similar to ULA Raginsky et al. [2017], our bound grows with the number of iterations k. We note that this result sheds light on the explicit dependency of the error with respect to the algorithm parameters (e.g. step-size) for a fixed number of

iterations, rather than explaining the asymptotic behavior when k goes to infinity. In the next sections, we will provide an overview of the proof of this theorem along with some remarks and comparisons to ULA.

5.3 Proof overview

Our proof strategy consists of bounding each of the four terms in (5.2) separately. Before bounding these terms, we first start by relating the expected discrepancies to the Wasserstein distance between two random processes. The result is formally presented in the following lemma and it extends the 2-Wasserstein continuity result given in Polyanskiy and Wu [2016] to Wasserstein distance with fractional orders.

Lemma 2. *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be two random variables on* $\mathbb{R}^d$ *which have* μ *and* ν *as the probability measures and let* g *be a function in* $C^1(\mathbb{R}^d, \mathbb{R})$*. Assume that for some* $c_1 > 0, c_2 \geq 0$ *and* $0 \leq \gamma < 1$*,*

$$\|\nabla g(\mathbf{x})\| \leq c_1 \|\mathbf{x}\|^\gamma + c_2, \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

and $\max\left\{\left(\mathbb{E}\|\mathbf{w}\|^{\gamma p}\right)^{\frac{1}{p}}, \left(\mathbb{E}\|\mathbf{v}\|^{\gamma p}\right)^{\frac{1}{p}}\right\} < \infty$*. Then, the following bound holds:*

$$\left| \int g \mathrm{d}\mu - \int g \mathrm{d}\nu \right| \leq C\, \mathcal{W}_q(\mu, \nu),$$

for some $C > 0$*.*

Lemma 2 lets us upperbound the first three terms of the right hand side of (5.2) by the Wasserstein distance between the appropriate stochastic processes, respectively $\mathcal{W}_q(\mu_{1t}, \mu_{2t})$, $\mathcal{W}_q(\mu_{1t}, \mu_{3t})$, and $\mathcal{W}_q(\mu_{3t}, \pi)$, where μ_{it} denotes the law of $\mathbf{x}_i(t)$.

The term $\mathcal{W}_q(\mu_{3t}, \pi)$ is related to the ergodicity of the process (5.5) and it has been shown that this distance diminishes exponentially for a considerably large class of Lévy diffusions Masuda [2007], Xie and Zhang [2017]. On the other hand, the term $\mathcal{W}_q(\mu_{1t}, \mu_{3t})$ is related to the numerical approximation of the Riesz derivatives, which is analyzed in Şimşekli [2017]. Therefore, in this study, we use the assumptions **A**5 and **A**6 for dealing with these terms, and focus on the term $\mathcal{W}_q(\mu_{1t}, \mu_{2t})$, which is related to the so-called 'weak-error' of the Euler scheme for the SDE (5.3). The existing estimates for such weak-errors are typically of order $C\eta^a$, where $a < 1$ and C is a constant that grows exponentially with t Mikulevičius and Zhang [2011]. The exponential growth with t is prohibitive in our case and one of our main technical contributions is that, in the sequel, we will prove a bound that grows *polynomially* with t, which substantially improves over the one with exponential growth.

We start by bounding $\mathcal{W}_q(\mu_{1t}, \mu_{2t})$ and $\mathcal{W}_q(\mu_{1t}, \mu_{3t})$. In order to do so, we prove the following lemma, which will be the key for our analysis.

Lemma 3. *For $\lambda \in (1,\infty)$, $i,j \in \{1,2,3\}$ and $i \neq j$, we have the following identity:*

$$\mathcal{W}_\lambda(\mu_{it},\mu_{jt}) = \inf\Big\{\Big(\mathbb{E}\Big[\int_0^t \lambda \left\|\Delta \mathbf{x}_{ij}(s)\right\|^{\lambda-2}\langle \Delta \mathbf{x}_{ij}(s), \Delta b_{ij}(s-)\rangle \mathrm{d}s\Big]\Big)^{1/\lambda}\Big\},$$

where the infimum is taken over the couplings whose marginals are μ_{it} and μ_{jt} and

$$\begin{aligned}\Delta \mathbf{x}_{ij}(s) &\triangleq \mathbf{x}_i(s) - \mathbf{x}_j(s),\\ \Delta b_{ij}(s-) &\triangleq b_i(\mathbf{x}_i(s-),\alpha) - b_j(\mathbf{x}_j(s-),\alpha).\end{aligned}$$

This result extends the recent study Gairing et al. [2018] and lets us relate the Wasserstein distance between the distributions of the random processes to their drift functions.

By using Lemma 3, we start by bounding the Wasserstein distance between μ_{1t} and μ_{2t}. The result is summarized in the following theorem.

Theorem 8. *Assume that the following condition holds: $0 < \eta \leq \frac{m}{M^2}$. Then, we have*

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq Cq\,\mathrm{Poly}(k,\eta,\beta,d),$$

for some $C > 0$.

The full statement of the proof and the explicit constants are provided in appendix. By only considering the leading terms of the bound provided in Theorem 8, we obtain the following corollary.

Corollary 1. *Suppose that $0 < \eta < \min\left\{1, \frac{m}{M^2}\right\}$. Then, the bound for the Wasserstein distance between the laws of $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ can be written as follows:*

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq C(k^2\eta + k^2\eta^{1+\gamma/\alpha}\beta^{-(q-1)\gamma/\alpha}d).$$

By combining Corollary 1 with Lemma 2, we obtain the following result, which provides an upperbound for the first term of the right hand side of (5.2).

Corollary 2. *For $0 < \eta < \frac{m}{M^2}$, there exists a constant $C > 0$ such that the following bound holds:*

$$\left|\mathbb{E}[f(\mathbf{x}_1(k\eta))] - \mathbb{E}[f(\mathbf{x}_2(k\eta))]\right| \leq C\Big(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}} + k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d\Big).$$

Remark For any $\varepsilon > 0$, if we choose $k \simeq \varepsilon^{-1}\mathrm{Poly}(\beta,d)$ and $\eta < \varepsilon^{2q+1}\mathrm{Poly}(\beta,d)$, then the bound in Corollary 2 scales as $\varepsilon\mathrm{Poly}(\beta,d)$. □

Next, by using a similar approach, we bound the distance between μ_{1t} and μ_{3t}. In the next theorem, we show that the error grows polynomially with the parameters.

Theorem 9. *We have the following estimate:*

$$\mathcal{W}_q^q(\mu_{1t}, \mu_{3t}) \leq Cq\text{Poly}(k, \eta, \beta, d)$$

By considering the leading terms of the bound in Theorem 9 and combining it with Lemma 2, we obtain the following corollaries.

Corollary 3. *There exists a constant $C \geq 0$ such that the following bound holds:*

$$\mathcal{W}_q^q(\mu_{1t}, \mu_{3t}) \leq C(k^{q+\gamma}\eta + k^{q+\gamma}\eta^q\beta^{-\frac{q-1}{\alpha}}d).$$

Corollary 4. *There exists a constant $C \geq 0$ such that the following inequality holds:*

$$|\mathbb{E}[f(\mathbf{x}_1(k\eta))] - \mathbb{E}[f(\mathbf{x}_3(k\eta))]| \leq C\Big(k^{\gamma+\frac{\gamma+q}{q}}\eta^{\gamma+\frac{1}{q}}\beta^{-\frac{\gamma}{\alpha}}d + k^{\gamma+\frac{\gamma+q}{q}}\eta^{\frac{1}{q}}\Big).$$

Remark For any $\varepsilon > 0$, if we choose $k \simeq \varepsilon^{-1}\text{Poly}(\beta, d)$ and $\eta < \varepsilon^{2q+\gamma q+\gamma}\text{Poly}(\beta, d)$, then the bound in Corollary 4 scales as $\varepsilon\text{Poly}(\beta, d)$. □

We now pass to the term $\mathbb{E}f(\mathbf{x}_3(k\eta)) - \mathbb{E}f(\hat{\mathbf{w}})$ of (5.2). Since we have that μ_{3t} exponentially converges to π in Wasserstein distance (Proposition 1), as a direct application of Lemma 2, we obtain the following result.

Lemma 4. *Let $\hat{\mathbf{w}}$ be a random variable drawn from the invariant measure $\pi \propto \exp(-\beta f)$ of (5.5). There exists a constant $C \geq 0$ such that the following bound holds:*

$$|\mathbb{E}[f(\mathbf{x}_3(t))] - \mathbb{E}[f(\hat{\mathbf{w}})]| \leq C\frac{b\beta + d}{m}\exp(-\lambda_*\beta^{-1}t).$$

Remark For any $\varepsilon > 0$, if we take $k\eta > \frac{\beta}{\lambda_*}\log\Big(\frac{1}{\varepsilon}\Big)$, then the bound in Lemma 4 can be scaled as $\varepsilon\text{Poly}(\beta, d)$. □

We finally bound the term $\mathbb{E}f(\hat{\mathbf{w}}) - f^*$, which is the expected suboptimality of a sample from π. By following a similar proof technique presented in Raginsky et al. [2017], we obtain the following result.

Lemma 5. *For $\beta > 0$, we have*

$$\mathbb{E}[f(\hat{\mathbf{w}})] - f^\star \leq \beta^{-1}\log\Big(\frac{(2e(b+\frac{d}{\beta}))^{d/2}\Gamma(\frac{d}{2}+1)\beta^d}{(dm)^{d/2}}\Big) + \frac{\beta^{-\gamma-1}Mc_\alpha^{-1}}{1+\gamma}.$$

Combining Corollary 2, Corollary 4, Lemma 4, and Lemma 5 proves Theorem 7.

5.4 Additional remarks

In this part, we first compare our global convergence result with those for Unadjusted Langevin Algorithm presented in Raginsky et al. [2017]. Then, we discuss the feasibility of the smoothness assumption **A**4 and provide an explicit condition such that assumption **A**4 holds.

5.4.1 Comparison with ULA

Let us compare this result with those for ULA presented in Raginsky et al. [2017], since they use a similar decomposition (as opposed to Xu et al. [2018]). The last two terms of the right hand side of the bound in Theorem 7 have less importance as they can be made arbitrarily small by increasing β. Besides, for β large enough, the first two terms in our bound can be combined in a single term that scales in the order of $k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}}$. The corresponding term for ULA is given as follows: $k\eta^{5/4}$, cf. Section 3.1 of Raginsky et al. [2017]. This observation shows that (5.1) has a worse dependency both on k and η, which is not surprising and indeed in-line with the existing literature Mikulevičius and Zhang [2011].

5.4.2 Discussion on smoothness assumptions

In this section we will discuss assumption **A**4 and provide a condition on γ such that **A**4 holds. Let us recall the four constraints given in **A**4:

$$(1/p + 1/q) = (1/p_1 + 1/q_1) = 1$$
$$\gamma p < 1, \quad \gamma q_1 < 1, \quad (q-1)p_1 < 1.$$

We will refer to these conditions as the *first*, *second*, *third*, and *fourth* conditions, respectively. Our aim is to find a condition on γ (more precisely, the maximum value of γ) such that there exist $p, q, p_1, q_1 > 0$ satisfying these four conditions.

First, suppose that $p > q_1$. Then, the maximum value of γ is decided by the second constraint. Since we want γ to be as large as possible, it is natural to choose a smaller p. We observe that, as we decrease p, due to the first and the fourth constraints, the value of q_1 needs to be increased. If we continue decreasing p, then q_1 continues to be increased and soon becomes strictly greater than p. At this moment, the maximum value of γ is decided by the third constraint, not by the second constraint anymore, and from this point on, it is more plausible to decrease q_1.

By this intuition, it is reasonable to choose p to be equal to q_1, which implies that $p_1 = q$. Accordingly, the fourth constraint becomes: $(q-1)q < 1$. By noting that $q > 1$, solving this constraint gives $1 < q < (1+\sqrt{5})/2$. Then by the first constraint, we have $p > (3+\sqrt{5})/2$, and the second constraint gives $\gamma < 1/p < (3-\sqrt{5})/2$.

This upper bound for γ is a number between 0.38 and 0.39 and tells us that there exist p, q, p_1, q_1 satisfying the four constraints if and only if $0 \leq \gamma < (3-\sqrt{5})/2$.

Let us take a closer look at Theorem 7. Since $\gamma(q+1) < (3-\sqrt{5})(3+\sqrt{5})/4 = 1$, we have $\gamma + \gamma/q = \gamma(q+1)/q < 1/q$. Hence,

$$1 + \max\{1/q, \gamma + \gamma/q\} = 1 + 1/q.$$

Let ε_1 and ε_2 be positive numbers such that

$$\begin{aligned} 1/q - \varepsilon_1 &= 2/(1+\sqrt{5}) = (\sqrt{5}-1)/2, \\ \gamma + \varepsilon_2 &= (3-\sqrt{5})/2. \end{aligned}$$

then, if $q = p_1$ is approximately equal to $(1+\sqrt{5})/2$ and γ is approximately equal to $(3-\sqrt{5})/2$, we imply that ε_1 and ε_2 become very small and

$$\begin{aligned} &1/q \approx (\sqrt{5}-1)/2, \\ &1/q + \gamma/(\alpha q) \approx (\sqrt{5}-1)/2 + (\sqrt{5}-2)/\alpha, \\ &(q-1)\gamma/(q\alpha) \approx (7-3\sqrt{5})/(2\alpha). \end{aligned}$$

For example, the values $\alpha = 1.65, \gamma = 0.38, p = q_1 = 2.63, q = p_1 = q/(q-1) \approx 1.613$ satisfy assumption **A**4. Hence, the bound in Theorem 7 can be expressed as follows:

Corollary 5. *Under conditions **A**1-7, for $\alpha = 1.65, \gamma = 0.38, p = q_1 = 2.63, q = p_1 = q/(q-1) \approx 1.613$ and for $0 < \eta < \frac{m}{M^2}$, there exists a positive constant C independent of k and η such that the following bound holds:*

$$\begin{aligned} \mathbb{E}[f(\mathbf{w}^k)] - f^* \leq & C\left\{ k^{1.62}\eta^{0.61} + \frac{k^{1.62}\eta^{0.75}d}{\beta^{0.0875}} + \frac{\beta b + d}{m}\exp(-\frac{\lambda_* k\eta}{\beta}) \right\} + \frac{Mc_\alpha^{-1}}{1.38\beta^{1.38}} \\ & + \frac{1}{\beta}\log\frac{(2e(b+\frac{d}{\beta}))^{\frac{d}{2}}\Gamma(\frac{d}{2}+1)\beta^d}{(dm)^{\frac{d}{2}}}. \end{aligned}$$

Proof. The result is a direct consequence of Theorem 7. □

As a final remark on this smoothness condition, we note that similar constraints are imposed on Lévy-driven SDEs in other studies as well Panloup [2008], Şimşekli [2017]. This is due to the fact that such SDEs often require better-behaved drifts in order to be able to compensate the jumps incurred by the Lévy motion.

5.5 Extensions

In this section, we discuss the implications of our results in the classical Monte Carlo sampling context. Then, we provide a similar global convergence result for the case where the gradient is replaced by a stochastic gradient.

5.5.1 Guarantees for posterior sampling

If our aim is only to draw samples from the distribution π, then, for a fixed k, we can bound the Wasserstein distance between the law of $\mathbf{w}^k$ and π. The result is stated as follows:

Corollary 6. *For $0 < \eta \leq \frac{m}{M^2}$, the following bound holds:*

$$\mathcal{W}_q(\mu_{2t}, \pi) \leq C\Big(k^{\frac{\max\{2,q+\gamma\}}{q}}\eta^{\frac{1}{q}} + k^{\frac{\max\{2,q+\gamma\}}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{q\alpha}}\beta^{-\frac{\gamma(q-1)}{q\alpha}}d^{\frac{1}{q}} + \beta e^{-\lambda_*\frac{k\eta}{\beta}}\Big).$$

As a typical use case, we can consider Bayesian posterior sampling, where we choose $\beta = 1$ and

$$f(X) = -(\log \mathrm{P}(Y|X) + \log \mathrm{P}(X)).$$

Here, Y denotes a dataset, $\mathrm{P}(Y|X)$ is the likelihood, $\mathrm{P}(X)$ denotes the prior density, and the target distribution π becomes the posterior distribution with density $\mathrm{P}(X|Y)$.

5.5.2 Extension to stochastic gradients

In many machine learning problems, the function f to be minimized has the following form:

$$f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^{(i)}(\mathbf{x}),$$

where i denotes different data points and n is the total number of data points. In large-scale applications, n can be very large, which renders the gradient computation infeasible. Therefore, at iteration k, we often approximate ∇f by its stochastic version that is defined as follows:

$$\nabla f_k(\mathbf{x}) \triangleq \frac{1}{n_s}\sum_{i\in\Omega_k} \nabla f^{(i)}(\mathbf{x}),$$

where Ω_k is a random subset of $\{1, \ldots, n\}$ with $|\Omega_k| = n_s \ll n$. The quantity $\nabla f_k(\mathbf{x})$ is often referred to as the 'stochastic gradient'. If the stochastic gradients satisfy a moment condition, then we have the following results:

Theorem 10. *Assume that for each i, the function $\mathbf{x} \mapsto f^{(i)}(\mathbf{x})$ satisfies the conditions **A**1-7. Let us replace ∇f by ∇f_k in (5.1). If, in addition, there exists $\delta \in [0, 1)$ for any k, such that*

$$\mathbb{E}_{\Omega_k}\|c_\alpha(\nabla f(\mathbf{x}) - \nabla f_k(\mathbf{x}))\|^{q_1} \leq \delta^{q_1} M^{q_1}\|\mathbf{x}\|^{\gamma q_1},$$

for $\mathbf{x} \in \mathbb{R}^d$, then we have the following bound:

$$\mathcal{W}_q^q(\mu_{1t}, \mu_{2t}) \leq C(1+\delta)(k^2\eta + k^2\eta^{1+\gamma/\alpha}\beta^{-\gamma(q-1)/\alpha}d).$$

Similar to our previous bounds, we can use Theorem 10 for obtaining a bound for the expected discrepancy, given as follows:

Corollary 7. *Under the same assumptions as in Theorem 10, we have the following bound:*

$$\left|\mathbb{E}[f(\mathbf{x}_1(k\eta))] - \mathbb{E}[f(\mathbf{x}_2(k\eta))]\right| \leq C(1+\delta)\Big(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}} + k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d\Big).$$

These results show that the guarantees for (5.1) still hold even under the presence of stochastic gradients.

Chapter 6

First exit time analysis

While approximating SGD as a continuous-time SDE brings a new perspective for analyzing SGD, this approach might not be accurate for any step-size η, and some theoretical concerns have already been raised for the validity of such approximations Yaida [2019]. Intuitively, one can expect that the metastable behavior of SGD would be similar to the behavior of its continuous-time limit only when the discretization step-size is small enough. Even though some theoretical results have been recently established for the discretizations of SDEs driven by Brownian motion Tzen et al. [2018], it is not clear that how the discretized Lévy SDEs behave in terms of metastability. In this chapter, we provide a formal theoretical analysis where we derive explicit conditions for the step-size such that the metastability behavior of the discrete-time system is guaranteed to be close to its continuous-time limit.

This chapter is based on the article [Nguyen et al., 2019a].

6.1 First exit times of continuous-time Lévy stable SDEs

Due to the discontinuities of the Lévy-driven SDEs, their metastability behaviors also differ significantly from their Brownian counterparts. In this section, we will briefly remind some important theoretical results, that we already mentioned in Section 4.4, about the following SDE:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \varepsilon \mathrm{dL}_t^{\alpha} \tag{6.1}$$

For simplicity, let us consider this SDE in dimension one, i.e. $d = 1$. In a relatively recent study Imkeller and Pavlyukevich [2006], the authors considered this SDE, where the potential function f is required to have a non-degenerate global minimum at the origin. Their result (Theorem 4) indicates that the first exit time of w_t needs only polynomial time with respect to the *width* of the basin and it does not depend on the

depth of the basin, whereas Brownian systems need exponential time in the height of the basin in order to exit from the basin Bovier et al. [2004], Imkeller et al. [2010b]. This difference is mainly due to the discontinuities of the Lévy motion, which enables it to 'jump out' of the basin, whereas the Brownian SDEs need to 'climb' the basin due to their continuity. Consequently, given that the gradient noise exhibits similar heavy-tailed behavior to an α-stable distributed random variable, this result can be considered as a proxy to understand the wide-minima behavior of SGD.

Let us remind that (see Section 4.4) this result has already been extended to $\mathbb{R}^d$ in Imkeller et al. [2010a]. Extension to state dependent noise has also been obtained in Pavlyukevich [2011]. We also note that the metastability phenomenon is closely related to the spectral gap of the forward operator corresponding to the SDE dynamics (see e.g. Bovier et al. [2004]) and it is known that this quantity scales like $\mathcal{O}(\varepsilon^{\alpha})$ for ε small which determines the dependency to ε in the first term of the exit time (4.14) due to Kramer's Law Berglund [2011], Burghoff and Pavlyukevich [2015]. Burghoff and Pavlyukevich [2015] showed that similar scaling in ε for the spectral gap would hold if we were to restrict the SDE dynamics to a discrete grid with a small enough grid size.

6.2 The main result

In this chapter, we consider a stochastic differential equation with both a Brownian term and a Lévy term, and its Euler discretization as follows Duan [2015]:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \varepsilon\sigma\mathrm{d}\mathrm{B}_t + \varepsilon\mathrm{d}\mathrm{L}_t^{\alpha} \tag{6.2}$$

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) + \varepsilon\sigma\eta^{1/2}Z_{k+1} + \varepsilon\eta^{1/\alpha}S_{k+1}, \tag{6.3}$$

with independent and identically distributed (i.i.d.) variables $Z_k \sim \mathcal{N}(0, I)$ where I is the identity matrix, the components of S_k are i.i.d with $\mathcal{S}\alpha\mathcal{S}(1)$ distribution, and ε is the amplitude of the noise. Here, we choose σ as a scalar for convenience. We also note that the participation of the Brownian term in (6.2) enables the use of a Girsanov-like change of measures, that is convenient for our analysis.

We formally define the *first exit times*, respectively for $\mathbf{w}_t$ and $\mathbf{w}^k$ as follows:

$$\tau_{\xi,a}(\varepsilon) \triangleq \inf\{t \geq 0 : \|\mathbf{w}_t - \bar{\mathbf{w}}\| \notin [0, a+\xi]\}, \tag{6.4}$$

$$\bar{\tau}_{\xi,a}(\varepsilon) \triangleq \inf\{k \in \mathbb{N} \,:\, \|\mathbf{w}^k - \bar{\mathbf{w}}\| \notin [0, a+\xi]\}. \tag{6.5}$$

where $\bar{\mathbf{w}}$ is some local minimum and the processes are initialized at $\mathbf{w}_0 \equiv \mathbf{w}^0$ such that $\|\mathbf{w}_0 - \bar{\mathbf{w}}\| \in [0, a]$.

Understanding the metastability behavior of SGD modeled by these dynamics requires understanding the first exit times for the continuous-time process $\mathbf{w}_t$ given by (6.2) and its discretization $\mathbf{w}^k$ (6.3). Similar to (4.12), we will study the first exit times defined by (6.4) and (6.5). Note that in the special case $\xi = 0$, we recover $\tau_{0,a}(\varepsilon) = \tau_a(\varepsilon)$ introduced previously in (4.12).

Our main goal is to obtain an explicit condition on the step-size, such that the first exit time of the continuous-time process $\tau_{\xi,a}(\varepsilon)$ (6.4) would be similar to the first exit time of its Euler discretization $\bar{\tau}_{\xi,a}(\varepsilon)$ (6.5). In equations (6.2) and (6.3), σ is chosen as a scalar for convenience; however, we believe that this analysis can be extended to the case where σ is a function of $\mathbf{w}_t$.

Let us now state the assumptions which will imply our result.

A 8. *The SDE* (6.2) *admits a unique strong solution.*

A 9. *Consider the process* $\mathrm{d}\hat{\mathbf{w}}_t = g(\hat{\mathbf{w}})\mathrm{d}t + \varepsilon\sigma\mathrm{dB}_t + \varepsilon\mathrm{dL}_t^\alpha$, *where* $\hat{\mathbf{w}} \equiv \{\hat{\mathbf{w}}_t\}_{t\geq 0}$ *denotes the whole process and the drift* g *is defined as follows*[1]*:*

$$g(\hat{\mathbf{w}}) \triangleq -\sum_{k=0}^{\infty} \nabla f(\hat{\mathbf{w}}_{k\eta}) \mathbb{I}_{[k\eta,(k+1)\eta)}(t).$$

Here, $\mathbb{I}$ *denotes the indicator function, i.e.* $\mathbb{I}_S(x) = 1$ *if* $x \in S$ *and* $\mathbb{I}_S(x) = 0$ *if* $x \notin S$. *Then, the process* $\phi_t \triangleq -\frac{g(\hat{\mathbf{w}})+\nabla f(\hat{\mathbf{w}}_t)}{\varepsilon\sigma}$ *satisfies:* $\mathbb{E}\exp\left(\frac{1}{2}\int_0^T \phi_t^2 \mathrm{d}t\right) < \infty$.

A 10. *The gradient of* f *is* γ*-Hölder continuous: There exists a constant* $M > 0$ *such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|^\gamma, \qquad \forall x, y \in \mathbb{R}^d.$$

A 11. *The gradient of* f *satisfies the following assumption:* $\|\nabla f(0)\| \leq B$.

A 12. *For some* $m > 0$ *and* $b \geq 0$, f *is* (m, b, γ)*-dissipative:*

$$\langle x, \nabla f(x)\rangle \geq m\|x\|^{1+\gamma} - b, \quad \forall x \in \mathbb{R}^d.$$

We note that, as opposed to the theory of SDEs driven by Brownian motion, the theory of Lévy-driven SDEs is still an active research field where even the existence of solutions with general drift functions is not well-established and the main contributions have appeared in the last decade Priola et al. [2012], Kulik [2019]. For this reason, **A**8 has been a common assumption in stochastic analysis, e.g. Imkeller and Pavlyukevich [2006], Imkeller et al. [2010a], Liang and Wang [2018]. Nevertheless, existence and uniqueness results have been very recently established in Kulik [2019] for SDEs with bounded Hölder drifts. Therefore **A**8 and **A**9 directly hold for bounded gradients and extending this result to Hölder and dissipative drifts is out of the scope of this study. On the other hand, the assumptions **A**10-**A**12 are standard conditions, which are often considered in non-convex optimization algorithms that are based on discretization of diffusions Raginsky et al. [2017], Xu et al. [2018], Erdogdu et al. [2018], Gao et al. [2018b,a], Şimşekli et al. [2018], Liutkus et al. [2019].

The next assumption identifies an explicit condition for the step-size, which is required to make sure that the discrete process well-approximates the continuous one.

[1] It is easy to verify that $\hat{\mathbf{w}}_{k\eta} = \mathbf{w}^k$ for all $k \in \mathbb{N}_+$ [Raginsky et al., 2017].

A 13. *For a given $\delta > 0$, $t = K\eta$, and for some $C > 0$, the step-size satisfies the following condition:*

$$0 < \eta \leq \min\Big\{1, \frac{m}{M^2}, \Big(\frac{\delta^2}{2K_1t^2}\Big)^{\frac{1}{\gamma^2+2\gamma-1}}, \Big(\frac{\delta^2}{2K_2t^2}\Big)^{\frac{1}{2\gamma}}, \Big(\frac{\delta^2}{2K_3t^2}\Big)^{\frac{\alpha}{2\gamma}}, \Big(\frac{\delta^2}{2K_4t^2}\Big)^{\frac{1}{\gamma}}\Big\},$$

where ε is as in (6.3)*, the constants m, M, b are defined by* ***A****10–* ***A****12 and*

$$K_1 = \mathcal{O}(d\varepsilon^{2\gamma^2-2}), \quad K_2 = \mathcal{O}(\varepsilon^{-2}), \quad K_3 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2}), \quad K_4 = \mathcal{O}(d^{2\gamma}\varepsilon^{2\gamma-2}).$$

More explicit forms of the constants are provided in appendix.

The main result of this chapter is presented in the following theorem, its proof can be found in appendix.

Theorem 11. *Under assumptions* ***A****8-* ***A****13, the following inequality holds:*

$$\mathbb{P}[\tau_{-\xi,a}(\varepsilon) > K\eta] - C_{K,\eta,\varepsilon,d,\xi} - \delta \leq \mathbb{P}[\bar{\tau}_{0,a}(\varepsilon) > K] \leq \mathbb{P}[\tau_{\xi,a}(\varepsilon) > K\eta] + C_{K,\eta,\varepsilon,d,\xi} + \delta,$$

where,

$$C_{K,\eta,\varepsilon,d,\xi} \triangleq \frac{C_1(K\eta(d\varepsilon+1)+1)^\gamma e^{M\eta}M\eta}{\xi} + 1 - \Big(1 - Cde^{-\xi^2 e^{-2M\eta}(\varepsilon\sigma)^{-2}/(16d\eta)}\Big)^K + 1 - \Big(1 - C_\alpha d^{1+\alpha/2}\eta e^{\alpha M\eta}\varepsilon^\alpha \xi^{-\alpha}\Big)^K,$$

for some constants C_1, C_α and C that does not depend on η or ε; M is given by ***A****10 and ε is as in* (6.2)–(6.3)*.*

Remark Our result (Theorem 11) shows that with sufficiently small discretization step η (as in assumption **A**13), the probability to exit a given neighborhood of the local optimum at a fixed time t of the discretization process approximates that of the continuous process. This result also provides an explicit condition for the step-size, which explains certain impacts of the other parameters of the problem, such as dimension d, noise amplitude ε, variance of Gaussian noise σ, towards the similarity of the discretization and continuous processes. Theorem 11 enables the use of the metastability results for Lévy-driven SDEs for their discretized counterpart. □

Exit time versus problem parameters. In Theorem 11, if we let η go to zero for any δ fixed, the constant $C_{K,\eta,\varepsilon,d,\xi}$ will also go to zero, and since δ can be chosen arbitrarily small, this implies that the probability of the first exit time for the discrete process and the continuous process will approach each other when the step-size gets smaller, as expected. If instead, we decrease d or ε, the quantity $C_{K,\eta,\varepsilon,d,\xi}$ also decreases monotonically, but it does not go to zero due to the first term in the expression of $C_{K,\eta,\varepsilon,d,\xi}$.

Exit time versus width of local minima. Popular activation functions used in deep learning such as ReLU functions are almost everywhere differentiable and therefore the

cost function has a well-defined Hessian almost everywhere (see e.g. Li and Yuan [2017]). The eigenvalues of the Hessian of the objective near local minima have also been studied in the literature (see e.g. Sagun et al. [2016], Papyan [2018]). If the Hessian around a local minimum is positive definite, the conditions for the multi-dimensional version of Theorem 4 in Imkeller et al. [2010a]) are satisfied locally around a local minimum. For local minima lying in wider valleys, the parameter a can be taken to be larger; in which case the expected exit time $\mathbb{E}\tau_{0,a}(\varepsilon) \sim \mathcal{O}(a^\alpha)$ will be larger by the formula (4.14). In other words, the SDE (4.6) spends more time to exit wider valleys. Theorem 11 shows that SGD modeled by the discretization of this SDE will also inherit a similar behavior if the step-size satisfies the conditions we provide (**A**13).

6.3 Proof overview

Relating the first exit times for $\mathbf{w}_t$ (6.2) and $\mathbf{w}^k$ (6.3) often requires obtaining bounds on the distance between $\mathbf{w}_{k\eta}$ and $\mathbf{w}^k$. For this purpose, for any given local minimum $\bar{\mathbf{w}}$ of f and $a > 0$, we define the following set

$$A \triangleq \Big\{(\mathbf{w}^1, \dots, \mathbf{w}^K) \in \mathbb{R}^d \times \dots \times \mathbb{R}^d : \max_{k \leq K} \|\mathbf{w}^k - \bar{\mathbf{w}}\| \leq a\Big\}, \tag{6.6}$$

which contains the sets of K points in $\mathbb{R}^d$, each point at a distance of at most a from the local minimum $\bar{\mathbf{w}}$. We also define the following set

$$N_a \triangleq \Big\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \bar{\mathbf{w}}\| \leq a\Big\}. \tag{6.7}$$

If $\|\mathbf{w}^k - \mathbf{w}_{k\eta}\|$ is small with high probability, then we expect that their first exit times from the set N_a will be close to each other as well with high probability. For objective functions with bounded gradients, in order to relate $\tau_{\xi,a}(\varepsilon)$ to $\bar{\tau}_{\xi,a}(\varepsilon)$, one can attempt to use the strong convergence of the Euler scheme (cf. Mikulevičius and Xu [2018] Proposition 1): $\lim_{\eta \to 0} \mathbb{E}\|\mathbf{w}^k - \mathbf{w}_{k\eta}\| = 0$. By using Markov's inequality, this result implies convergence in probability: for any $\delta > 0$ and $\varepsilon > 0$, there exists η such that $\mathbb{P}(\|\mathbf{w}^k - \mathbf{w}_{k\eta}\| > \varepsilon) < \delta/2$. Then, if $\mathbf{w}_{k\eta} \in N_a$ one of the following events must happen:

1. $\mathbf{w}^k \in N_a$,
2. $\mathbf{w}^k \notin N_a$ and $\|\mathbf{w}^k - \mathbf{w}_{k\eta}\| > \epsilon$ (with probability less than $\delta/2$),
3. $\mathbf{w}^k \notin N_a$ and distance from $\mathbf{w}^k$ to N_a is at most ε (with probability less than $\delta/2$).

By using this observation, we obtain: $\mathbb{P}[\mathbf{w}_{k\eta} \in N_a] \leq \mathbb{P}[\mathbf{w}^k \in N_a] + \delta$. Even though we could use this result in order to relate $\tau_{\xi,a}(\varepsilon)$ to $\bar{\tau}_{\xi,a}(\varepsilon)$, this approach would not yield a meaningful condition for η since the bounds for the strong error $\mathbb{E}\|\mathbf{w}^k - \mathbf{w}_{k\eta}\|$ often grows exponentially in general with k, which means η should be chosen exponentially

small for a given k. Therefore, in our strategy, we choose a different path where we do not use the strong convergence of the Euler scheme.

Our proof strategy is inspired by the recent study Tzen et al. [2018], where the authors analyze the empirical metastability of the Langevin equation which is driven by a Brownian motion. However, unlike the Brownian case that Tzen et al. [2018] was based on, some of the tools for analyzing Brownian SDEs do not exist for Lévy-driven SDEs, which increases the difficulty of our task.

We first define a *linearly interpolated* version of the discrete-time process $\{\mathbf{w}^k\}_{k\in\mathbb{N}_+}$, which will be useful in our analysis, given as follows:

$$\mathrm{d}\hat{\mathbf{w}}_t = b(\hat{\mathbf{w}})\mathrm{d}t + \varepsilon\sigma dB(t) + \varepsilon\mathrm{dL}_t^{\alpha}, \tag{6.8}$$

where $\hat{\mathbf{w}} \equiv \{\hat{\mathbf{w}}_t\}_{t\geq 0}$ denotes the whole process and the *drift* function $b(\hat{\mathbf{w}})$ is chosen as follows:

$$b(\hat{\mathbf{w}}) \triangleq -\sum_{k=0}^{\infty} \nabla f(\hat{\mathbf{w}}_{k\eta})\mathbb{I}_{[k\eta,(k+1)\eta)}(t).$$

Here, $\mathbb{I}$ denotes the indicator function, i.e. $\mathbb{I}_S(x) = 1$ if $x \in S$ and $\mathbb{I}_S(x) = 0$ if $x \notin S$. It is easy to verify that $\hat{\mathbf{w}}_{k\eta} = \mathbf{w}^k$ for all $k \in \mathbb{N}_+$ Dalalyan [2017b], Raginsky et al. [2017].

In our approach, we start by developing a Girsanov-like change of measures Tankov [2003] to express the Kullback-Leibler (KL) divergence between μ_t and $\hat{\mu}_t$, which is defined as follows:

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \triangleq \int \log \frac{\mathrm{d}\hat{\mu}_t}{\mathrm{d}\mu_t} \mathrm{d}\hat{\mu}_t,$$

where μ_t denotes the law of $\{\mathbf{w}_s\}_{s\in[0,t]}$, $\hat{\mu}_t$ denotes the law of $\{\hat{\mathbf{w}}_s\}_{s\in[0,t]}$, and $\mathrm{d}\mu_t/\mathrm{d}\hat{\mu}_t$ is the Radon–Nikodym derivative of μ_t with respect to $\hat{\mu}_t$. Here, we require **A**9 for the existence of a Girsanov transform between $\hat{\mu}_t$ and μ_t and for establishing an explicit formula for the transform. In appendix, we show that the KL divergence between μ_t and $\hat{\mu}_t$ can be written as:

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) = \frac{1}{2\varepsilon^2\sigma^2}\mathbb{E}\left[\int_0^t \|b(\hat{\mathbf{w}}) + \nabla f(\hat{\mathbf{w}}_s)\|^2 \mathrm{d}s\right]. \tag{6.9}$$

While this result has been known for SDEs driven by Brownian motion Øksendal and Sulem [2005], none of the references we are aware of expressed the KL divergence as in (6.9). We also note that one of the key reasons that allows us to obtain (6.9) is the presence of the Brownian motion in (6.2), i.e. $\sigma > 0$. For $\sigma = 0$ such a measure transformation cannot be performed Debussche and Fournier [2013].

In the next result, we show that if the step-size is chosen sufficiently small, the KL divergence between μ_t and $\hat{\mu}_t$ is bounded.

Theorem 12. *Assume that the conditions **A**8-**A**13 hold. Then the following inequality holds:*

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq 2\delta^2.$$

The proof technique is similar to the approach of Dalalyan [2017b], Raginsky et al. [2017]: the idea is to divide the integral in (6.9) into smaller pieces and bounding each piece separately. Once we obtain a bound on KL, by using an optimal coupling argument, the data processing inequality, and Pinsker's inequality, we obtain a bound for the total variation (TV) distance between μ_t and $\hat{\mu}_t$ as follows:

$$\mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \neq (\hat{\mathbf{w}}_\eta, \dots, \hat{\mathbf{w}}_{K\eta})] \leq \|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV} \leq \Big(\frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta})\Big)^{\frac{1}{2}}.$$

where $\mathbf{M}$ denotes the optimal coupling between $\{\mathbf{w}_s\}_{s\in[0,K\eta]}$ and $\{\hat{\mathbf{w}}_s\}_{s\in[0,K\eta]}$, i.e., the joint probability measure of $\{\mathbf{w}_s\}_{s\in[0,K\eta]}$ and $\{\hat{\mathbf{w}}_s\}_{s\in[0,K\eta]}$, which satisfies the following identity Lindvall [2002]:

$$\mathbb{P}_{\mathbf{M}}[\{\mathbf{w}_s\}_{s\in[0,K\eta]} \neq \{\hat{\mathbf{w}}_s\}_{s\in[0,K\eta]}] = \|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV}.$$

Combined with Theorem 12, this inequality implies the following useful result:

$$\mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A] - \delta \leq \mathbb{P}[\bar{\tau}_{0,a}(\varepsilon) > K] \leq \mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A] + \delta \tag{6.10}$$

where we used the fact that the event $(\hat{\mathbf{w}}(\eta), \dots, \hat{\mathbf{w}}(K\eta)) \in A$ is equivalent to the event $(\bar{\tau}_{0,a}(\varepsilon) > K)$. The remaining task is to relate the probability $\mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A]$ to $\mathbb{P}[\tau_{\xi,a}(\varepsilon) > K\eta]$. The event $(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A$ ensures that the process $\mathbf{w}_t$ does not leave the set N_a when $t = \eta, \dots, K\eta$; however, it does not indicate that the process remains in N_a when $t \in (k\eta, (k+1)\eta)$. In order to have a control over the whole process, we introduce the following event:

$$B \triangleq \Big\{ \max_{0\leq k\leq K-1} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathbf{w}_t - \mathbf{w}_{k\eta}\| \leq \xi \Big\},$$

such that the event $[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A] \cap B$ ensures that the process stays close to N_a for the whole time. By using this event, we can obtain the following inequalities:

$$\begin{aligned}\mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A] \leq& \mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A \cap B] + \mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in B^c] \\ =& \mathbb{P}[\tau_{\xi,a}(\varepsilon) > K\eta] + \mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in B^c].\end{aligned}$$

By using the same approach, we can obtain a lower bound on $\mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in A]$ as well. Hence, our final task reduces to bounding the term $\mathbb{P}[(\mathbf{w}_\eta, \dots, \mathbf{w}_{K\eta}) \in B^c]$, which we perform by using the weak reflection principles of Lévy processes Bayraktar et al. [2015]. This finally yields Theorem 11.

6.4 Numerical illustration

To illustrate our results, we perform the experiments on a synthetic setting and the experiments on real data: a multi-layer fully connected neural network with ReLu activations on the MNIST dataset.

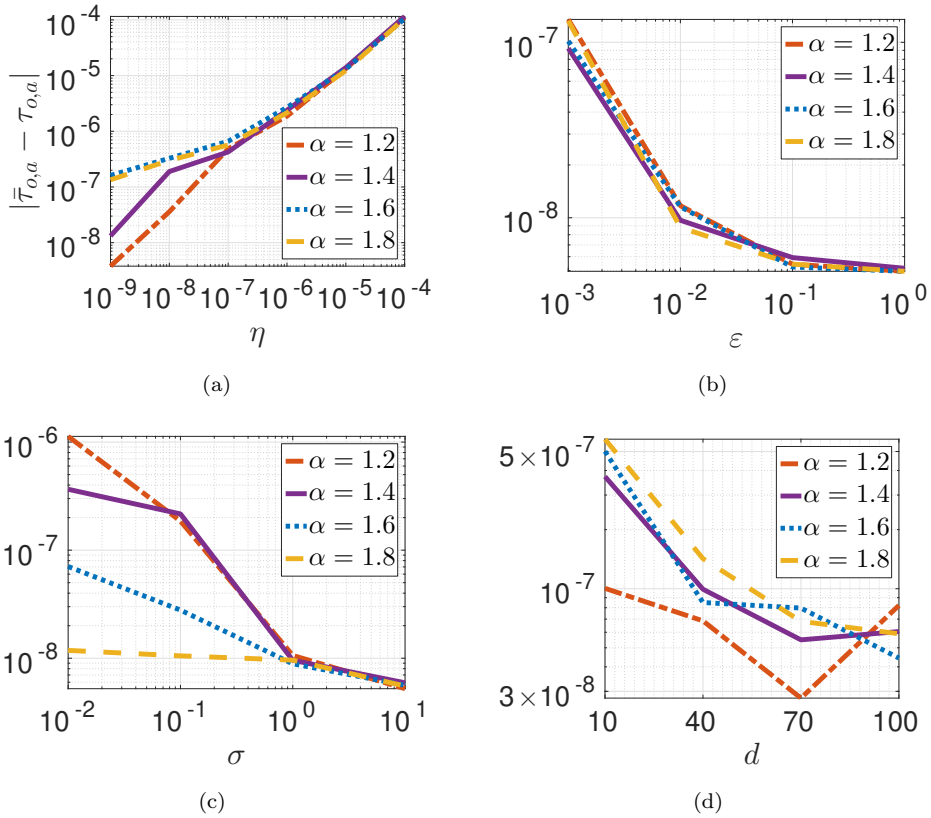


Figure 6.1: Synthetic experiments. The vertical axis is the quantity $|\bar{\tau}_{0,a} - \tau_{0,a}|$.

6.4.1 Synthetic data

We first conduct experiments on a synthetic problem, where the cost function is set to $f(x) = \frac{1}{2}\|x\|^2$. This corresponds to an Ornstein-Uhlenbeck-type process, which is commonly considered in metastability analyses Duan [2015]. This process locally satisfies the conditions **A**8-**A**12.

Since we cannot directly simulate the continuous-time process, we consider the stochastic process sampled from (6.3) with sufficiently small step-size as an approximation of the continuous scheme. Thus, we organize the experiments as follows. We first choose a very small step-size, i.e. $\eta = 10^{-10}$. Starting from an initial point $\mathbf{w}^0$ satisfying $\|\mathbf{w}^0\| < a$, we iterate (6.3) until we find the first K such that $\|\mathbf{w}^K\| > a$. We repeat this experiment 100 times, then we take the average $K\eta$ as the 'ground-truth' first exit time. We continue the experiments by calculating the first exit times for larger step-sizes (each repeated 100 times), and compute their distances to the ground truth. The detailed settings of the parameters for this experiment (Figure 6.1) can be found in appendix.

By Theorem 11, the distance between the first exit times of the discretization and the continuous processes depends on two terms $C_{K,\eta,\varepsilon,d,\bar{\varepsilon}}$ and δ, which are used for

explaining our experimental results.

We observe from Figure 6.1(a) that the error to the ground-truth first exit time is an increasing function of η, which directly matches our theoretical result. Figure 6.1(b) shows that, with small noise limit (e.g., in our settings, $\varepsilon < 1$ versus $\eta \approx 10^{-8}$), the error decreases with the parameter ε. By **A**13, with increased ε, we have the term δ to be reduced. On the other hand, $C_{K,\eta,\varepsilon,d,\bar{\varepsilon}}$ increases with ε. However, at small noise limit, this effect is dominated by the decrease of δ, that makes the error decrease overall. The decreasing speed then decelerates with larger ε, since, the product $\varepsilon\eta$ becomes so large that the increase of $C_{K,\eta,\varepsilon,d,\bar{\varepsilon}}$ starts to dominate the decrease of δ. Thus, it suggests that for a large ε, a very small step-size η would be required for reducing the distance between the first exit times of the processes. In Figure 6.1(c), the error decreases when the variance σ increases. The reason for the performance is the same as in (b), and can be explained by considering the expression of δ and $C_{K,\eta,\varepsilon,d,\bar{\varepsilon}}$ in the conclusion of Theorem 11.

In Figure 6.1(d), for small dimension, with the same exit time interval, when we increase d, both processes escape the interval earlier, with smaller exit times. Hence, the distance between their exit times becomes smaller. With larger d, the increasing effect of δ and $C_{K,\eta,\varepsilon,d,\bar{\varepsilon}}$ starts to dominate the above 'early-escape' effect, thus, the decreasing speed of the error diminish. We observe that the error even slightly increases when $\alpha = 1.2$ and d grows from 70 to 100.

6.4.2 Neural networks

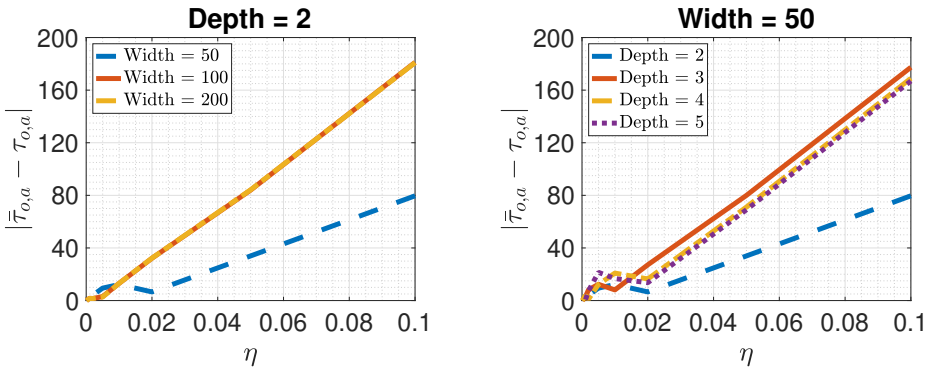


Figure 6.2: Results of the neural network experiments.

In our second set of experiments, we consider the real data setting used in Şimşekli et al. [2019]: a multi-layer fully connected neural network with ReLu activations on the MNIST dataset. We adapted the code provided in Şimşekli et al. [2019] and we provide our version in `https://github.com/umutsimsekli/sgd_first_exit_time`. For this model, we followed a similar methodology: we monitored the first exit time by varying the η, the number of layers (depth), and the number of neurons per layer (width). Since

a local minimum is not analytically available, we first trained the networks with SGD until a vicinity of a local minimum is reached with at least 90% accuracy, then we measured the first exit times with $a = 1$ and $\varepsilon = 0.1$. In order to have a prominent level of gradient noise, we set the minibatch size $b = 10$ and we did not add explicit Gaussian or Lévy noise. The result is given in Figure 6.2. We observe that, even with pure gradient noise, the error in the exit time behaves very similarly to the one that we observed in Figure 6.1(a), hence supporting our theory. We further observe that, the error has a better dependency when the width and depth are relatively small, whereas the slope of the error increases for larger width and depth. This result shows that, to inherit the metastability properties of the continuous-time SDE, we need to use a smaller η as we increase the size of the network. Note that this result does not conflict with Figure 6.1(d), since changing the width and depth does not simply change d, it also changes the landscape of the problem.

Chapter 7

Conclusion and future work

In this thesis, we investigated the tail behavior of the gradient noise in deep neural networks and empirically showed that the gradient noise is highly non-Gaussian. We analyzed the global convergence property of SGD for non-convex optimization via a stochastic process, which can be seen as a perturbed version of the gradient descent algorithm with heavy-tailed α-stable noise, for non-convex optimization and proved finite-time bounds for its expected suboptimality. Our results agreed with the existing related work, and showed that the weak-error of this algorithm increases faster than unadjusted Langevin algorithm (ULA), which suggests using smaller step-sizes.

In terms of metastability, we derived explicit conditions for the step-size such that the discrete-time SGD can inherit the metastability behavior of its continuous-time limit.

These outcomes enabled us to analyze SGD as a stochastic differential equation (SDE) driven by a Lévy motion and establish a bridge between SGD and existing theoretical results, which provides more insights on the behavior of SGD, especially in terms of choosing wide minima.

Future directions

Our study also brings up the following questions:

1. We observe that the tail-index might depend on the current state $\mathbf{w}^k$, which suggests analyzing SGD as a 'stable-like process' where the tail-index can depend on time [Bass, 1988]. However, the metastability behavior of these processes are not clear at the moment and its theory is still in an early phase [Kuhwald and Pavlyukevich, 2016].

2. At the initial point, in the over-parametrized regime with large batch sizes, the noise can in fact be of Gaussian nature (Figure 4.2). However, this property is destroyed quickly (see Neal [1996], Der and Lee [2006], Lee et al. [2018] for a discussion on the infinite width networks, and Panigrahi et al. [2019] for a discussion on the early phases

and large batches). We note that such Gaussianity heavily depends on the structure of the data, initialization scheme, and the size of the network in a sensitive way and may hold in only certain regimes or in specific cases. We think that identifying the crossover between the Gaussian and non-Gaussian regimes depending on the architecture and data is an important open problem.

3. Even though the general heavy-tailed behavior remains unchanged with the choice of the loss function, we still observe different behaviors in terms of relation to generalization (see Figure 4.11). We note that our results are related to the findings of [Martin and Mahoney, 2019], which modeled the weight matrices as heavy-tailed random matrices and investigated the density of the singular values of those matrices. Their empirical results on various different types of neural networks show that when the batch size gets smaller, the training process is able to catch finer-scale correlations from the data, leading to more strongly-correlated models between the layers of the network and that the entries of the weight matrices and the density of its singular values have heavier tails. Our results in Section 4.6.4 are partially consistent with the findings of [Martin and Mahoney, 2019]. Their results combined with ours would shed more light into the heavy-tailedness of the SGD iterates and generalization properties of SGD algorithms. In the future, we would like to investigate the underlying deeper connections between the heavy-tailed behavior and generalization further from both a mathematical and experimental perspectives.

4. We have empirically observed a heavy-tailed behavior in the stochastic gradient noise; however, it is still not (rigorously) clear what the underlying mechanism that drives this heavy-tailed behavior is. One possible idea is to consider a deep network with l layers $f(\mathbf{w}) \triangleq \mathbf{w}^{(l)}\sigma_l(\mathbf{w}^{(l-1)}\sigma_{l-1}(...\mathbf{w}^{(2)}\sigma_2(\mathbf{w}^{(1)}x_i)\dots))$ where we look for optimal weights $\mathbf{w}$ by minimizing the associated loss function using SGD algorithm, and then relate the stochastic gradient noise to a self-similar process which has a close connection with Lévy-driven SDE and α-stable distribution Pipiras and Taqqu [2017]. Additionally, there have been nice theoretical guarantees established for those self-similar processes Pipiras and Taqqu [2017] which could be useful for building a theoretical analysis of the heavy-tailed behavior of the stochastic gradient noise. Therefore, investigating this underlying mechanism would be a promising future direction.

5. Another interesting research direction concerns the multifractality, which is one of the important aspects of stochastic processes such as fractional Brownian motion, Lévy motion, etc. These processes are used for modeling phenomena in nature Bobrov et al. [2005], Pavlov et al. [2018], Teotia and Kumar [2011], which are often very irregular. Multifractal analysis consists of studying the set of the irregular points, characterized by the Hölder exponent Yang et al. [2018], of a function f. Motivated by the first exit time problem from local minima, it is also intriguing to study the connection between FLA (3.3) and the (ir)regularity of function f by using multifractality: The

probability that $\mathbf{w}^k$ in (3.3) resides in a set of low regularity points of f and the role of index α towards this probability.

6. Besides first exit time and metastability, studying the minimal time needed for a stochastic optimization algorithm (e.g. equation (4.2)) to enter a region containing a local minimum is important as well. Such minimal time is called hitting time and has been studied in recent years Zhang et al. [2017b], Chen et al. [2020] for Stochastic Gradient Langevin Dynamics, which is obtained by replacing the gradient in ULA (3.2) by stochastic gradient. Thus, extending these results for FLA (3.3) is a natural research direction and would provide a better understanding of the heavy-tailed behavior in the stochastic gradient noise.

Appendix

Supplementary materials for Chapter 5

Proof of Lemma 1

Proof. Let $q(\mathbf{x}, t)$ be the probability density of $\mathbf{x}(t)$. By Proposition 1 in Schertzer et al. [2001] (see also Section 7 of the same study), the fractional Fokker-Planck equation associated with (5.5) is given as follows:

$$\partial_t q(\mathbf{x}, t) = -\sum_{i=1}^{d} \partial_i [(b(\mathbf{x}, \alpha))_i q(\mathbf{x}, t)] - \beta^{-1} \sum_{i=1}^{d} \mathcal{D}_i^{\alpha} q(\mathbf{x}, t).$$

Using Definition (5.6) of b, we have

$$\begin{aligned}
\partial_t q(\mathbf{x}, t) &= -\sum_{i=1}^{d} \partial_i \Big[\frac{\beta^{-1} \mathcal{D}_i^{\alpha-2}(-\beta\phi(\mathbf{x})\partial_i f(\mathbf{x}))}{\phi(\mathbf{x})} q(\mathbf{x}, t) \Big] - \beta^{-1} \sum_{i=1}^{d} \mathcal{D}_i^{\alpha} q(\mathbf{x}, t) \\
&= -\sum_{i=1}^{d} \partial_i \Big[\frac{\beta^{-1} \mathcal{D}_i^{\alpha-2}(-\beta\pi(\mathbf{x})\partial_i f(\mathbf{x}))}{\pi(\mathbf{x})} q(\mathbf{x}, t) \Big] - \beta^{-1} \sum_{i=1}^{d} \mathcal{D}_i^{\alpha} q(\mathbf{x}, t) \\
&= -\sum_{i=1}^{d} \partial_i \Big[\frac{\beta^{-1} \mathcal{D}_i^{\alpha-2}(\partial_i \pi(\mathbf{x}))}{\pi(\mathbf{x})} q(\mathbf{x}, t) \Big] - \beta^{-1} \sum_{i=1}^{d} \mathcal{D}_i^{\alpha} q(\mathbf{x}, t).
\end{aligned}$$

Here, we used $\pi(\mathbf{x}) = \phi(\mathbf{x}) / \int \phi(\mathbf{x}) \mathrm{d}\mathbf{x}$ in the second equality and $-\beta \partial_i f(\mathbf{x}) = \partial_i \log \pi(\mathbf{x}) = \frac{\partial_i \pi(\mathbf{x})}{\pi(\mathbf{x})}$ in the third equality. Next, by replacing q by π on the right hand side of the above equality, we have:

$$
\begin{aligned}
-\sum_{i=1}^{d}\partial_i\Big[\frac{\beta^{-1}\mathcal{D}_i^{\alpha-2}(\partial_i\pi(\mathbf{x}))}{\pi(\mathbf{x})}\pi(\mathbf{x},t)\Big]-\beta^{-1}\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}\pi(\mathbf{x},t)=&\\
=-\sum_{i=1}^{d}\partial_i\Big[\beta^{-1}\mathcal{D}_i^{\alpha-2}(\partial_i\pi(\mathbf{x}))\Big]-\beta^{-1}\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}\pi(\mathbf{x},t)&\\
=-\sum_{i=1}^{d}\partial_i\partial_i\Big[\beta^{-1}\mathcal{D}_i^{\alpha-2}(\pi(\mathbf{x}))\Big]-\beta^{-1}\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}\pi(\mathbf{x},t)&\\
=\sum_{i=1}^{d}\mathcal{D}_i^{2}[\beta^{-1}\mathcal{D}_i^{\alpha-2}(\pi(\mathbf{x}))]-\beta^{-1}\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}\pi(\mathbf{x},t)&\\
=\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}[\beta^{-1}\pi(\mathbf{x})]-\beta^{-1}\sum_{i=1}^{d}\mathcal{D}_i^{\alpha}\pi(\mathbf{x},t)&\\
=0.&
\end{aligned}
$$

Here, we used Proposition 1 in Şimşekli [2017], $\mathcal{D}^2u(x)=-\frac{\partial}{\partial x^2}u(x)$, and the semi-group property of the Riesz derivation $\mathcal{D}^a\mathcal{D}^bu(x)=\mathcal{D}^{a+b}u(x)$. This proves that π is an invariant measure of the Markov process $(\mathbf{x}(t))_{t\geq 0}$. □

Proof of Proposition 1

Proof. By Corollary 1.3 in Liang and Wang [2018], the assumptions imply that there exist constants $\bar{C}>0$ and $\bar{C}_1>0$ such that $\mathcal{W}_1(\mu_{3t},\pi)\leq\bar{C}\beta e^{-\bar{C}_1 t}$.

Let P_{3t} be the coupling of μ_{3t} and π that such that $\mathcal{W}_1(\mu_{3t},\pi)=\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|\mathrm{d}P_{3t}$. For $0<\lambda<1$, by Hölder inequality,

$$
\begin{aligned}
\mathcal{W}_\lambda^\lambda(\mu_{3t},\pi)\leq&\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^\lambda\mathrm{d}P_{3t}\\
\leq&\Big(\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|\mathrm{d}P_{3t}\Big)^\lambda\\
=&\mathcal{W}_1^\lambda(\mu_{3t},\pi)
\end{aligned}
$$

For $\alpha>\lambda>1$,

$$
\begin{aligned}
\mathcal{W}_\lambda^\lambda(\mu_{3t},\pi)\leq&\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^\lambda\mathrm{d}P_{3t}\\
\leq&\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^\delta\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^{\lambda-\delta}\mathrm{d}P_{3t}\\
\leq&\Big(\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|\mathrm{d}P_{3t}\Big)^\delta\Big(\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^{(\lambda-\delta)/(1-\delta)}\mathrm{d}P_{3t}\Big)^{1-\delta}\\
=&\mathcal{W}_1^\delta(\mu_{3t},\pi)\Big(\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^{(\lambda-\delta)/(1-\delta)}\mathrm{d}P_{3t}\Big)^{1-\delta},
\end{aligned}
$$

where we used Holder's inequality for $\delta<1$ such that $(\lambda-\delta)/(1-\delta)<\alpha$, and $\int\|\mathbf{x}_3(t)-\hat{\mathbf{w}}\|^{(\lambda-\delta)/(1-\delta)}\mathrm{d}P_{3t}$ is bounded by a constant, by assumption **A**5.

Finally, we have

$$\mathcal{W}_\lambda(\mu_{3t}, \pi) \leq C\beta e^{-C_1 t},$$

for some constants $C, C_1 > 0$ and for $0 < \lambda < \alpha$. This completes the proof.

□

Remark Let us consider the case where the dimension d is equal to 1 (the extension for $d > 1$ is similar). The first part of assumption **A**5 can be satisfied under the following (rather non-trivial) assumptions. Assume that there exist constants $P, C_1, C_2, C_3, C_4, C_5, C_6 > 0$ such that:

$$f'(z) > 0 \text{ if } z > P, \tag{7.1}$$

$$\int_{|z|\leq P} |\phi(z) f'(z)| \mathrm{d}z = C_1 > 0 \tag{7.2}$$

$$\int_{z<-P} \phi(z) |f'(z)| |z|^{1-\alpha} \mathrm{d}z = C_2 > 0 \tag{7.3}$$

$$\int_{z>P} \phi(z) f'(z) |z|^{1-\alpha} \mathrm{d}z = C_3 > 0, \tag{7.4}$$

$$\text{if } |z| \leq P : \Big| \frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big| \leq C_4 |x-y| \ \ \forall x, y \in \mathbb{R}, \tag{7.5}$$

$$\text{if } z < -P : \Big| \frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big| \leq C_5 |x-y| |z|^{1-\alpha} \ \ \forall x, y \in \mathbb{R}, \tag{7.6}$$

$$\text{if } z > P : \Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big) \leq C_6 |z|^{1-\alpha} (y-x) \ \ \forall x, y \in \mathbb{R} \text{ s.t } x > y, \tag{7.7}$$

$$C_1 C_4 + C_2 C_5 < C_3 C_6. \tag{7.8}$$

By definition of Riesz potential, we have:

$$\begin{aligned}
b(x) - b(y) &= \int_{\mathbb{R}} \frac{\phi(z) f'(z)}{\phi(x)|x-z|^{\alpha-1}} \mathrm{d}z - \int_{\mathbb{R}} \frac{\phi(z) f'(z)}{\phi(y)|y-z|^{\alpha-1}} \mathrm{d}z \\
&= \int_{\mathbb{R}} \phi(z) f'(z) \Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big) \mathrm{d}z \\
&= \int_{|z|\leq P} \phi(z) f'(z) \Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big) \mathrm{d}z \\
&\quad + \int_{z<-P} \phi(z) f'(z) \Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big) \mathrm{d}z \\
&\quad + \int_{z>P} \phi(z) f'(z) \Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}} - \frac{1}{\phi(y)|y-z|^{\alpha-1}} \Big) \mathrm{d}z.
\end{aligned}$$

By these assumptions, we estimate the first term on the right hand side in the above

expression of $b(x) - b(y)$, for $x > y$, as follows:

$$\begin{aligned}\Big|\int_{|z|\le P}\phi(z)f'(z)\Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}}-\frac{1}{\phi(y)|y-z|^{\alpha-1}}\Big)\mathrm{d}z\Big| \le \\ \le\int_{|z|\le P}|\phi(z)f'(z)|\Big|\frac{1}{\phi(x)|x-z|^{\alpha-1}}-\frac{1}{\phi(y)|y-z|^{\alpha-1}}\Big|\mathrm{d}z \\ \le\int_{|z|\le P}|\phi(z)f'(z)|C_4|x-y|\mathrm{d}z \\ =C_1C_4|x-y| \\ =C_1C_4(x-y).\end{aligned}$$

For the remaining terms, we have:

$$\begin{aligned}\Big|\int_{z<-P}\phi(z)f'(z)\Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}}-\frac{1}{\phi(y)|y-z|^{\alpha-1}}\Big)\mathrm{d}z\Big| \le \\ \le\int_{z<-P}|\phi(z)f'(z)|\Big|\frac{1}{\phi(x)|x-z|^{\alpha-1}}-\frac{1}{\phi(y)|y-z|^{\alpha-1}}\Big|\mathrm{d}z \\ \le\int_{z<-P}\phi(z)|f'(z)|C_5|z|^{1-\alpha}|x-y|\mathrm{d}z \\ =C_2C_5|x-y| \\ =C_2C_5(x-y),\end{aligned}$$

and

$$\begin{aligned}\int_{z>P}\phi(z)f'(z)\Big(\frac{1}{\phi(x)|x-z|^{\alpha-1}}-\frac{1}{\phi(y)|y-z|^{\alpha-1}}\Big)\mathrm{d}z \le \\ \le\int_{z>P}\phi(z)f'(z)C_6|z|^{1-\alpha}(y-x)\mathrm{d}z \\ =C_3C_6(y-x) \\ =-C_3C_6(x-y).\end{aligned}$$

By combining these estimates, we get, for $x > y$:

$$b(x) - b(y) \le (C_1C_4 + C_2C_5 - C_3C_6)(x-y).$$

Thus, $(b(x) - b(y))(x-y) \le (C_1C_4 + C_2C_5 - C_3C_6)(x-y)^2$. Since $C_1C_4 + C_2C_5 - C_3C_6 < 0$, this inequality for drift b makes the first part of assumption **A**5 hold. □

Proof of Lemma 2 In this section, we precise the statement of Lemma 2 and provide the proof.

Lemma 6. *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be two random variables on* $\mathbb{R}^d$ *which have* μ *and* ν *as the probability measures and let* g *be a function in* $C^1(\mathbb{R}^d, \mathbb{R})$*. Assume that for some* $c_1 > 0, c_2 \ge 0$ *and* $0 \le \gamma < 1$*,*

$$\|\nabla g(w)\| \le c_1\|w\|^\gamma + c_2, \qquad \forall w \in \mathbb{R}^d$$

then the following bound holds:

$$\Big|\int g\mathrm{d}\mu - \int g\mathrm{d}\nu\Big| \leq \Big(c_1\Big(\mathbb{E}\|\mathbf{w}\|^{\gamma p}\Big)^{\frac{1}{p}} + c_1\Big(\mathbb{E}\|\mathbf{v}\|^{\gamma p}\Big)^{\frac{1}{p}} + c_2\Big)\mathcal{W}_q(\mu,\nu).$$

Proof. We have

$$\begin{aligned}
g(v) - g(w) &= \int_0^1 \langle w - v, \nabla g((1-t)v + tw)\rangle \mathrm{d}t \\
&\leq \int_0^1 \|w - v\| \|\nabla g((1-t)v + tw)\| \mathrm{d}t && \text{(by Cauchy-Schwarz)} \\
&\leq \int_0^1 \|w - v\| (c_1((1-t)\|v\| + t\|w\|)^\gamma + c_2) \mathrm{d}t && \text{(by the assumption on } \nabla g) \\
&\leq \|w - v\| \Big(c_1(\|v\| + \|w\|)^\gamma + c_2\Big) \\
&\leq \|w - v\| (c_1\|v\|^\gamma + c_1\|w\|^\gamma + c_2). && \text{(by Lemma 16)}
\end{aligned}$$

Now let $\mathbf{P}$ be a joint probability distribution of μ and ν that achieves $\mathcal{W}_\lambda(\mu,\nu)$, that is, $\mathbf{P} = \mathcal{L}((\mathbf{w},\mathbf{v}))$ with $\mu = \mathcal{L}(\mathbf{w})$ and $\nu = \mathcal{L}(\mathbf{v})$. We have

$$\begin{aligned}
\int g\mathrm{d}\mu - \int g\mathrm{d}\nu &= \mathbb{E}_{\mathbf{P}}[g(\mathbf{w}) - g(\mathbf{v})] \\
&\leq [\mathbb{E}_{\mathbf{P}}(c_1\|\mathbf{w}\|^\gamma + c_1\|\mathbf{v}\|^\gamma + c_2)^p]^{\frac{1}{p}} [\mathbb{E}_{\mathbf{P}}\|\mathbf{w} - \mathbf{v}\|^q]^{\frac{1}{q}} \\
&\leq \Big(c_1\Big(\mathbb{E}_{\mathbf{P}}\|\mathbf{w}\|^{\gamma p}\Big)^{\frac{1}{p}} + c_1\Big(\mathbb{E}_{\mathbf{P}}\|\mathbf{v}\|^{\gamma p}\Big)^{\frac{1}{p}} + c_2\Big)\mathcal{W}_q(\mu,\nu),
\end{aligned}$$

where we have used Holder's inequality and Minkowski's inequality. □

Proof of Lemma 3

Proof. We define a real function F_λ as follows:

$$F_\lambda(y) \triangleq \|y\|^\lambda. \tag{7.9}$$

It is clear that F_λ is a C^1 function. Let $\mathbf{y}(t) \triangleq \mathbf{x}_1(t) - \mathbf{x}_2(t)$. By the chain rule,

$$\begin{aligned}
\mathrm{d}F_\lambda(\mathbf{y}(t)) &= \langle \nabla F_\lambda(\mathbf{y}(t)), b_1(\mathbf{x}_1(t-),\alpha) - b_2(\mathbf{x}_2(t-),\alpha)\rangle \mathrm{d}t \\
&= \lambda \|\mathbf{x}_1(t) - \mathbf{x}_2(t))\|^{\lambda-2} \langle \mathbf{x}_1(t) - \mathbf{x}_2(t), b_1(\mathbf{x}_1(t-),\alpha) - b_2(\mathbf{x}_2(t-),\alpha)\rangle \mathrm{d}t.
\end{aligned} \tag{7.10}$$

By integrating both sides of (7.10) with respect to t, we arrive at

$$\begin{aligned}
F_\lambda(\mathbf{y}(t)) =& F_\lambda(\mathbf{y}(0)) + \int_0^t \lambda \|\mathbf{x}_1(t) - \mathbf{x}_2(t))\|^{\lambda-2} \langle \mathbf{x}_1(t) - \mathbf{x}_2(t), b_1(\mathbf{x}_1(t-),\alpha) - b_2(\mathbf{x}_2(t-),\alpha)\rangle \mathrm{d}s \\
=& \int_0^t \lambda \|\mathbf{x}_1(t) - \mathbf{x}_2(t))\|^{\lambda-2} \langle \mathbf{x}_1(t) - \mathbf{x}_2(t), b_1(\mathbf{x}_1(t-),\alpha) - b_2(\mathbf{x}_2(t-),\alpha)\rangle \mathrm{d}s.
\end{aligned}$$

By definition of Wasserstein distance, we have

$$\mathcal{W}_\lambda(\mu_{1t},\mu_{2t}) = \inf\{(\mathbb{E}[F_\lambda(\mathbf{y}(t))])^{1/\lambda}\},$$

which is the desired result. □

Proof of Theorem 8 In this section, we first precise the statement of Theorem 8 and then provide the corresponding proof.

Theorem 13. *Let* $\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda \triangleq l_{\alpha,\lambda,d} < \infty$. *We also define the following quantities:*

$$\begin{aligned}
P_1(\eta) &\triangleq \Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{p_1}} + (c\eta)^{\frac{1}{p_1}} + (2\eta(b+m))^{\frac{(q-1)}{2}} + 2^{\frac{(q-1)}{2}}(\eta B)^{(q-1)} \\
&\quad + \Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)}{\alpha}} l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d} + \eta^{q-1}M^{q-1}\Big((2\eta(b+m))^{\frac{(q-1)\gamma}{2}} + 2^{\frac{(q-1)\gamma}{2}}(\eta B)^{(q-1)\gamma} \\
&\quad + \Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)\gamma}{\alpha}} l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1\gamma,d}\Big), \\
P_2(\eta) &\triangleq M\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{q_1}} + (c\eta)^{\frac{1}{q_1}} + (2\eta(b+m))^{\frac{\gamma}{2}} + 2^{\frac{\gamma}{2}}(\eta B)^\gamma + \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big), \\
Q_1(\eta) &\triangleq c^{\frac{1}{p_1}} + (\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1})^{\frac{1}{p_1}} + \eta^{q-1}\Big(M^{q-1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma})^{\frac{1}{p_1}} + B^{(q-1)}\Big) \\
&\quad + \Big(\frac{\eta}{\beta}\Big)^{\frac{q-1}{\alpha}} l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}, \\
Q_2 &\triangleq M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + Mc^{\frac{1}{q_1}}.
\end{aligned}$$

Under additional assumption on the step-size: $0 < \eta \le \frac{m}{M^2}$, *we have*

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \le q\eta\Big(k^2P_1(\eta)P_2(\eta) + k^{1+1/p_1}P_1(\eta)Q_2 + k^{1+1/q_1}P_2(\eta)Q_1(\eta) + kQ_1(\eta)Q_2\Big).$$

Proof. From Lemma 3, we have

$$\begin{aligned}
&\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) = \\
&= \mathbb{E}\Big[\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_2(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_2(s), b_1(\mathbf{x}_1(s-),\alpha) - b_2(\mathbf{x}_2(s-),\alpha)\rangle \mathrm{d}s\Big] \\
&= \sum_{j=0}^{k-1}\mathbb{E}\Big[\int_{j\eta}^{(j+1)\eta} q\,\|\mathbf{x}_1(s)-\mathbf{x}_2(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_2(s), b_1(\mathbf{x}_1(s-),\alpha) - b_2(\mathbf{x}_2(s-),\alpha)\rangle \mathrm{d}s\Big] \\
&\le \sum_{j=0}^{k-1}\mathbb{E}\Big[\int_{j\eta}^{(j+1)\eta} q\,\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{q-1} c_\alpha \|\nabla f(\mathbf{x}_1(s)) - \nabla f(\mathbf{x}_2(j\eta))\| \mathrm{d}s\Big] \\
&= q\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}\Big[\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{q-1} c_\alpha \|\nabla f(\mathbf{x}_1(s)) - \nabla f(\mathbf{x}_2(j\eta))\|\Big]\mathrm{d}s \\
&\le q\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\Big[\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big]^{\frac{1}{p_1}}\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s)) - \nabla f(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}}\mathrm{d}s,
\end{aligned}$$

where we have used Cauchy-Schwarz inequality in the third line and Holder's inequality in the last line.

Since $(q-1)p_1 < 1$ by assumption **A**4, using Lemma 16 twice, we have:

$$\begin{aligned}\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \leq& \Big(\mathbb{E}\|\mathbf{x}_1(s)\|^{(q-1)p_1}+\mathbb{E}\|\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}}\\ \leq& \Big[\mathbb{E}\Big(\|\mathbf{x}_1(s)\|^{(q-1)p_1}\Big)\Big]^{\frac{1}{p_1}}+\Big[\mathbb{E}\Big(\|\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)\Big]^{\frac{1}{p_1}}.\end{aligned}$$

Then, by applying Lemma 9 and Lemma 12 for $s \in [j\eta,(j+1)\eta)$, we obtain:

$$\begin{aligned}\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \leq&\\ \leq& \Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}+\Big[\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1}+j\Big((2\eta(b+m))^{\frac{(q-1)p_1}{2}}\\ &+2^{\frac{(q-1)p_1}{2}}(\eta B)^{(q-1)p_1}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)p_1}{\alpha}}l_{\alpha,(q-1)p_1,d}\Big)\\ &+(s-j\eta)^{(q-1)p_1}\Big(M^{(q-1)p_1}\Big(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma}+j\Big((2\eta(b+m))^{\frac{(q-1)p_1\gamma}{2}}\\ &+2^{\frac{(q-1)p_1\gamma}{2}}(\eta B)^{(q-1)p_1\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)p_1\gamma}{\alpha}}l_{\alpha,(q-1)p_1\gamma,d}\Big)\Big)+B^{(q-1)p_1}\Big)\\ &+\Big(\frac{s-j\eta}{\beta}\Big)^{\frac{(q-1)p_1}{\alpha}}l_{\alpha,(q-1)p_1,d}\Big]^{\frac{1}{p_1}}.\end{aligned}$$

Next, using Lemma 16, the inequalities $j < j+1$ and $s-j\eta \leq \eta$ for $s \in [j\eta,(j+1)\eta)$, we get

$$\begin{aligned}\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \leq&\\ \leq& \Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}+(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1})^{\frac{1}{p_1}}+(j+1)^{\frac{1}{p_1}}\Big((2\eta(b+m))^{\frac{(q-1)}{2}}\\ &+2^{\frac{(q-1)}{2}}(\eta B)^{(q-1)}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}\Big)+\eta^{q-1}\Big(M^{q-1}\Big((\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma})^{\frac{1}{p_1}}\\ &+(j+1)^{\frac{1}{p_1}}\Big((2\eta(b+m))^{\frac{(q-1)\gamma}{2}}+2^{\frac{(q-1)\gamma}{2}}(\eta B)^{(q-1)\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)\gamma}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1\gamma,d}\Big)\Big)\\ &+B^{(q-1)}\Big)+\Big(\frac{\eta}{\beta}\Big)^{\frac{q-1}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}.\end{aligned}$$

We note that $s < (j+1)\eta$ and $q-1 < \frac{1}{p_1}$ (from the assumptions). Hence,

$$\begin{aligned}\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1} \leq& \Big(c\Big((j+1)\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\frac{1}{p_1}}\\ \leq& (j+1)^{\frac{1}{p_1}}\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{p_1}}+c^{\frac{1}{p_1}},\end{aligned}$$

where the last inequality is an application of Lemma 16. By replacing this inequality

into the previous one and rearranging the terms, we have

$$
\begin{aligned}
&\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}}\\
&\quad\le c^{\frac{1}{p_1}}+(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1})^{\frac{1}{p_1}}+\eta^{q-1}\Big(M^{q-1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma})^{\frac{1}{p_1}}+B^{(q-1)}\Big)\\
&\qquad+\Big(\frac{\eta}{\beta}\Big)^{\frac{q-1}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}+(j+1)^{\frac{1}{p_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{p_1}}+(2\eta(b+m))^{\frac{(q-1)}{2}}\\
&\qquad+2^{\frac{(q-1)}{2}}(\eta B)^{(q-1)}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}+\eta^{q-1}M^{q-1}\Big((2\eta(b+m))^{\frac{(q-1)\gamma}{2}}\\
&\qquad+2^{\frac{(q-1)\gamma}{2}}(\eta B)^{(q-1)\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)\gamma}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1\gamma,d}\Big)\Big)\\
&\quad\le c^{\frac{1}{p_1}}+(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1})^{\frac{1}{p_1}}+\eta^{q-1}\Big(M^{q-1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma})^{\frac{1}{p_1}}+B^{(q-1)}\Big)\\
&\qquad+\Big(\frac{\eta}{\beta}\Big)^{\frac{q-1}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}+(j+1)^{\frac{1}{p_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{p_1}}+(c\eta)^{\frac{1}{p_1}}+(2\eta(b+m))^{\frac{(q-1)}{2}}\\
&\qquad+2^{\frac{(q-1)}{2}}(\eta B)^{(q-1)}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1,d}+\eta^{q-1}M^{q-1}\Big((2\eta(b+m))^{\frac{(q-1)\gamma}{2}}\\
&\qquad+2^{\frac{(q-1)\gamma}{2}}(\eta B)^{(q-1)\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)\gamma}{\alpha}}l^{\frac{1}{p_1}}_{\alpha,(q-1)p_1\gamma,d}\Big)\Big)\\
&\quad=Q_1(\eta)+(j+1)^{\frac{1}{p_1}}P_1(\eta).
\end{aligned}
$$

Here, we have used Lemma 16 in the last inequality. Now, consider the following quantity

$$
\begin{aligned}
\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s))-\nabla f(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}}&\le\Big[\mathbb{E}\Big(M\|\mathbf{x}_1(s)-\mathbf{x}_2(j\eta)\|^{\gamma}\Big)^{q_1}\Big]^{\frac{1}{q_1}}\\
&\le\Big[\mathbb{E}\Big(M\|\mathbf{x}_1(s)\|^{\gamma}+M\|\mathbf{x}_2(j\eta)\|^{\gamma}\Big)^{q_1}\Big]^{\frac{1}{q_1}}\\
&\le\Big[\mathbb{E}\Big(M^{q_1}\|\mathbf{x}_1(s)\|^{\gamma q_1}\Big)\Big]^{\frac{1}{q_1}}+\Big[\mathbb{E}\Big(M^{q_1}\|\mathbf{x}_2(j\eta)\|^{\gamma q_1}\Big)\Big]^{\frac{1}{q_1}},
\end{aligned}
$$

where we have used assumption **A**2, Lemma 16 and Minkowski's inequality.

By Lemma 9 and Lemma 12, we have

$$
\begin{aligned}
&\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))-c_\alpha\nabla f(\mathbf{x}_2(j\eta))\|^{q_1}\Big]^{\frac{1}{q_1}}\le\\
&\qquad\le M\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+\Big[M^{q_1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})\\
&\qquad\quad+M^{q_1}j\Big((2\eta(b+m))^{\frac{\gamma q_1}{2}}+2^{\frac{\gamma q_1}{2}}(\eta B)^{\gamma q_1}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma q_1}{\alpha}}l_{\alpha,\gamma q_1,d}\Big)\Big]^{\frac{1}{q_1}}.
\end{aligned}
$$

By using Lemma 16 and the inequality $j<j+1$, we have

$$
\begin{aligned}
&\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))-c_\alpha\nabla f(\mathbf{x}_2(j\eta))\|^{q_1}\Big]^{\frac{1}{q_1}}\le\\
&\qquad\le M\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}}\\
&\qquad\quad+M(j+1)^{\frac{1}{q_1}}\Big((2\eta(b+m))^{\frac{\gamma}{2}}+2^{\frac{\gamma}{2}}(\eta B)^{\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}}l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big).
\end{aligned}
$$

We note that $s < (j+1)\eta$ and $\gamma < \frac{1}{q_1}$ (from the assumptions). Hence,

$$\begin{aligned}\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma} \leq& \Big(c\Big((j+1)\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\frac{1}{q_1}} \\ \leq& (j+1)^{\frac{1}{q_1}}\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{q_1}} + c^{\frac{1}{q_1}},\end{aligned}$$

where the last inequality is an application of Lemma 16. By replacing this inequality into the previous one and rearranging the terms, we have

$$\begin{aligned}&\Big[\mathbb{E}\|c_\alpha \nabla f(\mathbf{x}_1(s)) - c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^{q_1}\Big]^{\frac{1}{q_1}} \leq \\ &\quad\leq M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + Mc^{\frac{1}{q_1}} + M(j+1)^{\frac{1}{q_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{q_1}} \\ &\qquad + (2\eta(b+m))^{\frac{\gamma}{2}} + 2^{\frac{\gamma}{2}}(\eta B)^{\gamma} + \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big) \\ &\quad\leq M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + Mc^{\frac{1}{q_1}} + M(j+1)^{\frac{1}{q_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{q_1}} \\ &\qquad + (c\eta)^{\frac{1}{q_1}} + (2\eta(b+m))^{\frac{\gamma}{2}} + 2^{\frac{\gamma}{2}}(\eta B)^{\gamma} + \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big) \\ &\quad= Q_2 + (j+1)^{\frac{1}{q_1}} P_2(\eta).\end{aligned}$$

Here, we used Lemma 16 in the last inequality. By combining the above inequalities, we get

$$\begin{aligned}&\mathbb{E}\Big[\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_2(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_2(s), b_1(\mathbf{x}_1(s-),\alpha) - b_2(\mathbf{x}_2(s-),\alpha)\rangle \mathrm{d}s\Big] \leq \\ &\quad\leq \sum_{j=0}^{k-1} q\eta\Big((j+1)P_1(\eta)P_2(\eta) + (j+1)^{\frac{1}{p_1}}P_1(\eta)Q_2 + (j+1)^{\frac{1}{q_1}}P_2(\eta)Q_1(\eta) + Q_1(\eta)Q_2\Big) \\ &\quad\leq q\eta\Big(k^2 P_1(\eta)P_2(\eta) + k^{1+1/p_1}P_1(\eta)Q_2 + k^{1+1/q_1}P_2(\eta)Q_1(\eta) + kQ_1(\eta)Q_2\Big).\end{aligned}$$

The final conclusion follows from this inequality. □

Proof of Corollary 1

Proof. In order to get the results from the bound obtained by Theorem 13, we take the max power of k and the min power of η among the terms containing k and η but not containing β. For the terms containing β, we take the max power of k, min power of η, min power of $1/\beta$ and max power of d. We get

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq C(k^2\eta + k^2\eta^{1+\min\{\gamma,q-1\}/\alpha}\beta^{-(q-1)\gamma/\alpha}d).$$

Since $\gamma < 1/p = (q-1)/q < q-1$, we finally obtain

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq C(k^2\eta + k^2\eta^{1+\gamma/\alpha}\beta^{-(q-1)\gamma/\alpha}d).$$

□

Proof of Corollary 2

Proof. The proof starts from the bound established in Corollary 8 then, follows the same lines of the proof of Corollary 1. □

Proof of Theorem 7

Proof. We have the decomposition:

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}^k)] - f^* =& \mathbb{E}[f(\mathbf{x}_2(k\eta))] - f^* \\ =& (\mathbb{E}[f(\mathbf{x}_2(k\eta))] - \mathbb{E}[f(\mathbf{x}_1(k\eta))]) + (\mathbb{E}[f(\mathbf{x}_1(k\eta))] - \mathbb{E}[f(\mathbf{x}_3(k\eta))]) \\ &+ (\mathbb{E}[f(\mathbf{x}_3(k\eta))] - \mathbb{E}[f(\hat{\mathbf{w}})]) + (\mathbb{E}[f(\hat{\mathbf{w}})] - f^*).\end{aligned}$$

By Corollary 2, Corollary 4, Lemma 4 and Lemma 5, there exists a constant C' independent of k, η and β such that

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}^k)] - f^* \leq& C'\Big(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}} + k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d + k^{\gamma+\frac{\gamma+q}{q}}\eta^{\gamma+\frac{1}{q}}\beta^{-\frac{\gamma}{\alpha}}d \\ &+ k^{\gamma+\frac{\gamma+q}{q}}\eta^{\frac{1}{q}} + \beta\frac{b+d/\beta}{m}\exp(-\lambda_*\beta^{-1}t)\Big) + \frac{\beta^{-\gamma-1}Mc_\alpha^{-1}}{1+\gamma} \\ &+ \beta^{-1}\log\Big(\frac{(2e(b+d/\beta))^{d/2}\Gamma(d/2+1)\beta^d}{(dm)^{d/2}}\Big).\end{aligned}$$

Here, we note that $k\eta = t$. then by taking the largest power of k, smallest powers of η and β^{-1} among the terms containing all of three parameters k, η and β, there exist a constant C satisfying the following inequality:

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}^k)] - f^* \leq& C\Big(k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}} + k^{1+\max\{\frac{1}{q},\gamma+\frac{\gamma}{q}\}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d \\ &+ \beta\frac{b+d/\beta}{m}\exp(-\lambda_*\beta^{-1}k\eta)\Big) + \frac{\beta^{-\gamma-1}Mc_\alpha^{-1}}{1+\gamma} \\ &+ \beta^{-1}\log\Big(\frac{(2e(b+d/\beta))^{d/2}\Gamma(d/2+1)\beta^d}{(dm)^{d/2}}\Big).\end{aligned}$$

□

Proof of Theorem 9 In this section, we precise the statement of Theorem 9 and provide the full proof.

Theorem 14. *We have the following estimate:*

$$\begin{aligned}\mathcal{W}_q^q(\mu_{1t},\mu_{3t}) \leq& qt\Big(M(c^{q-1}+c_b^{q-1})(c^\gamma+c_b^\gamma)\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1+\gamma} \\ &+ L(c^{q-1}+c_b^{q-1})\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1}\Big),\end{aligned}$$

where c and c_b are constants defined in Lemma 9 and Lemma 10.

Proof. From Lemma 3, we have

$$\begin{aligned}
\mathcal{W}_q^q(\mu_{1t},\mu_{3t}) =&\mathbb{E}\Big[\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_3(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_3(s), b_1(\mathbf{x}_1(s-),\alpha)-b(\mathbf{x}_3(s-),\alpha)\rangle \mathrm{d}s\Big]\\
=&\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_3(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_3(s), b_1(\mathbf{x}_1(s-),\alpha)-b(\mathbf{x}_3(s-),\alpha)\rangle \mathrm{d}s\\
\leq&\mathbb{E}\Big[\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{q-1}\|c_\alpha\nabla f(\mathbf{x}_1(s))+b(\mathbf{x}_3(s),\alpha)\|\mathrm{d}s\Big]\\
=&q\int_0^t \mathbb{E}\Big[\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{q-1}\|c_\alpha\nabla f(\mathbf{x}_1(s))+b(\mathbf{x}_3(s),\alpha)\|\Big]\mathrm{d}s\\
\leq&q\int_0^t \Big[\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{(q-1)p_1}\Big]^{\frac{1}{p_1}}\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))+b(\mathbf{x}_3(s),\alpha)\|^{q_1}\Big]^{\frac{1}{q_1}}\mathrm{d}s,
\end{aligned}$$

where we have used Cauchy-Schwarz inequality in the third line and Holder's inequality in the last line.

Since $(q-1)p_1<1$ by assumption **A**4, using Lemma 16 twice, we have:

$$\begin{aligned}
\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \leq&\Big(\mathbb{E}\|\mathbf{x}_1(s)|^{(q-1)p_1}+\mathbb{E}\|\mathbf{x}_3(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}}\\
\leq&\Big[\mathbb{E}\Big(\|\mathbf{x}_1(s)\|^{(q-1)p_1}\Big)\Big]^{\frac{1}{p_1}}+\Big[\mathbb{E}\Big(\|\mathbf{x}_3(s)\|^{(q-1)p_1}\Big)\Big]^{\frac{1}{p_1}}.
\end{aligned}$$

Then, by applying Lemma 9 and Lemma 10 we obtain:

$$\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \leq \Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}+\Big(c_b\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}.$$

Now, consider the following quantity

$$\begin{aligned}
&\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))+b(\mathbf{x}_3(s),\alpha)\|^{q_1}\Big]^{\frac{1}{q_1}} \leq\\
&\qquad\leq\Big[\mathbb{E}\big(\|c_\alpha\nabla f(\mathbf{x}_1(s))-c_\alpha\nabla f(\mathbf{x}_3(s))\|+\|c_\alpha\nabla f(\mathbf{x}_3(s))+b(\mathbf{x}_3(s),\alpha)\|\big)^{q_1}\Big]^{\frac{1}{q_1}}\\
&\qquad\leq\Big[\mathbb{E}\big(M\|\mathbf{x}_1(s)-\mathbf{x}_3(s)\|^{\gamma}+L\big)^{q_1}\Big]^{\frac{1}{q_1}}\\
&\qquad\leq\Big[\mathbb{E}\Big(M\|\mathbf{x}_1(s)\|^{\gamma}+M\|\mathbf{x}_3(s)\|^{\gamma}+L\Big)^{q_1}\Big]^{\frac{1}{q_1}}\\
&\qquad\leq\Big[\mathbb{E}\Big(M^{q_1}\|\mathbf{x}_1(s)\|^{\gamma q_1}\Big)\Big]^{\frac{1}{q_1}}+\Big[\mathbb{E}\Big(M^{q_1}\|\mathbf{x}_3(s)\|^{\gamma q_1}\Big)\Big]^{\frac{1}{q_1}}+L,
\end{aligned}$$

where we have used assumption **A**2, assumption **A**6, Lemma 16 and Minkowski's inequality. By Lemma 9 and Lemma 10, we have

$$\begin{aligned}
&\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))+b(\mathbf{x}_3(s),\alpha)\|^{q_1}\Big]^{\frac{1}{q_1}} \leq\\
&\qquad\leq M\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+M\Big(c_b\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+L.
\end{aligned}$$

By combining the above inequalities, we get

$$
\begin{aligned}
&\mathbb{E}\Big[\int_0^t q\,\|\mathbf{x}_1(s)-\mathbf{x}_3(s))\|^{q-2}\langle \mathbf{x}_1(s)-\mathbf{x}_3(s), b_1(\mathbf{x}_1(s-),\alpha)-b(\mathbf{x}_3(s-),\alpha)\rangle \mathrm{d}s\Big] \leq \\
&\leq q\int_0^t \Big(\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}+\Big(c_b\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{q-1}\Big)\Big(M\Big(c\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+ \\
&\quad +1\Big)+1\Big)\Big)^{\gamma}+M\Big(c_b\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+L\Big)\mathrm{d}s \\
&= q\int_0^t \Big(M(c^{q-1}+c_b^{q-1})(c^{\gamma}+c_b^{\gamma})\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1+\gamma}+L(c^{q-1}+c_b^{q-1})\Big(s\Big(\frac{d}{\beta^{1/\alpha}}+ \\
&\quad +1\Big)+1\Big)^{q-1}\Big)\mathrm{d}s \\
&\leq qt\Big(M(c^{q-1}+c_b^{q-1})(c^{\gamma}+c_b^{\gamma})\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1+\gamma}+L(c^{q-1}+c_b^{q-1})\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+ \\
&\quad +1\Big)^{q-1}\Big).
\end{aligned}
$$

The final conclusion follows from this inequality. □

Proof of Corollary 3

Proof. First, we replace t by $k\eta$. Then, by following the same lines of the proof of Corollary 1, we get

$$\mathcal{W}_q^q(\mu_{1t},\mu_{3t}) \leq C(k^{q+\gamma}\eta + k^{q+\gamma}\eta^q\beta^{-\frac{q-1}{\alpha}}d^{q-1+\gamma}).$$

By assumption **A**4, $q-1<1/p_1$ and $\gamma<1/q_1$. It implies that $d^{q-1+\gamma}<d^{1/p_1+1/q_1}=d$. Hence, we have

$$\mathcal{W}_q^q(\mu_{1t},\mu_{3t}) \leq C(k^{q+\gamma}\eta + k^{q+\gamma}\eta^q\beta^{-\frac{q-1}{\alpha}}d).$$

□

Proof of Corollary 4

Proof. By Lemma 2, Lemma 9 and Lemma 10, we have

$$
\begin{aligned}
c_\alpha|\mathbb{E}[f(\mathbf{x}_1(t))]-\mathbb{E}[f(\mathbf{x}_3(t))]| \leq& \\
&\leq\Big(M\big(\mathbb{E}\|\mathbf{x}_1(t)\|^{\gamma p}\big)^{\frac{1}{p}}+M\big(\mathbb{E}\|\mathbf{x}_3(t)\|^{\gamma p}\big)^{\frac{1}{p}}+B\Big)\mathcal{W}_q(\mu_{1t},\mu_{3t}) \\
&\leq\Big(M\Big(c\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+M\Big(c_b\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+B\Big)\mathcal{W}_q(\mu_{1t},\mu_{3t}).
\end{aligned}
$$

Then by Theorem 9, we have

$$
\begin{aligned}
&c_\alpha|\mathbb{E}[f(\mathbf{x}_1(t))] - \mathbb{E}[f(\mathbf{x}_3(t))]| \\
&\qquad \leq \Big(M\Big(c\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^\gamma + M\Big(c_b\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^\gamma + \\
&\qquad\quad + B\Big)\Bigg(qt\Big(M(c^{q-1}+c_b^{q-1})(c^\gamma+c_b^\gamma)\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1+\gamma} + \\
&\qquad\quad + L(c^{q-1}+c_b^{q-1})\Big(t\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)^{q-1}\Big)\Bigg)^{\frac{1}{q}}.
\end{aligned}
$$

Applying Lemma 16 twice, we get

$$
\begin{aligned}
&c_\alpha|\mathbb{E}[f(\mathbf{x}_1(t))] - \mathbb{E}[f(\mathbf{x}_3(t))]| \leq \\
&\leq \Big(M(c^\gamma+c_b^\gamma)\Big(\frac{t^\gamma d^\gamma}{\beta^{\gamma/\alpha}}+t^\gamma+1\Big)+B\Big)\Bigg((qt)^{1/q}\Big(M^{1/q}(c^{q-1}+c_b^{q-1})^{1/q}(c^\gamma+c_b^\gamma)^{1/q} \\
&\quad \Big(\frac{td}{\beta^{1/\alpha}}+t+1\Big)^{(q-1+\gamma)/q} + L^{1/q}(c^{q-1}+c_b^{q-1})^{1/q}\Big(\frac{td}{\beta^{1/\alpha}}+t+1\Big)^{(q-1)/q}\Big)\Bigg) \\
&\leq \Big(M(c^\gamma+c_b^\gamma)\Big(\frac{t^\gamma d^\gamma}{\beta^{\gamma/\alpha}}+t^\gamma+1\Big)+B\Big)\Bigg((qt)^{1/q}\Big(M^{1/q}(c^{q-1}+c_b^{q-1})^{1/q}(c^\gamma+c_b^\gamma)^{1/q} \\
&\quad \Big(\frac{(td)^{(q-1+\gamma)/q}}{\beta^{(q-1+\gamma)/(q\alpha)}}+t^{(q-1+\gamma)/q}+1\Big) + L^{1/q}(c^{q-1}+c_b^{q-1})^{1/q}\Big(\frac{(td)^{(q-1)/q}}{\beta^{(q-1)/(q\alpha)}}+t^{(q-1)/q}+1\Big)\Big)\Bigg).
\end{aligned}
$$

Now, by replacing $t = k\eta$ we find that, among the terms containing β, the largest power of d, the largest power of k and the smallest power of η are $\gamma + \frac{q-1+\gamma}{q}$, $\gamma + \frac{\gamma+q}{q}$ and $\gamma + \frac{1}{q}$, respectively. For the smallest power of β^{-1}, we need to compare the following quantities: γ/α, $(q-1+\gamma)/(q\alpha)$ and $(q-1)/(q\alpha)$.

It is obvious that $(q-1+\gamma)/(q\alpha) > (q-1)/(q\alpha)$. Next, from the relation $\gamma < 1/p = (q-1)/q$, we have $\gamma/\alpha < (q-1)/(q\alpha)$. Thus, the smallest power of β^{-1} is γ/α. Hence, we have the following bound:

$$
c_\alpha|\mathbb{E}[f(\mathbf{x}_1(t))] - \mathbb{E}[f(\mathbf{x}_3(t))]| \leq C\Big(k^{\gamma+\frac{\gamma+q}{q}}\eta^{\gamma+\frac{1}{q}}\beta^{-\frac{\gamma}{\alpha}}d^{\gamma+\frac{q-1+\gamma}{q}} + k^{\gamma+\frac{\gamma+q}{q}}\eta^{\frac{1}{q}}\Big),
$$

for some constant $C > 0$. For the power of d, using that $\gamma < 1/p$, $q-1 < 1/p_1$ and $\gamma < 1/q_1$ we have

$$
\gamma + \frac{q-1+\gamma}{q} \leq 1/p + \frac{1/p_1 + 1/q_1}{q} = 1/p + 1/q = 1.
$$

Finally, we have

$$
c_\alpha|\mathbb{E}[f(\mathbf{x}_1(t))] - \mathbb{E}[f(\mathbf{x}_3(t))]| \leq C\Big(k^{\gamma+\frac{\gamma+q}{q}}\eta^{\gamma+\frac{1}{q}}\beta^{-\frac{\gamma}{\alpha}}d + k^{\gamma+\frac{\gamma+q}{q}}\eta^{\frac{1}{q}}\Big).
$$

□

Proof of Lemma 4

Proof. By Lemma 2, we have

$$c_\alpha|\mathbb{E}[f(\mathbf{x}_3(t))] - \mathbb{E}[f(\hat{\mathbf{w}})]| \leq \Big(M\big(\mathbb{E}\|\mathbf{x}_3(t)\|^{\gamma p}\big)^{\frac{1}{p}} + M\big(\mathbb{E}\|\hat{\mathbf{w}}\|^{\gamma p}\big)^{\frac{1}{p}} + B\Big)\mathcal{W}_q(\mu_{3t}, \pi).$$

Assumption **A**7 says that $\mathbb{E}\|\hat{\mathbf{w}}\|^{\gamma p}$ is bounded by a constant depending on b, m and β. In addition, by Proposition 1, $\lim_{t\to\infty} \mathcal{W}_{\gamma p}(\mu_{3t}, \pi) = 0$, and by Theorem 7.12 in Villani [2003], it follows that

$$\lim_{t\to\infty} \mathbb{E}\|\mathbf{x}_3(t)\|^{\gamma p} = \mathbb{E}\|\hat{\mathbf{w}}\|^{\gamma p}.$$

Thus, $\mathbb{E}\|\mathbf{x}_3(t)\|^{\gamma p}$ is bounded by a constant independent of t. Finally, since $q < \alpha$, by propostion 1 again, $\mathcal{W}_q(\mu_{3t}, \pi) \leq C\beta e^{-\lambda_* t/\beta}$. Hence, using the bound in assumption **A**7, there exists constant C such that

$$|\mathbb{E}[f(\mathbf{x}_3(t))] - \mathbb{E}[f(\hat{\mathbf{w}})]| \leq C\beta\frac{b + d/\beta}{m}\exp(-\lambda_*\beta^{-1}t).$$

□

Proof of Lemma 5

Proof. The proof is adapted from Raginsky et al. [2017], Section 3.5. First, we have the decomposition:

$$\begin{aligned}\mathbb{E}[f(\hat{\mathbf{w}})] =& \int_{\mathbb{R}^d} f(w)\frac{\exp(-\beta f(w))}{\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v}\mathrm{d}w \\ =& \frac{1}{\beta}\Big(-\int_{\mathbb{R}^d}\frac{\exp(-\beta f(w))}{\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v}\log\frac{\exp(-\beta f(w))}{\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v}\mathrm{d}w - \log\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v\Big).\end{aligned}$$

The first term in the parentheses is the differential entropy of the probability density of $\hat{\mathbf{w}}$, which has a finite second moment (due to assumption **A**7). Hence, it is upper-bounded by the differential entropy of a Gaussian density with the same second moment:

$$-\int_{\mathbb{R}^d}\frac{\exp(-\beta f(w))}{\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v}\log\frac{\exp(-\beta f(w))}{\int_{\mathbb{R}^d}\exp(-\beta f(v))\mathrm{d}v}\mathrm{d}w \leq \frac{d}{2}\log\Big(\frac{2\boldsymbol{\pi}e(b + d/\beta)}{dm}\Big).$$

Here, $\boldsymbol{\pi}$ denotes the Archimedes' constant. By Lemma 8, we have

$$-\log\int_{\mathbb{R}^d}\exp(-\beta f(w))\mathrm{d}w \leq \beta f(\mathbf{w}^*) + \frac{\beta^{-\gamma}Mc_\alpha^{-1}}{1+\gamma} - \log\Big(\frac{\boldsymbol{\pi}^{d/2}\beta^{-d}}{\Gamma(d/2+1)}\Big).$$

Then, it implies that

$$\begin{aligned}\mathbb{E}[f(\hat{\mathbf{w}})] \leq& \frac{d\beta^{-1}}{2}\log\Big(\frac{2\boldsymbol{\pi}e(b + d/\beta)}{dm}\Big) + f(\mathbf{w}^*) + \frac{\beta^{-\gamma-1}Mc_\alpha^{-1}}{1+\gamma} - \beta^{-1}\log\Big(\frac{\boldsymbol{\pi}^{d/2}\beta^{-d}}{\Gamma(d/2+1)}\Big) \\ =& f(\mathbf{w}^*) + \frac{\beta^{-\gamma-1}Mc_\alpha^{-1}}{1+\gamma} + \beta^{-1}\log\Big(\frac{(2e(b + d/\beta))^{d/2}\Gamma(d/2+1)\beta^d}{(dm)^{d/2}}\Big),\end{aligned}$$

which leads to desired result. □

Proof of Corollary 6

Proof. By triangular inequality, we have

$$\mathcal{W}_q(\mu_{2t},\pi) \le \mathcal{W}_q(\mu_{2t},\mu_{1t}) + \mathcal{W}_q(\mu_{1t},\mu_{3t}) + \mathcal{W}_q(\mu_{3t},\pi).$$

Then, using Corollary 1, Corollary 3 and Proposition 1, we get

$$\begin{aligned}
&\mathcal{W}_q(\mu_{2t},\pi) \le \\
&\le C\Big((k^2\eta + k^2\eta^{1+\gamma/\alpha}\beta^{-\gamma(q-1)/\alpha}d)^{1/q} + (k^{q+\gamma}\eta + k^{q+\gamma}\eta^q\beta^{-(q-1)/\alpha}d)^{1/q} + \beta e^{-\lambda_* k\eta/\beta}\Big) \\
&\le C\Big(k^{2/q}\eta^{1/q} + k^{2/q}\eta^{1/q+\gamma/(q\alpha)}\beta^{-\gamma(q-1)/(q\alpha)}d^{1/q} + k^{1+\gamma/q}\eta^{1/q} + k^{1+\gamma/q}\eta\beta^{-(q-1)/(q\alpha)}d^{1/q} \\
&\quad + \beta e^{-\lambda_* k\eta/\beta}\Big),
\end{aligned}$$

where, we used Lemma 16 for the second inequality. Then, similar to the proof of Corollary 1, we obtain

$$\begin{aligned}
&\mathcal{W}_q(\mu_{2t},\pi) \le \\
&\qquad \le C\Big(k^{\max\{2,q+\gamma\}/q}\eta^{1/q} + k^{\max\{2,q+\gamma\}/q}\eta^{1/q+\gamma/(q\alpha)}\beta^{-\gamma(q-1)/(q\alpha)}d^{1/q} + \beta e^{-\lambda_* k\eta/\beta}\Big).
\end{aligned}$$

□

Proof of Theorem 10

Proof. Since each function $x \mapsto f^{(i)}(x)$ satisfies assumptions **A**1-7, it is easy to check that f_k also satisfies these assumptions (with the same constants and the same parameters) for all k. Then by repeating exactly the same lines as in the proof of Lemma 12, we obtain the same estimates for the moments of $\mathbf{x}_2$. Now by following the same steps as in the proof of Theorem 13, we first have

$$\begin{aligned}
&\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \le \\
&\le q\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\Big[\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big]^{\frac{1}{p_1}}\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s)) - \nabla f_k(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}}\mathrm{d}s,
\end{aligned}$$

then

$$\begin{aligned}
&\Big(\mathbb{E}\|\mathbf{x}_1(s)-\mathbf{x}_2(s)\|^{(q-1)p_1}\Big)^{\frac{1}{p_1}} \le \\
&\quad \le c^{\frac{1}{p_1}} + (\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1})^{\frac{1}{p_1}} + \eta^{q-1}\Big(M^{q-1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{(q-1)p_1\gamma})^{\frac{1}{p_1}} + B^{(q-1)}\Big) \\
&\qquad + \Big(\frac{\eta}{\beta}\Big)^{\frac{q-1}{\alpha}} l_{\alpha,(q-1)p_1,d}^{\frac{1}{p_1}} + (j+1)^{\frac{1}{p_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{p_1}} + (c\eta)^{\frac{1}{p_1}} + (2\eta(b+m))^{\frac{(q-1)}{2}} \\
&\qquad + 2^{\frac{(q-1)}{2}}(\eta B)^{(q-1)} + \Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)}{\alpha}} l_{\alpha,(q-1)p_1,d}^{\frac{1}{p_1}} + \eta^{q-1}M^{q-1}\Big((2\eta(b+m))^{\frac{(q-1)\gamma}{2}} \\
&\qquad + 2^{\frac{(q-1)\gamma}{2}}(\eta B)^{(q-1)\gamma} + \Big(\frac{\eta}{\beta}\Big)^{\frac{(q-1)\gamma}{\alpha}} l_{\alpha,(q-1)p_1\gamma,d}^{\frac{1}{p_1}}\Big)\Big) \\
&\quad = Q_1(\eta) + (j+1)^{\frac{1}{p_1}}P_1(\eta),
\end{aligned}$$

where $P_1(\eta)$ and $Q_1(\eta)$ are defined in Theorem 13. Now, by Minkowski's inequality, we have

$$\begin{aligned}\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s))-\nabla f_k(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}} =& \\ =&\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s)) - \nabla f(\mathbf{x}_2(j\eta)) + \nabla f(\mathbf{x}_2(j\eta)) \\ &- \nabla f_k(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}} \\ \leq&\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_1(s)) - \nabla f(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}} + \Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_2(j\eta)) \\ &- \nabla f_k(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}}.\end{aligned}$$

As in the proof of Theorem 13, the following inequality holds:

$$\begin{aligned}\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))-c_\alpha\nabla f(\mathbf{x}_2(j\eta))\|^{q_1}\Big]^{\frac{1}{q_1}} \leq& \\ \leq& M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + Mc^{\frac{1}{q_1}} + M(j+1)^{\frac{1}{q_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{q_1}} \\ &+ (c\eta)^{\frac{1}{q_1}} + (2\eta(b+m))^{\frac{\gamma}{2}} + 2^{\frac{\gamma}{2}}(\eta B)^\gamma + \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big) \\ =& Q_2 + (j+1)^{\frac{1}{q_1}} P_2(\eta),\end{aligned}$$

where $P_2(\eta)$ and Q_2 are defined in theorem 13. Using the additional assumption, lemma 12, and lemma 16, we get

$$\begin{aligned}\Big[\mathbb{E}\|c_\alpha(\nabla f(\mathbf{x}_2(j\eta))-\nabla f_k(\mathbf{x}_2(j\eta)))\|^{q_1}\Big]^{\frac{1}{q_1}} \leq& \\ \leq& \delta\Big[\mathbb{E}\Big(M^{q_1}\|\mathbf{x}_2(j\eta)\|^{\gamma q_1}\Big)\Big]^{\frac{1}{q_1}} \\ \leq& \delta\Big[M^{q_1}(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1}) + M^{q_1} j\Big((2\eta(b+m))^{\frac{\gamma q_1}{2}} + 2^{\frac{\gamma q_1}{2}}(\eta B)^{\gamma q_1} \\ &+ \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma q_1}{\alpha}} l_{\alpha,\gamma q_1,d}\Big)\Big]^{\frac{1}{q_1}} \\ \leq& \delta M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + \delta M(j+1)^{\frac{1}{q_1}}\Big((2\eta(b+m))^{\frac{\gamma}{2}} + 2^{\frac{\gamma}{2}}(\eta B)^\gamma \\ &+ \Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big).\end{aligned}$$

By combining the two above inequalities, we obtain

$$\begin{aligned}\Big[\mathbb{E}\|c_\alpha\nabla f(\mathbf{x}_1(s))-c_\alpha\nabla f(\mathbf{x}_2(j\eta))\|^{q_1}\Big]^{\frac{1}{q_1}} \leq& \\ \leq& (1+\delta)M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma q_1})^{\frac{1}{q_1}} + Mc^{\frac{1}{q_1}} + M(j+1)^{\frac{1}{q_1}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{q_1}} \\ &+ (c\eta)^{\frac{1}{q_1}} + (1+\delta)(2\eta(b+m))^{\frac{\gamma}{2}} + (1+\delta)2^{\frac{\gamma}{2}}(\eta B)^\gamma \\ &+ (1+\delta)\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}} l^{\frac{1}{q_1}}_{\alpha,\gamma q_1,d}\Big) \\ =& Q_2' + (j+1)^{\frac{1}{q_1}} P_2'(\eta).\end{aligned}$$

Finally, we have

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq q\eta\Big(k^2P_1(\eta)P_2'(\eta)+k^{1+1/p_1}P_1(\eta)Q_2'+k^{1+1/q_1}P_2'(\eta)Q_1(\eta)+kQ_1(\eta)Q_2'\Big).$$

By considering the additional term δ, we arrive at the following bound:

$$\mathcal{W}_q^q(\mu_{1t},\mu_{2t}) \leq C(1+\delta)(k^2\eta+k^2\eta^{1+\gamma/\alpha}\beta^{-\gamma(q-1)/\alpha}d).$$

□

Proof of corollary 7

Proof. By lemma 2,

$$\begin{aligned}c_\alpha\big|\mathbb{E}[f(\mathbf{x}_1(k\eta))]-\mathbb{E}[f(\mathbf{x}_2(k\eta))]\big| \leq \\ \leq\Big(M\Big(\mathbb{E}\|\mathbf{x}_1(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+M\Big(\mathbb{E}\|\mathbf{x}_2(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+B\Big)\mathcal{W}_q(\mu_{1t},\mu_{2t}).\end{aligned}$$

Then, by following the same proof as in corollary 8, corollary 1 and using theorem 10, we get

$$c_\alpha\big|\mathbb{E}[f(\mathbf{x}_1(k\eta))]-\mathbb{E}[f(\mathbf{x}_2(k\eta))]\big| \leq C(1+\delta)\Big(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}}+k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d\Big).$$

□

Technical results

Corollary 8. *Along with $P_1(\eta),P_2(\eta),Q_1(\eta),Q_2$ in Lemma 13, we define, in addition, the following quantities:*

$$\begin{aligned}&P_3(\eta)\triangleq M\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{p}}+(c\eta)^{\frac{1}{p}}+(2\eta(b+m))^{\frac{\gamma}{2}}+2^{\frac{\gamma}{2}}(\eta B)^\gamma+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}}l^{\frac{1}{p}}_{\alpha,\gamma p,d}\Big)\\ &Q_3\triangleq M(\mathbb{E}\|X_2(0)\|^{\gamma p})^{\frac{1}{p}}+Mc^{\frac{1}{p}}+B.\end{aligned}$$

For $0<\eta<\frac{m}{M^2}$, we have the following bound:

$$\begin{aligned}&c_\alpha\big|\mathbb{E}[f(\mathbf{x}_1(k\eta))]-\mathbb{E}[f(\mathbf{x}_2(k\eta))]\big| \leq\\ &\leq(q\eta)^{\frac{1}{q}}\Big(k^{1+\frac{1}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}}P_3(\eta)+k^{1+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}}P_3(\eta)+k^{1+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}}P_3(\eta)\\ &\quad+k(Q_1(\eta)Q_2)^{\frac{1}{q}}P_3(\eta)+k^{\frac{2}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}}Q_3+k^{\frac{1}{q}+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}}Q_3\\ &\quad+k^{\frac{1}{q}+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}}Q_3+k^{\frac{1}{q}}(Q_1(\eta)Q_2)^{\frac{1}{q}}Q_3\Big).\end{aligned}$$

Proof. By Lemma 2,

$$c_\alpha\big|\mathbb{E}[f(\mathbf{x}_1(k\eta))]-\mathbb{E}[f(\mathbf{x}_2(k\eta))]\big| \leq$$
$$\leq\Big(M\Big(\mathbb{E}\|\mathbf{x}_1(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+M\Big(\mathbb{E}\|\mathbf{x}_2(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+B\Big)\mathcal{W}_q(\mu_{1t},\mu_{2t}).$$

Using Lemma 9 and Lemma 13, we have

$$\Big(M\Big(\mathbb{E}\|\mathbf{x}_1(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+M\Big(\mathbb{E}\|\mathbf{x}_2(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+B\Big)\leq$$
$$\leq M\Big(c\Big(k\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+M\Big[(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma p})$$
$$+k\Big((2\eta(b+m))^{\frac{\gamma p}{2}}+2^{\frac{\gamma p}{2}}(\eta B)^{\gamma p}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma p}{\alpha}}l_{\alpha,\gamma p,d}\Big)\Big]^{\frac{1}{p}}+B.$$

By using Lemma 16, we obtain

$$\Big(M\Big(\mathbb{E}\|\mathbf{x}_1(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+M\Big(\mathbb{E}\|\mathbf{x}_2(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+B\Big)\leq$$
$$\leq M\Big(c\Big(k\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}+M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma p})^{\frac{1}{p}}$$
$$+Mk^{\frac{1}{p}}\Big((2\eta(b+m))^{\frac{\gamma}{2}}+2^{\frac{\gamma}{2}}(\eta B)^{\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}}l^{\frac{1}{p}}_{\alpha,\gamma p,d}\Big)+B.$$

We note that $\gamma<\frac{1}{p}$. Hence,

$$\Big(c\Big(k\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\gamma}\leq\Big(c\Big(k\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)+1\Big)\Big)^{\frac{1}{p}}$$
$$\leq k^{\frac{1}{p}}\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{p}}+c^{\frac{1}{p}},$$

where the last inequality is an application of Lemma 16. By replacing this inequality into the previous one and rearranging the terms, we have

$$\Big(M\Big(\mathbb{E}\|\mathbf{x}_1(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+M\Big(\mathbb{E}\|\mathbf{x}_2(k\eta)\|^{\gamma p}\Big)^{\frac{1}{p}}+B\Big)\leq$$
$$\leq M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma p})^{\frac{1}{p}}+Mc^{\frac{1}{p}}+B+Mk^{\frac{1}{p}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}+1\Big)\Big)^{\frac{1}{p}}$$
$$+(2\eta(b+m))^{\frac{\gamma}{2}}+2^{\frac{\gamma}{2}}(\eta B)^{\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}}l^{\frac{1}{p}}_{\alpha,\gamma p,d}\Big)$$
$$\leq M(\mathbb{E}\|\mathbf{x}_2(0)\|^{\gamma p})^{\frac{1}{p}}+Mc^{\frac{1}{p}}+B+Mk^{\frac{1}{p}}\Big(\Big(c\eta\Big(\frac{d}{\beta^{1/\alpha}}\Big)\Big)^{\frac{1}{p}}$$
$$+(c\eta)^{\frac{1}{p}}+(2\eta(b+m))^{\frac{\gamma}{2}}+2^{\frac{\gamma}{2}}(\eta B)^{\gamma}+\Big(\frac{\eta}{\beta}\Big)^{\frac{\gamma}{\alpha}}l^{\frac{1}{p}}_{\alpha,\gamma p,d}\Big)$$
$$=Q_3+k^{\frac{1}{p}}P_3(\eta).$$

Here, we have used lemma 16 in the last inequality. Next, by Lemma 13 and Lemma 16,

$$\mathcal{W}_q(\mu_{1t},\mu_{2t})\leq(q\eta)^{\frac{1}{q}}\Big(k^2P_1(\eta)P_2(\eta)+k^{1+1/p_1}P_1(\eta)Q_2+k^{1+1/q_1}P_2(\eta)Q_1(\eta)+kQ_1(\eta)Q_2\Big)^{\frac{1}{q}}$$
$$\leq(q\eta)^{\frac{1}{q}}\Big(k^{\frac{2}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}}+k^{\frac{1}{q}+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}}+k^{\frac{1}{q}+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}}+$$
$$+k^{\frac{1}{q}}(Q_1(\eta)Q_2)^{\frac{1}{q}}\Big).$$

By combining the above two inequalities, we get

$$
\begin{aligned}
&c_\alpha\big|\mathbb{E}[f(\mathbf{x}_1(k\eta))] - \mathbb{E}[f(\mathbf{x}_2(k\eta))]\big| \leq \\
&\leq (q\eta)^{\frac{1}{q}}\Big(Q_3 + k^{\frac{1}{p}}P_3(\eta)\Big)\Big(k^{\frac{2}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}} + k^{\frac{1}{q}+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}} + k^{\frac{1}{q}+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}} + \\
&\quad + k^{\frac{1}{q}}(Q_1(\eta)Q_2)^{\frac{1}{q}}\Big) \\
&= (q\eta)^{\frac{1}{q}}\Big(k^{1+\frac{1}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}}P_3(\eta) + k^{1+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}}P_3(\eta) + k^{1+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}}P_3(\eta) \\
&\quad + k(Q_1(\eta)Q_2)^{\frac{1}{q}}P_3(\eta) + k^{\frac{2}{q}}(P_1(\eta)P_2(\eta))^{\frac{1}{q}}Q_3 + k^{\frac{1}{q}+\frac{1}{qp_1}}(P_1(\eta)Q_2)^{\frac{1}{q}}Q_3 + \\
&\quad + k^{\frac{1}{q}+\frac{1}{qq_1}}(P_2(\eta)Q_1(\eta))^{\frac{1}{q}}Q_3 + k^{\frac{1}{q}}(Q_1(\eta)Q_2)^{\frac{1}{q}}Q_3\Big).
\end{aligned}
$$

□

The following lemma is an extension of Lemma 1.2.3 in Nesterov [2013] to functions with Hölder continuous gradients.

Lemma 7. *Under assumption **A**2, the following inequality holds for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$*:*

$$
c_\alpha|f(\mathbf{x}) - f(\mathbf{y}) - \langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| \leq \frac{M}{1+\gamma}\|\mathbf{x} - \mathbf{y}\|^{1+\gamma}.
$$

Proof. Let $g(t) \triangleq c_\alpha f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$. Then, $g'(t) = c_\alpha\langle\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y}\rangle$ and $\int_0^1 g'(t)\mathrm{d}t = g(1) - g(0) = c_\alpha(f(\mathbf{x}) - f(\mathbf{y}))$. We have

$$
\begin{aligned}
c_\alpha|f(\mathbf{x}) - f(\mathbf{y}) - \langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| &= \Big|\int_0^1 g'(t)\mathrm{d}t - c_\alpha\langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\Big| \\
&= \Big|\int_0^1 c_\alpha\langle\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y}\rangle\mathrm{d}t - c_\alpha\langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\Big| \\
&= \Big|\int_0^1 c_\alpha\langle\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\mathrm{d}t\Big|.
\end{aligned}
$$

By Cauchy-Schwarz inequality and assumption **A**2, we have

$$
\begin{aligned}
c_\alpha|f(\mathbf{x}) - f(\mathbf{y}) - \langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| &\leq \int_0^1 c_\alpha\|\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y})\|\|\mathbf{x} - \mathbf{y}\|\mathrm{d}t \\
&\leq \int_0^1 Mt^\gamma\|\mathbf{x} - \mathbf{y}\|^\gamma\|\mathbf{x} - \mathbf{y}\|\mathrm{d}t \\
&= \frac{M}{1+\gamma}\|\mathbf{x} - \mathbf{y}\|^{1+\gamma}.
\end{aligned}
$$

□

Lemma 8. *The normalized factor of* π *is bounded below, i.e.,*

$$
\log\int_{\mathbb{R}^d}\exp(-\beta f(w))\mathrm{d}w \geq -\beta f(\mathbf{w}^*) - \frac{\beta^{-\gamma}Mc_\alpha^{-1}}{1+\gamma} + \log\Big(\frac{\boldsymbol{\pi}^{d/2}\beta^{-d}}{\Gamma(d/2+1)}\Big),
$$

where $\boldsymbol{\pi}$ *denotes the Archimedes' constant.*

Proof. We start by writing:

$$\log \int_{\mathbb{R}^d} \exp(-\beta f(w))\mathrm{d}w = -\beta f(\mathbf{w}^*) + \log \int_{\mathbb{R}^d} \exp\big(-\beta(f(w) - f(\mathbf{w}^*))\big)\mathrm{d}w$$
$$\geq -\beta f(\mathbf{w}^*) + \log \int_{\mathbb{R}^d} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\|w - \mathbf{w}^*\|^{1+\gamma}\Big)\mathrm{d}w.$$

Here, we used Lemma 7, with $\nabla f(\mathbf{w}^*) = 0$. For the second term on the right hand side, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\|w - \mathbf{w}^*\|^{1+\gamma}\Big)\mathrm{d}w &= \int_{\|w\|\leq\beta^{-1}} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\|w\|^{1+\gamma}\Big)\mathrm{d}w \\
&\quad + \int_{\|w\|\geq\beta^{-1}} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\|w\|^{1+\gamma}\Big)\mathrm{d}w \\
&\geq \int_{\|w\|\leq\beta^{-1}} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\beta^{-1-\gamma}\Big)\mathrm{d}w + 0 \\
&= \exp\Big(-\frac{\beta^{-\gamma} M c_\alpha^{-1}}{1+\gamma}\Big)\int_{\|w\|\leq\beta^{-1}} 1\mathrm{d}w \\
&= \exp\Big(-\frac{\beta^{-\gamma} M c_\alpha^{-1}}{1+\gamma}\Big)\frac{\boldsymbol{\pi}^{d/2}\beta^{-d}}{\Gamma(d/2+1)},
\end{aligned}$$

where, Γ denotes the Gamma function and $\boldsymbol{\pi}$ denotes Archimedes' constant (here, it is not the invariant distribution). Hence,

$$\log \int_{\mathbb{R}^d} \exp\Big(-\frac{\beta M c_\alpha^{-1}}{1+\gamma}\|w - \mathbf{w}^*\|^{1+\gamma}\Big)\mathrm{d}w \geq -\frac{\beta^{-\gamma} M c_\alpha^{-1}}{1+\gamma} + \log\Big(\frac{\boldsymbol{\pi}^{d/2}\beta^{-d}}{\Gamma(d/2+1)}\Big).$$

By combining the above inequalities, we have the desired result. □

Lemma 9. *For $\lambda \in (0,1)$, there exists a constant c depending on m, b, α, such that*

$$\mathbb{E}\Big(\|\mathbf{x}_1(t)\|^\lambda\Big)^{\frac{1}{\lambda}} \leq c\Big(t(d\beta^{-1/\alpha}+1)+1\Big), \qquad \forall t>0, \beta \geq 1, 1<\alpha<2.$$

Proof. We follow exactly the same proof as Lemma 7.1 in Xie and Zhang [2017], with some modifications. Let $h(x) \triangleq (1+\|x\|^2)^{1/2}$. By Itô's formula, we have

$$\begin{aligned}
\mathrm{d}h(\mathbf{x}_1(t)) = &\Bigg(\langle b_1(\mathbf{x}_1(t)), \nabla h(\mathbf{x}_1(t))\rangle + \int_{\mathbb{R}^d}\Big(h(\mathbf{x}_1(t)+\beta^{-1/\alpha}x) - h(\mathbf{x}_1(t)) \\
&- \mathbb{I}_{\|x\|<1}\langle\beta^{-1/\alpha}x, \nabla h(\mathbf{x}_1(t))\rangle\Big)\nu(\mathrm{d}x)\Bigg)\mathrm{d}t + \mathrm{d}M(t), \qquad (7.11)
\end{aligned}$$

where $M(t)$ is a local martingale. Noticing that $\partial_i h(x) = x_i(1+\|x\|^2)^{-1/2}/2$ and using assumption **A**3, we have

$$\begin{aligned}\langle b_1(x), \nabla h(x)\rangle =& \langle b_1(x), x\rangle (1+\|x\|^2)^{-1/2}/2 \\ \leq& (-m\|x\|^{1+\gamma} + b)(1+\|x\|^2)^{-1/2}/2 \\ =& (-m(\|x\|^{1+\gamma}+1) + m + b)(1+\|x\|^2)^{-1/2}/2.\end{aligned}$$

Since $(\|x\|^2+1)^{(1+\gamma)/2} \leq (\|x\|^{1+\gamma}+1)$ by Lemma 16, it follows that

$$\begin{aligned}\langle b_1(x), \nabla h(x)\rangle \leq& (-m(\|x\|^2+1)^{(1+\gamma)/2} + m + b)(1+\|x\|^2)^{-1/2}/2 \\ =& (-m(\|x\|^2+1)^{\gamma/2} + (m+b)(1+\|x\|^2)^{-1/2})/2 \\ \leq& (-m(\|x\|^2+1)^{\gamma/2} + m + b)/2 \\ =& (-mh(x)^\gamma + m + b)/2.\end{aligned}$$

On the other hand, observing that

$$|h(x+y) - h(x)| \leq \|y\| \int_0^1 \|\nabla h(x+sy)\| \mathrm{d}s \leq \|y\|/2,$$

and

$$h(x+y) - h(x) - \langle y, \nabla h(x)\rangle \leq \|y\|^2/2,$$

we have

$$\begin{aligned}\int_{\mathbb{R}^d} \Big(h(\mathbf{x}_1(t)+x) - h(\mathbf{x}_1(t)) - &\mathbb{I}_{\|x\|<1}\langle x, \nabla h(\mathbf{x}_1(t))\rangle\Big)\nu(\mathrm{d}x) \leq \\ \leq& \frac{1}{2\beta^{2/\alpha}} \int_{\|x\|<1} \|x\|^2 \nu(\mathrm{d}x) + \frac{1}{2\beta^{1/\alpha}} \int_{\|x\|\geq 1} \|x\| \nu(\mathrm{d}x) \\ \leq& C\frac{d}{\beta^{1/\alpha}},\end{aligned}$$

where the last inequality is due to Lemma 15. By integrating (7.11) and combining the above inequalities, we have

$$\begin{aligned}h(\mathbf{x}_1(t)) - h(\mathbf{x}_1(0)) \leq& \int_0^t \Big((-mh(\mathbf{x}_1(s))^\gamma + m + b)/2 + C\frac{d}{\beta^{1/\alpha}}\Big)\mathrm{d}s + M(t) \\ \leq& \int_0^t \Big((m+b)/2 + C\frac{d}{\beta^{1/\alpha}}\Big)\mathrm{d}s + M(t).\end{aligned}$$

By Lemma 3.8 in Xie and Zhang [2017], for $\lambda \in (0,1)$,

$$\mathbb{E}\Big(\sup_{s\in[0,t]} h(\mathbf{x}_1(s))^\lambda\Big) \leq c_\lambda \Big(\mathbb{E}h(\mathbf{x}_1(0)) + ((m+b)/2 + C\frac{d}{\beta^{1/\alpha}})t\Big)^\lambda.$$

This leads to the conclusion since $h(x) \geq \|x\|$. □

Lemma 10. *For $\lambda \in (0,1)$, there exists a constant c_b depending on L, m, b, α, such that*

$$\mathbb{E}\Big(\|\mathbf{x}_3(t)\|^{\lambda}\Big)^{\frac{1}{\lambda}} \leq c_b\Big(t(d\beta^{-1/\alpha}+1)+1\Big), \qquad \forall t>0, \beta \geq 1, 1<\alpha<2.$$

Proof. The proof is similar to the proof of Lemma 9. □

Lemma 11. *Let X be a scalar symmetric α-stable distribution with $\alpha < 2$, i.e. $X \sim \mathcal{S}\alpha\mathcal{S}(1)$ (see Definition 3), then, for $-1 < \lambda < \alpha$,*

$$\mathbb{E}(|X|^{\lambda}) = \frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}.$$

Proof. The proof follows from Theorem 3 in Shanbhag and Sreehari [1977] (see also equation (13) in Matsui et al. [2016]). □

Corollary 9. *The quantity $l_{\alpha,\lambda,d} \triangleq \mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda}$ is finite for $0 \leq \lambda < \alpha$. For details, we have*

(a) If $1 < \lambda < \alpha$, then

$$\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda} \leq d^{\lambda}\Big(\frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big).$$

(b) If $0 \leq \lambda \leq 1$, then

$$\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda} \leq d\Big(\frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big).$$

Proof. Since $\mathrm{L}^{\alpha}(1)$, by definition, is a d-dimensional vector whose components are i.i.d symmetric α-stable distributions $L_i^{\alpha}(1)$ for $i \in \{1, \ldots, d\}$, we have

$$\|\mathrm{L}^{\alpha}(1)\| \leq \sum_{i=1}^{d} |L_i^{\alpha}(1)|$$

(a) $1 < \lambda < \alpha$. By using Minkowski's inequality and Lemma 11,

$$\begin{aligned}(\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda})^{1/\lambda} \leq& \Big(\mathbb{E}\Big[\Big(\sum_{i=1}^{d}|L_i^{\alpha}(1)|\Big)^{\lambda}\Big]\Big)^{1/\lambda}\\ \leq& \sum_{i=1}^{d}(\mathbb{E}|L_i^{\alpha}(1)|^{\lambda})^{1/\lambda}\\ =& d\Big(\frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big)^{1/\lambda}.\end{aligned}$$

Thus, we have

$$\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda \le d^\lambda\Big(\frac{2^\lambda\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big).$$

(b) $0 \le \lambda \le 1$. By using Lemma 16 and Lemma 11 ,

$$\begin{aligned}\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda \le& \mathbb{E}\Big[\Big(\sum_{i=1}^d |L_i^\alpha(1)|\Big)^\lambda\Big]\\ \le& \sum_{i=1}^d \mathbb{E}|L_i^\alpha(1)|^\lambda\\ =& d\Big(\frac{2^\lambda\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big).\end{aligned}$$

□

Lemma 12. *Let us denote the value* $\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda$ *by* $l_{\alpha,\lambda,d} < \infty$. *For* $0 < \eta \le \frac{m}{M^2}$ *and* $s \in [j\eta, (j+1)\eta)$, *we have the following estimates:*

(a) If $1 < \lambda < \alpha$ *and* $1 < \gamma\lambda < \alpha$ *then*

$$\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda \le B_{j,\lambda} \triangleq \Big(\Big(\mathbb{E}\|\mathbf{x}_2(0)\|^\lambda\Big)^{\frac{1}{\lambda}} + j\Big((2\eta(b+m))^{\frac12} + 2^{\frac12}\eta B + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}}\Big)\Big)^\lambda,$$

$$\mathbb{E}\|\mathbf{x}_2(s)\|^\lambda \le \Big(B_{j,\lambda}^{\frac{1}{\lambda}} + (s-j\eta)\Big(MB_{j,\gamma\lambda}^{\frac{1}{\lambda}} + B\Big) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}}\Big)^\lambda.$$

(b) If $0 \le \lambda \le 1$ *then*

$$\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda \le \bar{B}_{j,\lambda} \triangleq \mathbb{E}\|\mathbf{x}_2(0)\|^\lambda + j\Big((2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda + \Big(\frac{\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d}\Big),$$

$$\mathbb{E}\|\mathbf{x}_2(s)\|^\lambda \le \bar{B}_{j,\lambda} + (s-j\eta)^\lambda\Big(M^\lambda \bar{B}_{j,\gamma\lambda} + B^\lambda\Big) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d}.$$

(c) If $1 < \lambda < \alpha$ *and* $0 \le \gamma\lambda \le 1$ *then*

$$\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda \le B_{j,\lambda},$$

$$\mathbb{E}\|\mathbf{x}_2(s)\|^\lambda \le \Big(B_{j,\lambda}^{\frac{1}{\lambda}} + (s-j\eta)\Big(M\bar{B}_{j,\gamma\lambda}^{\frac{1}{\lambda}} + B\Big) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}}\Big)^\lambda.$$

Proof. Starting from

$$\mathbf{x}_2((j+1)\eta) = \mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta)) + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} \mathrm{L}^\alpha(1),$$

we have either (by Minkowski, if $\lambda > 1$)

$$\Big(\mathbb{E}\|\mathbf{x}_2((j+1)\eta)\|^\lambda\Big)^{\frac{1}{\lambda}} \le \Big(\mathbb{E}\|\mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^\lambda\Big)^{\frac{1}{\lambda}} + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} \Big(\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda\Big)^{\frac{1}{\lambda}}, \tag{7.12}$$

or (by Lemma 16, if $0 \le \lambda \le 1$)

$$\mathbb{E}\|\mathbf{x}_2((j+1)\eta)\|^\lambda \le \mathbb{E}\|\mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^\lambda + \Big(\frac{\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} \mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda. \tag{7.13}$$

We have

$$\begin{aligned}
&\|\mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^\lambda = \\
&\qquad = \|\mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^{2\times\frac{\lambda}{2}} &(7.14)\\
&\qquad = \Big(\|\mathbf{x}_2(j\eta)\|^2 - 2\eta c_\alpha \langle \mathbf{x}_2(j\eta), \nabla f(\mathbf{x}_2(j\eta))\rangle + \eta^2 \|c_\alpha \nabla f(\mathbf{x}_2(j\eta)\|^2\Big)^{\frac{\lambda}{2}} \\
&\qquad \le \Big(\|\mathbf{x}_2(j\eta)\|^2 - 2\eta(m\|\mathbf{x}_2(j\eta)\|^{1+\gamma} - b) + \eta^2(2M^2\|\mathbf{x}_2(j\eta)\|^{2\gamma} + 2B^2)\Big)^{\frac{\lambda}{2}}, &(7.15)
\end{aligned}$$

where we used assumption **A**3 and Lemma 13. For $0 < \eta \le \frac{m}{M^2}$,

$$2\eta m(\|X_2(j\eta)\|^{1+\gamma} + 1) \ge 2\eta^2 M^2 \|X_2(j\eta)\|^{2\gamma}. \quad \text{(since } 1+\gamma > 2\gamma \text{ and } \eta m > \eta^2 M^2\text{)}$$

Using this inequality we have

$$\begin{aligned}
\|\mathbf{x}_2(j\eta) - \eta c_\alpha \nabla f(\mathbf{x}_2(j\eta))\|^\lambda &\le \Big(\|\mathbf{x}_2(j\eta)\|^2 + 2\eta(b+m) + 2\eta^2 B^2\Big)^{\frac{\lambda}{2}} \\
&\le \|\mathbf{x}_2(j\eta)\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda. \quad \text{(by Lemma 16)} \qquad (7.16)
\end{aligned}$$

Consider the case where $\lambda > 1$. By (7.12) and (7.16),

$$\begin{aligned}
\Big(\mathbb{E}\|\mathbf{x}_2((j+1)\eta)\|^\lambda\Big)^{\frac{1}{\lambda}} \le& \\
&\le \Big(\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda\Big)^{\frac{1}{\lambda}} + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} \Big(\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda\Big)^{\frac{1}{\lambda}} \\
&\le \Big(\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda\Big)^{\frac{1}{\lambda}} + (2\eta(b+m))^{\frac{1}{2}} + 2^{\frac{1}{2}}\eta B + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}} \quad \text{(by Lemma 16)} \\
&\le \Big(\mathbb{E}\|\mathbf{x}_2(0)\|^\lambda\Big)^{\frac{1}{\lambda}} + (j+1)\Big((2\eta(b+m))^{\frac{1}{2}} + 2^{\frac{1}{2}}\eta B + \Big(\frac{\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}}\Big).
\end{aligned}$$

For the case where $0 \le \lambda \le 1$, by (7.13) and (7.16),

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_2((j+1)\eta)\|^\lambda &\le \mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda + \Big(\frac{\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d} \\
&\le \mathbb{E}\|\mathbf{x}_2(0)\|^\lambda + (j+1)\Big((2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda + \Big(\frac{\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d}\Big).
\end{aligned}$$

Now, from the identification, for $s \in [j\eta, (j+1)\eta)$,

$$\mathbf{x}_2(s) = \mathbf{x}_2(j\eta) + (s - j\eta)c_\alpha \nabla f(\mathbf{x}_2(j\eta)) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} \mathrm{L}^\alpha(1),$$

we have

$$\begin{aligned}
\|\mathbf{x}_2(s)\| &\le \|\mathbf{x}_2(j\eta)\| + (s-j\eta)c_\alpha \|\nabla f(\mathbf{x}_2(j\eta))\| + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} \|\mathrm{L}^\alpha(1)\| \\
&\le \|\mathbf{x}_2(j\eta)\| + (s-j\eta)(M\|\mathbf{x}_2(j\eta)\|^\gamma + B) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} \|\mathrm{L}^\alpha(1)\|.
\end{aligned}$$

For $\lambda > 1$,

$$\Big(\mathbb{E}\|\mathbf{x}_2(s)\|^\lambda\Big)^{\frac{1}{\lambda}} \le \Big(\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda\Big)^{\frac{1}{\lambda}} + (s-j\eta)\Big(M\Big(\mathbb{E}\|\mathbf{x}_2(j\eta)\|^{\gamma\lambda}\Big)^{\frac{1}{\lambda}} + B\Big) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}}.$$

For $\lambda \le 1$,

$$\mathbb{E}\|\mathbf{x}_2(s)\|^\lambda \le \mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda + (s-j\eta)^\lambda\Big(M^\lambda \mathbb{E}\|\mathbf{x}_2(j\eta)\|^{\gamma\lambda} + B^\lambda\Big) + \Big(\frac{s-j\eta}{\beta}\Big)^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d}.$$

By replacing the estimate of $\mathbb{E}\|\mathbf{x}_2(j\eta)\|^\lambda$, we obtain the desired result.

□

Lemma 13. *Under assumptions **A**1 and **A**2 we have*

$$c_\alpha \|\nabla f(w)\| \le M\|w\|^\gamma + B, \qquad \forall w \in \mathbb{R}^d.$$

Proof. By assumption **A**2 we have

$$c_\alpha \|\nabla f(w) - \nabla f(0)\| \le M\|w - 0\|^\gamma.$$

Since $c_\alpha\|\nabla f(0)\| \le B$ by assumption **A**1, the conclusion follows. □

Lemma 14. *For the function b defined in Lemma 1, we have, for $w \in \mathbb{R}^d$,*

$$\|b(w)\| \le M\|w\|^\gamma + (B+L),$$
$$\langle w, b(w)\rangle \le (L-m)\|w\|^{1+\gamma} + (b+L).$$

Proof. From assumption **A**6, it implies that

$$\|b(w)\| \le c_\alpha\|\nabla f(w)\| + L.$$

Then, by Lemma 13,

$$\|b(w)\| \le M\|w\|^\gamma + (B+L).$$

Next, by Cauchy-Schwarz inequality and assumption **A**6, we have

$$\langle w, b(w) + c_\alpha \nabla f(w)\rangle \le \|w\|L.$$

Then, by assumption **A**3,

$$\begin{aligned}\langle w, b(w)\rangle \le& - c_\alpha\langle w, \nabla f(w)\rangle + \|w\|L\\ \le& - m\|w\|^{1+\gamma} + b + \|w\|L\\ \le& - m\|w\|^{1+\gamma} + b + (\|w\|^{1+\gamma} + 1)L\\ =&(L-m)\|w\|^{1+\gamma} + (b+L).\end{aligned}$$

Here, we have used the inequality $\|w\| \le \|w\|^{1+\gamma} + 1$. □

Lemma 15. *Let ν be the Lévy measure of a d-dimensional Lévy process L^α whose components are independent scalar symmetric α-stable Lévy processes $L_1^\alpha, \ldots, L_d^\alpha$. Then there exists a constant $C > 0$ such that the following inequality holds with $\beta \geq 1$ and $2 > \alpha > 1$:*

$$\frac{1}{\beta^{2/\alpha}} \int_{\|x\|<1} \|x\|^2 \nu(\mathrm{d}x) + \frac{1}{\beta^{1/\alpha}} \int_{\|x\|\geq 1} \|x\| \nu(\mathrm{d}x) \leq C \frac{d}{\beta^{1/\alpha}}.$$

Proof. Using Lemma 4.1 in Kallsen and Tankov [2006], we have

$$\begin{aligned}
\int_{\|x\|<1} \|x\|^2 \nu(\mathrm{d}x) &= \sum_{i=1}^{d} \int_{|x_i|<1} |x_i|^2 \frac{1}{|x_i|^{1+\alpha}} \mathrm{d}x_i \\
&= \sum_{i=1}^{d} \frac{2}{2-\alpha} \\
&= \frac{2d}{2-\alpha}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\int_{\|x\|\geq 1} \|x\| \nu(\mathrm{d}x) &= \sum_{i=1}^{d} \int_{|x_i|\geq 1} |x_i| \frac{1}{|x_i|^{1+\alpha}} \mathrm{d}x_i \\
&= \sum_{i=1}^{d} \frac{2}{\alpha-1} \\
&= \frac{2d}{\alpha-1}.
\end{aligned}$$

Combining these two equalities, we have the desired conclusion. □

Lemma 16. *For $a, b \geq 0$ and $0 \leq \gamma \leq 1$, we have the following inequality:*

$$(a+b)^\gamma \leq a^\gamma + b^\gamma.$$

Proof. If $a = b = 0$, the inequality is trivial. Hence, let us assume that $a > b \geq 0$. We have

$$\begin{aligned}
\left(1 + \frac{b}{a}\right)^\gamma &\leq 1 + \gamma \frac{b}{a} && \text{(by Bernoulli's inequality)} \\
&\leq 1 + \frac{b}{a} && \text{(since } 0 \leq \gamma \leq 1 \text{ and } \frac{b}{a} \geq 0\text{)} \\
&\leq 1 + \left(\frac{b}{a}\right)^\gamma. && \text{(since } 0 \leq \gamma \leq 1 \text{ and } 0 \leq \frac{b}{a} \leq 1\text{)}
\end{aligned}$$

By multiplying both sides by $a^\gamma > 0$, we have the conclusion. □

Supplementary materials for Chapter 6

More details on assumption A 13

In this section, we provide the precise expressions of the constants given in Assumption **A**13. For a given $\delta > 0$, $t = K\eta$, and for some $C > 0$, the step-size satisfies the following condition:

$$0 < \eta \leq \min\Big\{1, \frac{m}{M^2}, \Big(\frac{\delta^2}{2K_1t^2}\Big)^{\frac{1}{\gamma^2+2\gamma-1}}, \Big(\frac{\delta^2}{2K_2t^2}\Big)^{\frac{1}{2\gamma}}, \Big(\frac{\delta^2}{2K_3t^2}\Big)^{\frac{\alpha}{2\gamma}}, \Big(\frac{\delta^2}{2K_4t^2}\Big)^{\frac{1}{\gamma}}\Big\},$$

where ε is as in (6.3), the constants m, M, b are defined by **A**10– **A**12 and

$$\begin{aligned}
K_1 &\triangleq \frac{CM^{2+2\gamma}3^\gamma}{\varepsilon^2\sigma^2}\max\Big\{(2(b+m))^{\gamma^2}, 2^{\gamma^2}B^{2\gamma^2}, d\varepsilon^{2\gamma^2}R_1, d\varepsilon^{2\gamma^2}R_2\Big\},\\
K_2 &\triangleq \frac{CM^{2+2\gamma}3^\gamma}{2\varepsilon^2\sigma^2}\Big(\mathbb{E}\|\mathbf{w}_0\|^{2\gamma^2} + B^2/M^2\Big),\\
K_3 &\triangleq \frac{M^2 3^\gamma \varepsilon^{2\gamma-2}d^{2\gamma}}{2\sigma^2}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big),\\
K_4 &\triangleq \frac{M^2 3^\gamma \varepsilon^{2\gamma-2}d^{2\gamma}}{2\sigma^2}\Big(2^\gamma\Gamma(\frac{2\gamma+1}{2})/\sqrt{\pi}\Big),
\end{aligned}$$

with

$$R_1 \triangleq \Big(\frac{2^{2\gamma^2}\Gamma((1+2\gamma^2)/2)\Gamma(1-2\gamma^2/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma^2)}\Big), R_2 \triangleq \Big(2^{\gamma^2}\Gamma\left(\frac{2\gamma^2+1}{2}\right)/\sqrt{\pi}\Big)\Big).$$

Proof of Theorem 11

Proof. Note that $(\mathbf{w}^1, \ldots, \mathbf{w}^K) \in A$ is equivalent to $\bar{\tau}_{0,a}(\varepsilon) > K$. Hence, from Lemma 20, the remaining task is to upper-bound $\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A]$:

$$\begin{aligned}
\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] \leq& \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A \cap B] + \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in B^c]\\
\leq& \mathbb{P}[\tau_{\xi,a}(\varepsilon) > K\eta] + \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in B^c],
\end{aligned}$$

and to lower-bound it:

$$\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] \geq \mathbb{P}[\tau_{-\xi,a}(\varepsilon) > K\eta] - \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in B^c].$$

By Lemma 17, the final result follows. □

Lemma 17. *There exist constants C, C_1 and C_α such that:*

$$\begin{aligned}
\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in& B^c] \leq\\
&\leq \frac{C_1(K\eta(d\varepsilon+1)+1)^\gamma e^{M\eta}M\eta}{\xi} + 1 - \Big(1 - Cde^{-\xi^2 e^{-2M\eta}(\varepsilon\sigma)^{-2}/(16d\eta)}\Big)^K\\
&\quad + 1 - \Big(1 - C_\alpha d^{1+\alpha/2}\eta e^{\alpha M\eta}\varepsilon^\alpha\xi^{-\alpha}\Big)^K,
\end{aligned}$$

Proof. We have for $t \in [k\eta, (k+1)\eta]$,

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}_{k\eta}\| &\leq \int_{k\eta}^{t} \|\nabla f(\mathbf{w}_s)\| \mathrm{d}s + \varepsilon\sigma\|\mathrm{B}(t) - \mathrm{B}(k\eta)\| + \varepsilon\|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\| \\
&\leq \int_{k\eta}^{t} \|\nabla f(\mathbf{w}_s) - \nabla f(\mathbf{w}_{k\eta})\| \mathrm{d}s + \eta\|\nabla f(\mathbf{w}_{k\eta})\| + \varepsilon\sigma\|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \\
&\quad + \varepsilon\|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\| \\
&\leq \int_{k\eta}^{t} M\|\mathbf{w}_s - \mathbf{w}_{k\eta}\|^\gamma \mathrm{d}s + \eta(M\|\mathbf{w}_{k\eta}\|^\gamma + B) + \varepsilon\sigma\|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \\
&\quad + \varepsilon\|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\|.
\end{aligned}$$

For $\gamma < 1$, using that $\|\mathbf{w}_s - \mathbf{w}_{k\eta}\|^\gamma \leq \|\mathbf{w}_s - \mathbf{w}_{k\eta}\| + 1$, we get:

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}_{k\eta}\| \leq \int_{k\eta}^{t} M\|\mathbf{w}_s - \mathbf{w}_{k\eta}\| \mathrm{d}s + \eta(M\|\mathbf{w}_{k\eta}\|^\gamma + B + M) \\
+ \varepsilon\sigma \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\| + \varepsilon \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\|.
\end{aligned}$$

Then the Gronwall Lemma gives:

$$\begin{aligned}
\sup_{t\in[k\eta,(k+1)\eta]} \|\mathbf{w}_t - \mathbf{w}_{k\eta}\| \leq & e^{M\eta}\Big[\eta(M\|\mathbf{w}_{k\eta}\|^\gamma + B + M) + \varepsilon\sigma \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \\
& + \varepsilon \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\|\Big].
\end{aligned}$$

Hence,

$$\begin{aligned}
\max_{0\leq k\leq K-1} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathbf{w}_t - \mathbf{w}_{k\eta}\| \leq & e^{M\eta}\Big[\eta(M \max_{0\leq k\leq K-1} \|\mathbf{w}_{k\eta}\|^\gamma + B + M) \\
& + \varepsilon\sigma \max_{0\leq k\leq K} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \\
& + \varepsilon \max_{0\leq k\leq K-1} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\|\Big].
\end{aligned}$$

By Lemma 7.1 in Xie and Zhang [2017], Lemma 9 and Markov's inequality, for any $u > 0$, we have:

$$\mathbb{P}[\max_{0\leq k\leq K-1} \|\mathbf{w}_{k\eta}\|^\gamma \geq u] \leq \frac{\mathbb{E}[\max_{0\leq k\leq K-1} \|\mathbf{w}_{k\eta}\|^\gamma]}{u} \leq \frac{C_1(K\eta(d\varepsilon+1)+1)^\gamma}{u},$$

where C_1 is a constant independent of K, η, ε and d. By Lemma 19, we have:

$$\mathbb{P}[\max_{k\in[0,\ldots,K-1]} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \geq u] \leq 1 - \Big(1 - Cde^{-u^2/(d\eta)}\Big)^K$$

and

$$\mathbb{P}[\max_{k\in[0,\ldots,K-1]} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\| \geq u] \leq 1 - \Big(1 - C_\alpha d^{1+\alpha/2}\eta u^{-\alpha}\Big)^K.$$

Finally, we get:

$$\begin{aligned}
\mathbb{P}[(\mathbf{w}_\eta,\dots,\mathbf{w}_{K\eta}) \in & B^c] \leq \\
&\leq \mathbb{P}[\max_{0\leq k\leq K-1} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathbf{w}_t - \mathbf{w}_{k\eta}\| > \xi] \\
&\leq \mathbb{P}[e^{M\eta}\eta M \max_{0\leq k\leq K-1} \|\mathbf{w}_{k\eta}\|^\gamma \geq \xi/4] + \mathbb{P}[e^{M\eta}\eta(B+M) \geq \xi/4] \\
&\quad + \mathbb{P}[e^{M\eta} \max_{k\in[0,\dots,K-1]} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\| \geq (\varepsilon\sigma)^{-1}\xi/4] \\
&\quad + \mathbb{P}[e^{M\eta} \max_{k\in[0,\dots,K-1]} \sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\| \geq \varepsilon^{-1}\xi/4] \\
&\leq \frac{C_1(K\eta(d\varepsilon+1)+1)^\gamma e^{M\eta} M\eta}{\xi} + 1 - \left(1 - Cde^{-\xi^2 e^{-2M\eta}(\varepsilon\sigma)^{-2}/(16d\eta)}\right)^K \\
&\quad + 1 - \left(1 - C_\alpha d^{1+\alpha/2}\eta e^{\alpha M\eta}\varepsilon^\alpha \xi^{-\alpha}\right)^K.
\end{aligned}$$

□

Now we prove the following lemma.

Lemma 18. *For any $u > 0$, $\eta > 0$ and $K \in \mathbb{N}^*$, there exist constants C and C_α such that:*

$$\max_{k\in[0,\dots,K-1]} \mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]} \|B(t) - B(k\eta)\| \geq u] \leq Cde^{-u^2/(d\eta)}.$$

$$\max_{k\in[0,\dots,K-1]} \mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)\| \geq u] \leq C_\alpha d^{1+\alpha/2}\eta u^{-\alpha}.$$

Proof. To prove the results, we begin with the known results for Brownian motion and α-stable Lévy motion:

$$\begin{aligned}
\mathbb{P}[|[\mathrm{B}(1)]_i| \geq u] &\leq Ce^{-u^2}, \\
\mathbb{P}[|[\mathrm{L}^\alpha(1)]_i| \geq u] &\leq C_\alpha u^{-\alpha},
\end{aligned}$$

where C and C_α are positive constants, $[\mathrm{B}(1)]_i$ and $[\mathrm{L}^\alpha(1)]_i$ denote the i-th component of the motions respectively, for i from 1 to d. By reflection principle for Brownian motion and α-stable Lévy motion, we have

$$\begin{aligned}
\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]} |[\mathrm{B}(t) - \mathrm{B}(k\eta)]_i| \geq u] &\leq 2\mathbb{P}[|[\mathrm{B}(\eta)]_i| \geq u] = 2\mathbb{P}[|[\mathrm{B}(1)]_i| \geq u/\eta^{1/2}], \\
\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]} |[\mathrm{L}^\alpha(t) - \mathrm{L}^\alpha(k\eta)]_i| \geq u] &\leq 2\mathbb{P}[|[\mathrm{L}^\alpha(\eta)]_i| \geq u] = 2\mathbb{P}[|[\mathrm{L}^\alpha(1)]_i| \geq u/\eta^{1/\alpha}].
\end{aligned}$$

Since $\sup_{t\in[k\eta,(k+1)\eta]} \|\mathrm{B}(t) - \mathrm{B}(k\eta)\|^2 \leq \sum_{i=1}^d \sup_{t\in[k\eta,(k+1)\eta]} |[\mathrm{B}(t) - \mathrm{B}(k\eta)]_i|^2$, we

have

$$\begin{aligned}\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|\geq u]=&\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|^2\geq u^2]\\ \leq&\sum_{i=1}^{d}\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}|[\mathrm{B}(t)-\mathrm{B}(k\eta)]_i|^2\geq u^2/d]\\ \leq&\sum_{i=1}^{d}\mathbb{P}[|[\mathrm{B}(1)]_i|\geq u/(d\eta)^{1/2}]\\ \leq&Cde^{-u^2/(d\eta)}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{L}^\alpha(t)-\mathrm{L}^\alpha(k\eta)\|\geq u]\leq&\sum_{i=1}^{d}\mathbb{P}[|[\mathrm{L}^\alpha(1)]_i|\geq u/(d^{1/2}\eta^{1/\alpha})]\\ \leq&C_\alpha d^{1+\alpha/2}\eta u^{-\alpha}.\end{aligned}$$

The constants C and C_α do not depend on k, hence we have the conclusion. □

Lemma 19. *The following estimates hold:*

$$\mathbb{P}[\max_{k\in[0,\dots,K-1]}\sup_{t\in[k\eta,(k+1)\eta]}\|B(t)-B(k\eta)\|\geq u]\leq 1-\left(1-Cde^{-u^2/(d\eta)}\right)^K,$$

$$\mathbb{P}[\max_{k\in[0,\dots,K-1]}\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{L}^\alpha(t)-\mathrm{L}^\alpha(k\eta)\|\geq u]\leq 1-\left(1-C_\alpha d^{1+\alpha/2}\eta u^{-\alpha}\right)^K.$$

Proof. We have

$$\begin{aligned}\mathbb{P}[\max_{k\in[0,\dots,K-1]}\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|\geq u]=&\\ =&1-\mathbb{P}[\max_{k\in[0,\dots,K-1]}\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|<u]\\ =&1-\prod_{k=0}^{K-1}\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|<u]\\ =&1-\prod_{k=0}^{K-1}\left(1-\mathbb{P}[\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{B}(t)-\mathrm{B}(k\eta)\|\geq u]\right)\\ \leq&1-\prod_{k=0}^{K-1}\left(1-Cde^{-u^2/(d\eta)}\right)\\ =&1-\left(1-Cde^{-u^2/(d\eta)}\right)^K.\end{aligned}$$

Similarly, we have

$$\mathbb{P}[\max_{k\in[0,\dots,K-1]}\sup_{t\in[k\eta,(k+1)\eta]}\|\mathrm{L}^\alpha(t)-\mathrm{L}^\alpha(k\eta)\|\geq u]\leq 1-\left(1-C_\alpha d^{1+\alpha/2}\eta u^{-\alpha}\right)^K.$$

□

Lemma 20. *Suppose that assumptions **A**10 and **A**11 hold. Then, for any $\delta > 0$, we have:*

$$\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] - \delta \leq \mathbb{P}[(\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta}) \in A] \leq \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] + \delta,$$

provided that

$$0 < \eta \leq \min\Big\{1, \frac{m}{M^2}, \Big(\frac{\delta^2}{2K_1t^2}\Big)^{\frac{1}{\gamma^2+2\gamma-1}}, \Big(\frac{\delta^2}{2K_2t^2}\Big)^{\frac{1}{2\gamma}}, \Big(\frac{\delta^2}{2K_3t^2}\Big)^{\frac{\alpha}{2\gamma}}, \Big(\frac{\delta^2}{2K_4t^2}\Big)^{\frac{1}{\gamma}}\Big\},$$

Proof. By optimal coupling between two probability measure (Lindvall [2002], Theorem 5.2), there exists a coupling $\mathbf{M}$ of $(\mathbf{w}_s)_{0\leq s\leq K\eta}$ and $(\hat{\mathbf{w}}_s)_{0\leq s\leq K\eta}$ such that

$$\mathbb{P}_{\mathbf{M}}[(\mathbf{w}_s)_{0\leq s\leq K\eta} \neq (\hat{\mathbf{w}}_s)_{0\leq s\leq K\eta}] = \|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV},$$

where TV denotes the total variation distance. By Pinsker's inequality, we also have

$$\|\mu_{K\eta} - \hat{\mu}_{K\eta}\|_{TV}^2 \leq \frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta}).$$

Then,

$$\begin{aligned}\mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \neq (\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta})] \leq& \mathbb{P}_{\mathbf{M}}[(\mathbf{w}_s)_{0\leq s\leq K\eta} \neq (\hat{\mathbf{w}}_s)_{0\leq s\leq K\eta}] \\ \leq& \Big(\frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta})\Big)^{1/2}.\end{aligned}$$

From the following inequalities

$$\begin{aligned}&\mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] - \mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \neq (\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta})] \leq \mathbb{P}_{\mathbf{M}}[(\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta}) \in A] \\ &\mathbb{P}_{\mathbf{M}}[(\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta}) \in A] \leq \mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] + \mathbb{P}_{\mathbf{M}}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \neq (\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta})],\end{aligned}$$

we arrive at

$$\begin{aligned}&\mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] - \Big(\frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta})\Big)^{1/2} \leq \mathbb{P}[(\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta}) \in A] \\ &\mathbb{P}[(\hat{\mathbf{w}}_\eta, \ldots, \hat{\mathbf{w}}_{K\eta}) \in A] \leq \mathbb{P}[(\mathbf{w}_\eta, \ldots, \mathbf{w}_{K\eta}) \in A] + \Big(\frac{1}{2}\mathrm{KL}(\hat{\mu}_{K\eta}, \mu_{K\eta})\Big)^{1/2}.\end{aligned}$$

By Theorem 12, we have the desired inequalities. □

Proof of Theorem 12

First, we derive a Girsanov-type change of measure Øksendal and Sulem [2005], Tankov [2003] for the SDE considered in (6.2). Let $\mathbb{P}$ denote the law of $\mathbf{w}_t$ and $\mathbb{Q}$ be an equivalent measure defined by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{F}_T} = \exp\left(\int_0^T \phi_t \mathrm{dB}_t - \frac{1}{2}\int_0^T \phi_t^2 \mathrm{d}t\right), \tag{7.17}$$

where $\mathcal{F}_T$ denotes the filtration up to time T. Then the process B^ϕ defined by $\mathrm{B}^\phi(t) = \mathrm{B}(t) - \int_0^t \phi_s \mathrm{d}s$ is a $\mathbb{Q}$-Brownian motion. With the choice of ϕ_t given in **A**9, we see that

$\mathbf{w}$ satisfies $\mathrm{d}\mathbf{w}_t = b(\mathbf{w})\mathrm{d}t + \varepsilon\sigma \mathrm{dB}^{\phi}(t) + \varepsilon \mathrm{dL}^{\alpha}(t)$. Since this equation has a unique solution (constructed explicitly with the Euler scheme), we conclude that $\mathbf{w}$ has the same law under $\mathbb{Q}$ as $\hat{\mathbf{w}}$ under $\mathbb{P}$.

We thus have:

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) = \mathrm{KL}(\mathbb{P}_t, \mathbb{Q}_t) = \mathbb{E}^{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\Big|_{\mathcal{F}_t}\right] = \frac{1}{2\varepsilon^2\sigma^2}\mathbb{E}^{\mathbb{P}}\left[\int_0^t \|b(\hat{\mathbf{w}}) + \nabla f(\hat{\mathbf{w}}_s)\|^2 \mathrm{d}s\right] \tag{7.18}$$

By using the same steps of the proof of Raginsky et al. [2017][Lemma 3.6], we obtain

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) = \frac{1}{2\varepsilon^2\sigma^2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbb{E}\|\nabla f(\hat{\mathbf{w}}_s) - \nabla f(\hat{\mathbf{w}}(j\eta))\|^2 \,\mathrm{d}s \tag{7.19}$$

$$\leq \frac{M^2}{2\varepsilon^2\sigma^2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta} \mathbb{E}\|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{j\eta}\|^{2\gamma} \,\mathrm{d}s. \tag{7.20}$$

Next, we prove the following theorem.

Theorem 15. *Under assumptions **A**10 and **A**11 we have, for* $0 < \eta \leq \min\{1, \frac{m}{M^2}\}$,

$$\begin{aligned}
&\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq \\
&\leq \frac{M^2 3^{\gamma}}{2\varepsilon^2\sigma^2} k\eta \Bigg(C M^{2\gamma}\eta^{2\gamma}\Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^{2^{\gamma^2}} \\
&\quad + \frac{k-1}{2}\Big((2\eta(b+m))^{\gamma^2} + 2^{\gamma^2}(\eta B)^{2^{\gamma^2}} + \varepsilon^{2^{\gamma^2}}\eta^{\frac{2\gamma^2}{\alpha}} d\Big(\frac{2^{2^{\gamma^2}}\Gamma((1+2\gamma^2)/2)\Gamma(1-2\gamma^2/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma^2)}\Big) \\
&\quad + \varepsilon^{2^{\gamma^2}}\eta^{\gamma^2} d\Big(2^{\gamma^2}\frac{\Gamma\left(\frac{2\gamma^2+1}{2}\right)}{\sqrt{\pi}}\Big)\Big) + \frac{B^2}{M^2}\Big) + (\varepsilon\eta^{1/\alpha})^{2\gamma} d^{2\gamma}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big) \\
&\quad + (\varepsilon\eta^{1/2})^{2\gamma} d^{2\gamma}\Big(2^{\gamma}\frac{\Gamma\left(\frac{2\gamma+1}{2}\right)}{\sqrt{\pi}}\Big)\Bigg) \\
&\leq K_1 k^2 \eta^{1+2\gamma+\gamma^2} + K_2 k\eta^{1+2\gamma} + K_3 k\eta^{1+\frac{2\gamma}{\alpha}} + K_4 k\eta^{1+\gamma},
\end{aligned}$$

where

$$\begin{aligned}
K_1 &\triangleq \frac{CM^{2+2\gamma}3^{\gamma}}{\varepsilon^2\sigma^2}\max\Big\{(2(b+m))^{\gamma^2}, 2^{\gamma^2}B^{2^{\gamma^2}}, \varepsilon^{2^{\gamma^2}} d\Big(\frac{2^{2^{\gamma^2}}\Gamma((1+2\gamma^2)/2)\Gamma(1-2\gamma^2/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma^2)}\Big), \\
&\qquad \varepsilon^{2^{\gamma^2}} d\Big(2^{\gamma^2}\frac{\Gamma\left(\frac{2\gamma^2+1}{2}\right)}{\sqrt{\pi}}\Big)\Big)\Big\}, \\
K_2 &\triangleq \frac{CM^{2+2\gamma}3^{\gamma}}{2\varepsilon^2\sigma^2}\Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^{2^{\gamma^2}} + \frac{B^2}{M^2}\Big), \\
K_3 &\triangleq \frac{M^2 3^{\gamma}\varepsilon^{2\gamma-2}d^{2\gamma}}{2\sigma^2}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big), \\
K_4 &\triangleq \frac{M^2 3^{\gamma}\varepsilon^{2\gamma-2}d^{2\gamma}}{2\sigma^2}\Big(2^{\gamma}\frac{\Gamma\left(\frac{2\gamma+1}{2}\right)}{\sqrt{\pi}}\Big).
\end{aligned}$$

Proof. Let us consider the term $\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{j\eta}$, for $s \in [j\eta, (j+1)\eta]$:

$$\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{j\eta} = -(s - j\eta)\nabla f(\hat{\mathbf{w}}_{j\eta}) + \varepsilon(L_s - L_{j\eta}) + \varepsilon(B_s - B_{j\eta}) \tag{7.21}$$

$$\triangleq T_1 + T_2 + T_3 \tag{7.22}$$

Using this equation and (7.20), we obtain:

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq \frac{M^2}{2\varepsilon^2\sigma^2} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}\|T_1 + T_2 + T_3\|^{2\gamma} \, \mathrm{d}s \tag{7.23}$$

$$\leq \frac{M^2}{2\varepsilon^2\sigma^2} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}\Big(\|T_1 + T_2 + T_3\|^2\Big)^{\gamma} \, \mathrm{d}s \tag{7.24}$$

$$\leq \frac{M^2}{2\varepsilon^2\sigma^2} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}\Big(3\|T_1\|^2 + 3\|T_2\|^2 + 3\|T_3\|^2\Big)^{\gamma} \, \mathrm{d}s \tag{7.25}$$

$$\leq \frac{M^2 3^{\gamma}}{2\varepsilon^2\sigma^2} \sum_{j=0}^{k-1} \int_{j\eta}^{(j+1)\eta} \mathbb{E}\Big(\|T_1\|^{2\gamma} + \|T_2\|^{2\gamma} + \|T_3\|^{2\gamma}\Big) \, \mathrm{d}s \tag{7.26}$$

where (7.25) is obtained from $(a+b)^{\gamma} \leq a^{\gamma} + b^{\gamma}$ since $\gamma \in (0,1)$ and $a, b \geq 0$.

Since $2\gamma > 1$, we have by Corollary 9

$$\begin{aligned}
\mathbb{E}\|T_2\|^{2\gamma} =& \mathbb{E}\|\varepsilon(s - j\eta)^{1/\alpha}\mathrm{L}^{\alpha}(1))\|^{2\gamma} \\
\leq& (\varepsilon\eta^{1/\alpha})^{2\gamma}\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{2\gamma} \\
\leq& (\varepsilon\eta^{1/\alpha})^{2\gamma} d^{2\gamma}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big),
\end{aligned}$$

and by Corollary 10,

$$\begin{aligned}
\mathbb{E}\|T_3\|^{2\gamma} =& \mathbb{E}\|\varepsilon(s - j\eta)^{1/2}\mathrm{B}(1))\|^{2\gamma} \\
\leq& (\varepsilon\eta^{1/2})^{2\gamma}\mathbb{E}\|\mathrm{B}(1)\|^{2\gamma} \\
\leq& (\varepsilon\eta^{1/2})^{2\gamma} d^{2\gamma}\left(2^{\gamma}\frac{\Gamma\left(\frac{2\gamma+1}{2}\right)}{\sqrt{\pi}}\right),
\end{aligned}$$

By definition, we have

$$\mathbb{E}\|T_1\|^{2\gamma} = \mathbb{E}\|(s - j\eta)\nabla f(\hat{\mathbf{w}}_{j\eta})\|^{2\gamma} \tag{7.27}$$

$$\leq \eta^{2\gamma}\mathbb{E}\|\nabla f(\hat{\mathbf{w}}_{j\eta})\|^{2\gamma} \tag{7.28}$$

$$\leq \eta^{2\gamma}\mathbb{E}(M\|\hat{\mathbf{w}}_{j\eta}\|^{\gamma} + B)^{2\gamma} \tag{7.29}$$

$$\leq CM^{2\gamma}\eta^{2\gamma}\mathbb{E}\left(\|\hat{\mathbf{w}}_{j\eta}\|_{\gamma}^{\gamma} + \Big(\frac{B^{1/\gamma}}{M^{1/\gamma}}\Big)^{\gamma}\right)^{2\gamma} \tag{7.30}$$

$$\leq CM^{2\gamma}\eta^{2\gamma}\mathbb{E}\left(\|\hat{\mathbf{w}}'_{j\eta}\|_{\gamma}^{\gamma}\right)^{2\gamma} \tag{7.31}$$

where we used the equivalence of ℓ_p-norms and $\hat{\mathbf{w}}'_{j\eta}$ is the concatenation of $\hat{\mathbf{w}}_{j\eta}$ and $\frac{B^{1/\gamma}}{M^{1/\gamma}}$. We then obtain

$$\mathbb{E}\|T_1\|^{2\gamma} \leq CM^{2\gamma}\eta^{2\gamma}\mathbb{E}\|\hat{\mathbf{w}}' -_{j\eta}\|_{\gamma}^{2\gamma^2} \tag{7.32}$$

$$\leq CM^{2\gamma}\eta^{2\gamma}\mathbb{E}\|\hat{\mathbf{w}}'_{j\eta}\|_{2\gamma^2}^{2\gamma^2} \tag{7.33}$$

$$= CM^{2\gamma}\eta^{2\gamma}\mathbb{E}\Big(\|\hat{\mathbf{w}}_{j\eta}\|_{2\gamma^2}^{2\gamma^2} + \frac{B^2}{M^2}\Big) \tag{7.34}$$

$$\leq CM^{2\gamma}\eta^{2\gamma}\Big(\mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^{2\gamma^2} + \frac{B^2}{M^2}\Big). \tag{7.35}$$

By combining the above inequalities and Lemma 23, we obtain

$$\begin{aligned}
&\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq \\
&\leq \frac{M^2 3^\gamma}{2\varepsilon^2\sigma^2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}\Big(\|T_1\|^{2\gamma} + \|T_2\|^{2\gamma} + \|T_3\|^{2\gamma}\Big)\,\mathrm{d}s \\
&\leq \frac{M^2 3^\gamma}{2\varepsilon^2\sigma^2}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\Bigg(CM^{2\gamma}\eta^{2\gamma}\Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^{2\gamma^2} \\
&\quad + j\Big((2\eta(b+m))^{\gamma^2} + 2^{\gamma^2}(\eta B)^{2\gamma^2} + \varepsilon^{2\gamma^2}\eta^{\frac{2\gamma^2}{\alpha}}d\Big(\frac{2^{2\gamma^2}\Gamma((1+2\gamma^2)/2)\Gamma(1-2\gamma^2/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma^2)}\Big) \\
&\quad + \varepsilon^{2\gamma^2}\eta^{\gamma^2}d\Big(2^{\gamma^2}\frac{\Gamma\left(\frac{2\gamma^2+1}{2}\right)}{\sqrt{\pi}}\Big)\Big) + \frac{B^2}{M^2}\Big) + (\varepsilon\eta^{1/\alpha})^{2\gamma}d^{2\gamma}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big) \\
&\quad + (\varepsilon\eta^{1/2})^{2\gamma}d^{2\gamma}\Big(2^{\gamma}\frac{\Gamma\left(\frac{2\gamma+1}{2}\right)}{\sqrt{\pi}}\Big)\Bigg)\,\mathrm{d}s \\
&= \frac{M^2 3^\gamma}{2\varepsilon^2\sigma^2}k\eta\Bigg(CM^{2\gamma}\eta^{2\gamma}\Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^{2\gamma^2} \\
&\quad + \frac{k-1}{2}\Big((2\eta(b+m))^{\gamma^2} + 2^{\gamma^2}(\eta B)^{2\gamma^2} + \varepsilon^{2\gamma^2}\eta^{\frac{2\gamma^2}{\alpha}}d\Big(\frac{2^{2\gamma^2}\Gamma((1+2\gamma^2)/2)\Gamma(1-2\gamma^2/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma^2)}\Big) \\
&\quad + \varepsilon^{2\gamma^2}\eta^{\gamma^2}d\Big(2^{\gamma^2}\frac{\Gamma\left(\frac{2\gamma^2+1}{2}\right)}{\sqrt{\pi}}\Big)\Big) + \frac{B^2}{M^2}\Big) + (\varepsilon\eta^{1/\alpha})^{2\gamma}d^{2\gamma}\Big(\frac{2^{2\gamma}\Gamma((1+2\gamma)/2)\Gamma(1-2\gamma/\alpha)}{\Gamma(1/2)\Gamma(1-\gamma)}\Big) \\
&\quad + (\varepsilon\eta^{1/2})^{2\gamma}d^{2\gamma}\Big(2^{\gamma}\frac{\Gamma\left(\frac{2\gamma+1}{2}\right)}{\sqrt{\pi}}\Big)\Bigg).
\end{aligned}$$

By defining the constants K_1, K_2, K_3 and K_4 as in the statement of the theorem, we directly have the conclusion. □

The proof of Theorem 12 is given bellow.

Proof. By Theorem 15, we have

$$\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq K_1 k^2\eta^{1+2\gamma+\gamma^2} + K_2 k\eta^{1+2\gamma} + K_3 k\eta^{1+\frac{2\gamma}{\alpha}} + K_4 k\eta^{1+\gamma}.$$

We can easily check that, for example, if $0 < \eta \leq \left(\frac{\delta^2}{2K_1 t^2}\right)^{\frac{1}{\gamma^2+2\gamma-1}}$, then $K_1 k^2 \eta^{1+2\gamma+\gamma^2} \leq \frac{\delta^2}{2}$. By the same arguments, we finally have

$$\begin{aligned}\mathrm{KL}(\hat{\mu}_t, \mu_t) \leq & \frac{\delta^2}{2} + \frac{\delta^2}{2} + \frac{\delta^2}{2} + \frac{\delta^2}{2} \\ = & 2\delta^2.\end{aligned}$$

This finalizes the proof. □

Technical results

Lemma 21. *Under assumptions **A**10 and **A**11 we have*

$$\|\nabla f(w)\| \leq M\|w\|^\gamma + B, \quad \forall w \in \mathbb{R}^d.$$

Proof. By assumption **A**10 we have

$$\|\nabla f(w) - \nabla f(0)\| \leq M\|w - 0\|^\gamma.$$

Since $\|\nabla f(0)\| \leq B$ by assumption **A**11, the conclusion follows. □

For the moments of the noise B(1), we have the following corollary.

Lemma 22. *Let X be a scalar standard Gaussian random variable. Then, for $\lambda > -1$, we have*

$$\mathbb{E}(|X|^\lambda) = 2^{\lambda/2} \frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}},$$

where Γ denotes the Gamma function.

Proof. The result is a direct consequence of equation (17) in Winkelbauer [2012]. □

Corollary 10. *Let $B(1)$ be a d-dimensional vector whose components are i.i.d standard Gaussian random variable. The quantity $\mathbb{E}\|B(1)\|^\lambda$ is finite for $\lambda > -1$. For details, we have*

(a) If $1 < \lambda < \alpha$, then

$$\mathbb{E}\|B(1)\|^\lambda \leq d^\lambda \left(2^{\lambda/2} \frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\right).$$

(b) If $0 \leq \lambda \leq 1$, then

$$\mathbb{E}\|B(1)\|^\lambda \leq d \left(2^{\lambda/2} \frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\right).$$

Proof. Since B(1), by definition, is a d-dimensional vector whose components are i.i.d standard Gaussian random variable $B_i(1)$ for $i \in \{1, \dots, d\}$, we have

$$\|\mathrm{B}(1)\| \leq \sum_{i=1}^{d} |B_i(1)|.$$

(a) $1 < \lambda < \alpha$. By using Minkowski's inequality and Lemma 22,

$$(\mathbb{E}\|\mathrm{B}(1)\|^{\lambda})^{1/\lambda} \leq \Big(\mathbb{E}\Big[\Big(\sum_{i=1}^{d} |B_i(1)|\Big)^{\lambda}\Big]\Big)^{1/\lambda} \leq \sum_{i=1}^{d} (\mathbb{E}|B_i(1)|^{\lambda})^{1/\lambda} = d\left(2^{\lambda/2}\frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\right)^{1/\lambda}.$$

Thus, we have

$$\mathbb{E}\|\mathrm{B}(1)\|^{\lambda} \leq d^{\lambda}\left(2^{\lambda/2}\frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\right).$$

(b) $0 \leq \lambda \leq 1$.

$$\mathbb{E}\|\mathrm{B}(1)\|^{\lambda} \leq \mathbb{E}\Big[\Big(\sum_{i=1}^{d} |B_i(1)|\Big)^{\lambda}\Big] \leq \sum_{i=1}^{d} \mathbb{E}|B_i(1)|^{\lambda} = d\left(2^{\lambda/2}\frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\right).$$

□

Lemma 23. *For $0 < \eta \leq \frac{m}{M^2}$ and $s \in [j\eta, (j+1)\eta)$, we have the following estimates:*

(a) If $1 < \lambda < \alpha$ then

$$\mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^{\lambda} \leq \Bigg(\Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^{\lambda}\Big)^{\frac{1}{\lambda}} + j\Big((2\eta(b+m))^{\frac{1}{2}} + 2^{\frac{1}{2}}\eta B + \varepsilon\eta^{\frac{1}{\alpha}} d\Big(\frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big)^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{2}} d\Big(2^{\lambda/2}\frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\Big)^{\frac{1}{\lambda}}\Big)\Bigg)^{\lambda}.$$

(b) If $0 \leq \lambda \leq 1$ then

$$\mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^{\lambda} \leq \mathbb{E}\|\hat{\mathbf{w}}_0\|^{\lambda} + j\Big((2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^{\lambda} + \varepsilon^{\lambda}\eta^{\frac{\lambda}{\alpha}} d\Big(\frac{2^{\lambda}\Gamma((1+\lambda)/2)\Gamma(1-\lambda/\alpha)}{\Gamma(1/2)\Gamma(1-\lambda/2)}\Big) + \varepsilon^{\lambda}\eta^{\frac{\lambda}{2}} d\Big(2^{\lambda/2}\frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\pi}}\Big)\Big).$$

Proof. Let us denote the value $\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda}$ by $l_{\alpha,\lambda,d} < \infty$ and the value $\mathbb{E}\|\mathrm{B}(1)\|^{\lambda}$ by $b_{\lambda,d} < \infty$. Starting from

$$\hat{\mathbf{w}}_{(j+1)\eta} = \hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta}) + \varepsilon\eta^{\frac{1}{\alpha}}\mathrm{L}^{\alpha}(1) + \varepsilon\eta^{\frac{1}{2}}\mathrm{B}(1),$$

we have either, by Minkowski, for $\lambda > 1$,

$$\Big(\mathbb{E}\|\hat{\mathbf{w}}_{(j+1)\eta}\|^{\lambda}\Big)^{\frac{1}{\lambda}} \leq \Big(\mathbb{E}\|\hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta})\|^{\lambda}\Big)^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{\alpha}}\Big(\mathbb{E}\|\mathrm{L}^{\alpha}(1)\|^{\lambda}\Big)^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{2}}\Big(\mathbb{E}\|\mathrm{B}(1)\|^{\lambda}\Big)^{\frac{1}{\lambda}}, \tag{7.36}$$

or for $0 \leq \lambda \leq 1$),

$$\mathbb{E}\|\hat{\mathbf{w}}_{(j+1)\eta}\|^\lambda \leq \mathbb{E}\|\hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta})\|^\lambda + \varepsilon^\lambda \eta^{\frac{\lambda}{\alpha}} \mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda + \varepsilon^\lambda \eta^{\frac{\lambda}{2}} \mathbb{E}\|\mathrm{B}(1)\|^\lambda. \tag{7.37}$$

Consider the first term on the right side:

$$\begin{aligned}
\|\hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta})\|^\lambda &= \|\hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta})\|^{2\times\frac{\lambda}{2}} \\
&= \Big(\|\hat{\mathbf{w}}_{j\eta}\|^2 - 2\eta\langle\hat{\mathbf{w}}_{j\eta}, \nabla f(\hat{\mathbf{w}}_{j\eta}\rangle + \eta^2\|\nabla f(\hat{\mathbf{w}}_{j\eta}\|^2\Big)^{\frac{\lambda}{2}} \\
&\leq \Big(\|\hat{\mathbf{w}}_{j\eta}\|^2 - 2\eta(m\|\hat{\mathbf{w}}_{j\eta}\|^{1+\gamma} - b) + \eta^2(2M^2\|\hat{\mathbf{w}}_{j\eta}\|^{2\gamma} + 2B^2)\Big)^{\frac{\lambda}{2}},
\end{aligned} \tag{7.38}$$

where we used assumption **A**12 and Lemma 21. For $0 < \eta \leq \frac{m}{M^2}$,

$$2\eta m(\|\hat{\mathbf{w}}_{j\eta}\|^{1+\gamma} + 1) \geq 2\eta^2 M^2 \|\hat{\mathbf{w}}_{j\eta}\|^{2\gamma}. \qquad \text{(since } 1+\gamma > 2\gamma \text{ and } \eta m > \eta^2 M^2\text{)}$$

Using this inequality we have

$$\begin{aligned}
\|\hat{\mathbf{w}}_{j\eta} - \eta\nabla f(\hat{\mathbf{w}}_{j\eta})\|^\lambda &\leq \Big(\|\hat{\mathbf{w}}_{j\eta}\|^2 + 2\eta(b+m) + 2\eta^2 B^2\Big)^{\frac{\lambda}{2}} \\
&\leq \|\hat{\mathbf{w}}_{j\eta}\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda.
\end{aligned} \tag{7.39}$$

Consider the case where $\lambda > 1$. By (7.36) and (7.39),

$$\begin{aligned}
&\Big(\mathbb{E}\|\hat{\mathbf{w}}_{(j+1)\eta}\|^\lambda\Big)^{\frac{1}{\lambda}} \leq \\
&\leq \Big(\mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda\Big)^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{\alpha}}\Big(\mathbb{E}\|\mathrm{L}^\alpha(1)\|^\lambda\Big)^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{2}}\Big(\mathbb{E}\|\mathrm{B}(1)\|^\lambda\Big)^{\frac{1}{\lambda}} \\
&\leq \Big(\mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^\lambda\Big)^{\frac{1}{\lambda}} + (2\eta(b+m))^{\frac{1}{2}} + 2^{\frac{1}{2}}\eta B + \varepsilon\eta^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{2}} b_{\lambda,d}^{\frac{1}{\lambda}} \\
&\leq \Big(\mathbb{E}\|\hat{\mathbf{w}}_0\|^\lambda\Big)^{\frac{1}{\lambda}} + (j+1)\Big((2\eta(b+m))^{\frac{1}{2}} + 2^{\frac{1}{2}}\eta B + \varepsilon\eta^{\frac{1}{\alpha}} l_{\alpha,\lambda,d}^{\frac{1}{\lambda}} + \varepsilon\eta^{\frac{1}{2}} b_{\lambda,d}^{\frac{1}{\lambda}}\Big).
\end{aligned}$$

For the case where $0 \leq \lambda \leq 1$, by (7.37) and (7.39),

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{w}}_{(j+1)\eta}\|^\lambda &\leq \mathbb{E}\|\hat{\mathbf{w}}_{j\eta}\|^\lambda + (2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda + \varepsilon^\lambda\eta^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d} + \varepsilon^\lambda\eta^{\frac{\lambda}{2}} b_{\lambda,d} \\
&\leq \mathbb{E}\|\hat{\mathbf{w}}_0\|^\lambda + (j+1)\Big((2\eta(b+m))^{\frac{\lambda}{2}} + 2^{\frac{\lambda}{2}}(\eta B)^\lambda + \varepsilon^\lambda\eta^{\frac{\lambda}{\alpha}} l_{\alpha,\lambda,d} + \varepsilon^\lambda\eta^{\frac{\lambda}{2}} b_{\lambda,d}\Big).
\end{aligned}$$

By using Corollary 9 and Corollary 10, we have the desired results.

□

Details of the simulations in Section 6.4

The detailed settings of the parameters for the synthetic experiment (Figure 6.1) are as follows.

Figure 6.1(a) $d = 10$, $\alpha \in \{1.2, 1.4, 1.6, 1.8\}$, $\varepsilon = 0.1$, $\sigma = 1$, $a = 4 \times 10^{-4}$.
Figure 6.1(b) $d = 10$, $\alpha \in \{1.2, 1.4, 1.6, 1.8\}$, $\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 10\}$, $\sigma = 1$, $a = 4 \times 10^{-6}$.
Figure 6.1(c) $d = 10$, $\alpha \in \{1.2, 1.4, 1.6, 1.8\}$, $\varepsilon = 0.1$, $\sigma \in \{10^{-2}, 10^{-1}, 1, 10\}$, $a = 4 \times 10^{-5}$.
Figure 6.1(d) $d \in \{10, 40, 70, 100\}$, $\alpha \in \{1.2, 1.4, 1.6, 1.8\}$, $\varepsilon = 0.1$, $\sigma = 1$, $a = 4 \times 10^{-4}$.

Bibliography

M. S. Advani and A. M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

S. I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

M. A. Armenta and P. M. Jodoin. The representation theory of neural networks. *arXiv preprint arXiv:2007.12213*, 2020.

M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. B. Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, volume 80, pages 314–323, Stockholm Sweden, 10–15 Jul 2018.

J. Baker, P. Fearnhead, E. B Fox, and C. Nemeth. sgmcmc: An R package for stochastic gradient Markov chain Monte Carlo. *arXiv preprint arXiv:1710.00578*, 2017.

J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

R. F. Bass. Uniqueness in law for pure jump Markov processes. *Probability Theory and Related Fields*, 79(2):271–287, 1988.

E. Bayraktar, S. Nadtochiy, et al. Weak reflection principle for Lévy processes. *The Annals of Applied Probability*, 25(6):3251–3294, 2015.

N. Berglund. Kramers' law: Validity, derivations and generalisations. *arXiv preprint arXiv:1106.5799*, 2011.

D. P. Bertsekas. Incremental least squares methods and the extended Kalman filter. *SIAM Journal on Optimization*, 6(3):807–822, 1996.

T Birdal and U Şimşekli. Probabilistic permutation synchronization using the Riemannian structure of the Birkhoff polytope. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11097–11108, 2019.

T. Birdal, U. Şimşekli, M. O. Eken, and S. Ilic. Bayesian pose graph optimization via Bingham distributions and tempered geodesic MCMC. In *Advances in Neural Information Processing Systems*, pages 308–319, 2018.

N. Y. Bobrov, N. A. Smirnova, F. Vallianatos, and J. P. Makris. Multifractal analysis: a method to investigate non-stationary properties of geophysical processes. In *Proceeding of the 2005 WSEAS International Conference on ENGINEERING EDUCATION (Eds: D. Triantis & F. Vallianatos)*, pages paper–507, 2005.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

R. F. Brcich, D. R. Iskander, and A. M. Zoubir. The stability test for symmetric alpha-stable distributions. *IEEE Transactions on Signal Processing*, 53(3):977–986, 2005.

T. Burghoff and I. Pavlyukevich. Spectral analysis for a discrete metastable system driven by Lévy flights. *Journal of Statistical Physics*, 161(1):171–196, 2015.

G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

C. Çelik and M. Duman. Crank–Nicolson method for the fractional diffusion equation with the Riesz fractional derivative. *Journal of Computational Physics*, 231(4):1743–1750, 2012.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pages 2269–2277, 2015.

X. Chen, S. S. Du, and X. T. Tong. On stationary-point hitting time and ergodicity of stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 21 (68):1–41, 2020.

N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728, 2016.

U. Şimşekli, R. Badeau, A. T. Cemgil, and G. Richard. Stochastic quasi-Newton Langevin Monte Carlo. In *ICML*, 2016.

U. Şimşekli, C. Yildiz, T. H. Nguyen, A. T. Cemgil, and G. Richard. Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. In *ICML*, pages 4674–4683, 2018.

U. Şimşekli, L. Sagun, and Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *ICML*, 2019.

S. Cunningham, H. Ridley, J. Weinel, and R. Picking. Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks. *Personal and Ubiquitous Computing*, pages 1–14, 2020.

A. S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *Proceedings of the 2017 Conference on Learning Theory*, 2017a.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017b.

H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann. Escaping saddles with stochastic gradients. In *ICML*, pages 1155–1164, 2018.

M. V. Day. On the exponential exit law in the small parameter exit problem. *Stochastics*, 8(4):297–323, 1983.

L. De Haan and L. Peng. Comparison of tail index estimators. *Statistica Neerlandica*, 52(1):60–70, 1998.

A. Debussche and N. Fournier. Existence of densities for stable-like driven SDE's with Hölder continuous coefficients. *Journal of Functional Analysis*, 264(8):1757–1778, 2013.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

N. Dehmamy, A. L. Barabási, and R. Yu. Understanding the representation power of graph neural networks in learning graph topology. In *Advances in Neural Information Processing Systems*, pages 15413–15423, 2019.

A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

R. Der and D. D. Lee. Beyond Gaussian processes: On the distributions of infinite networks. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 275–282. MIT Press, 2006.

T. Dozat. Incorporating nesterov momentum into adam. 2016.

J. Duan. *An Introduction to Stochastic Dynamics*. Cambridge University Press, New York, 2015.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *arXiv preprint arXiv:1507.05021*, 2015.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

A. Durmus, U. Şimşekli, E. Moulines, R. Badeau, and G. Richard. Stochastic gradient Richardson-Romberg Markov Chain Monte Carlo. In *NIPS*, 2016.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9693–9702, 2018.

H. Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

M. I. Freidlin and A. D. Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

J. Gairing, M. Högele, and T. Kosenkova. Transportation distances and noise sensitivity of multiplicative Lévy SDE with applications. *Stochastic Processes and their Applications*, 128(7):2153–2178, 2018.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Global convergence of stochastic gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv e-prints*, art. arXiv:1809.04618, Sep 2018a.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. *arXiv e-prints*, art. arXiv:1812.07725, Dec 2018b.

M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart. The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *arXiv preprint arXiv:1809.09349*, 2018.

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in R^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, pages 1163–1174, 1975.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012a.

G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning. *Coursera, video lectures*, 264(1), 2012b.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

C. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

P. Imkeller and I. Pavlyukevich. First exit times of SDEs driven by stable Lévy processes. *Stochastic Processes and their Applications*, 116(4):611–642, 2006.

P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in rd perturbed by multifractal Lévy noise. *Journal of Statistical Physics*, 141(1):94–119, 2010a.

P. Imkeller, I. Pavlyukevich, and T. Wetzel. The hierarchy of exit times of Lévy-driven Langevin equations. *The European Physical Journal Special Topics*, 191(1):211–222, 2010b.

A. Jacot-Guillarmod, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589. 2018.

M. Jas, T. Dupré La Tour, U. Şimşekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems*, pages 1099–1108, 2017.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

B. Jourdain, S. Méléard, and W. A. Woyczynski. Lévy flights in evolutionary ecology. *Journal of Mathematical Biology*, 65(4):677–707, 2012.

J. Kallsen and P. Tankov. Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis*, 97(7):1551–1572, 2006.

N. S. Keskar and R. Socher. Improving generalization performance by switching from Adam to SGD. *arXiv preprint arXiv:1712.07628*, 2017.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

A. I. Khan and S. Al-Habsi. Machine learning in computer vision. *Procedia Computer Science*, 167:1444–1451, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

I. Kuhwald and I. Pavlyukevich. Bistable behaviour of a jump-diffusion driven by a periodic stable-like additive process. *Discrete & Continuous Dynamical Systems-Series B*, 21(9), 2016.

A. M. Kulik. On weak uniqueness and distributional properties of a solution to an SDE with α-stable noise. *Stochastic Processes and their Applications*, 129(2):473–506, 2019.

M. Kunze. Stochastic differential equations. Lecture notes, University of Ulm, 2012.

D. Lamberton and G. Pages. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. *Stochastics and dynamics*, 3(04): 435–451, 2003.

P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

J. H. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. *ICLR*, 2018.

S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard. Alpha-stable multichannel audio source separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 576–580. IEEE, 2017.

P. Lévy. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1937.

J. Li, Y. Sun, J. Su, T. Suzuki, and F. Huang. Understanding generalization in deep learning via tensor methods. *arXiv preprint arXiv:2001.05070*, 2020.

Q. Li, C. Tai, and E. Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110, 2017.

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 597–607. Curran Associates, Inc., 2017.

M. Liang and J. Wang. Gradient estimates and ergodicity for SDEs driven by multiplicative Lévy noises via coupling. *arXiv preprint arXiv:1801.05936*, 2018.

T. Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2015.

A. Liutkus, U. Şimşekli, S. Majewski, A. Durmus, and F. R. Stoter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, 2019.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017.

A. S. Lundervold and A. Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Y. A. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2899–2907, 2015.

B. B. Mandelbrot. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*. Springer Science & Business Media, 2013.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. *arXiv preprint arXiv:1901.08276*, 2019.

D. Masters and C. Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

H. Masuda. Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stochastic processes and their applications*, 117(1):35–56, 2007.

M. Matsui, Z. Pawlas, et al. Fractional absolute moments of heavy tailed distributions. *Brazilian Journal of Probability and Statistics*, 30(2):272–298, 2016.

P. Mertikopoulos and M. Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

R. Mikulevičius and F. Xu. On the rate of convergence of strong Euler approximation for SDEs driven by lévy processes. *Stochastics*, 90(4):569–604, 2018.

R. Mikulevičius and C. Zhang. On the rate of convergence of weak Euler approximation for nondegenerate SDEs driven by Lévy processes. *Stochastic Processes and their Applications*, 121(8):1720–1748, 2011.

S. Mittnik and S. T. Rachev. Tail estimation of the stable index α. *Applied Mathematics Letters*, 9(3):53–56, 1996.

M. Mohammadi, A. Mohammadpour, and H. Ogata. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

T. H. Nguyen, U. Şimşekli, M. Gurbuzbalaban, and G. Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Neurips*, 2019a.

T. H. Nguyen, U. Şimşekli, and G. Richard. Non-asymptotic analysis of fractional Langevin Monte Carlo for non-convex optimization. In *ICML*, 2019b.

B. K. Øksendal and A. Sulem. *Applied stochastic control of jump diffusions*, volume 498. Springer, 2005.

M. D. Ortigueira. Riesz potential operators and inverses via fractional centred derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

M. D. Ortigueira, T. M. Laleg-Kirati, and J. A. T. Machado. Riesz potential versus fractional Laplacian. *Journal of Statistical Mechanics*, (09), 2014.

J. Padmanabhan and M. J. Johnson Premkumar. Machine learning in automatic speech recognition: A survey. *IETE Technical Review*, 32(4):240–251, 2015.

A. Panigrahi, R. Somani, N. Goyal, and P. Netrapalli. Non-Gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

F. Panloup. Recursive computation of the invariant measure of a stochastic differential equation driven by a Lévy process. *The Annals of Applied Probability*, 18(2):379–426, 2008.

Vardan Papyan. The full spectrum of deep net Hessians at scale: Dynamics with sample size. *arXiv preprint arXiv:1811.07062*, 2018.

D. S. Park, J. Sohl-Dickstein, Q. V Le, and S. L. Smith. The effect of network width on stochastic gradient descent and generalization: an empirical study. *arXiv preprint arXiv:1905.03776*, 2019.

V. Paulauskas and M. Vaičiulis. Once more on comparison of tail index estimators. *arXiv preprint arXiv:1104.1242*, 2011.

A. N. Pavlov, O. N. Pavlova, A. S. Abdurashitov, O. A. Sindeeva, O. V. Semyachkina-Glushkovskaya, and J. Kurths. Characterizing scaling properties of complex signals with missed data segments using the multifractal analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(1):013124, 2018.

I. Pavlyukevich. Cooling down lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 40(41):12299, 2007.

I. Pavlyukevich. First exit times of solutions of stochastic differential equations driven by multiplicative Lévy noise with heavy tails. *Stochastics and Dynamics*, 11(02n03): 495–519, 2011.

J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

V. Pipiras and M. S. Taqqu. *Long-range dependence and self-similarity*, volume 45. Cambridge university press, 2017.

T. Poggio, A. Banburski, and Q. Liao. Theoretical issues in deep networks: Approximation, optimization and generalization. *arXiv preprint arXiv:1908.09375*, 2019.

Y. Polyanskiy and Y. Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtskỳ. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15, 2020.

E. Priola et al. Pathwise uniqueness for singular SDEs driven by stable processes. *Osaka Journal of Mathematics*, 49(2):421–447, 2012.

H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Y. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13 (2):206–219, 2019.

M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pages 2847–2854. PMLR, 2017.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1674–1703, 2017.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, December 2002. ISSN 13875841.

C. Sacco, A. B. Radwan, A. Anderson, R. Harik, and E. Gregory. Machine learning in composites manufacturing: A case study of automated fiber placement inspection. *Composite Structures*, page 112514, 2020.

L. Sagun, V. U. Güney, G. Ben Arous, and Y. LeCun. Explorations on high dimensional landscapes. *International Conference on Learning Representations Workshop Contribution, arXiv:1412.6615*, 2015.

L. Sagun, L. Bottou, and Y. LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *ICLR 2018 Workshop Contribution, arXiv:1706.04454*, 2017.

G. Samorodnitsky and M. Grigoriu. Tails of solutions of certain nonlinear stochastic differential equations driven by heavy tailed Lévy motions. *Stochastic Processes and their Applications*, 105(1):69 – 97, 2003. ISSN 0304-4149.

G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC press, 1994.

D. Schertzer, M. Larchevêque, J. Duan, V. V. Yanovsky, and S. Lovejoy. Fractional Fokker–Planck equation for nonlinear stochastic differential equations driven by non-Gaussian Lévy stable noises. *Journal of Mathematical Physics*, 42(1):200–212, 2001.

N. N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

D. N. Shanbhag and M. Sreehari. On certain self-decomposable distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 38(3):217–222, 1977.

B. Shetty. Natural language processing (NLP) for machine learning. *Retrieved November*, 24:2018, 2018.

U. Şimşekli. Fractional Langevin Monte carlo: Exploring Lévy driven stochastic differential equations for Markov chain Monte Carlo. In *ICML*, pages 3200–3209, 2017.

U. Şimşekli, A. Liutkus, and A. T. Cemgil. Alpha-stable matrix factorization. *IEEE Signal Processing Letters*, 22(12):2289–2293, 2015.

U. Şimşekli, H. Erdoğan, S. Leglaive, A. Liutkus, R. Badeau, and G. Richard. Alpha-stable low-rank plus residual decomposition for speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 651–655. IEEE, 2018.

U. Şimşekli, M. Gürbüzbalaban, T. H. Nguyen, G. Richard, and L. Sagun. On the heavy-tailed theory of Stochastic Gradient Descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

U. Şimşekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. *arXiv preprint arXiv:2006.09313*, 2020.

S. L. Smith, P. J. Kindermans, C. Ying, and Q. V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

D. Soydaner. A comparison of optimization algorithms for deep learning. *International Journal of Pattern Recognition and Artificial Intelligence*, page 2052013, 2020.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

P. Tankov. *Financial modelling with jump processes*. Chapman and Hall/CRC, 2003.

M. Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

S. S. Teotia and D. Kumar. Role of multifractal analysis in understanding the preparation zone for large size earthquake in the North-Western Himalaya region. *Nonlinear Processes in Geophysics*, 18(1):111–118, 2011.

T. Tieleman and G. Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical Report.*, 2017.

B. Tzen, T. Liang, and M. Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. In *Proceedings of the 2018 Conference on Learning Theory*, 2018.

J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

C. Villani. Topics in optimal transportation, volume 58 of Graduate. *Studies in Mathematics*, 2003.

E. R. Weeks, T. H. Solomon, J. S. Urbach, and H. L. Swinney. Observation of anomalous diffusion and Lévy flights. In *Lévy flights and related topics in physics*, pages 51–71. Springer, 1995.

M. Welling and Y. W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, June 2011.

A. Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

W. A. Woyczyński. Lévy processes in the physical sciences. In *Lévy processes*, pages 241–266. Springer, 2001.

L. Wu, C. Ma, and E. Weinan. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8289–8298, 2018.

K. Xenou, G. Chalkiadakis, and S. Afantenos. Deep reinforcement learning in strategic board game environments. In *European Conference on Multi-Agent Systems*, pages 233–248. Springer, 2018.

L. Xiao, J. Pennington, and S. S. Schoenholz. Disentangling trainability and generalization in deep neural networks.

L. Xie and X. Zhang. Ergodicity of stochastic differential equations with jumps and singular coefficients. *arXiv preprint arXiv:1705.07402*, 2017.

C. Xing, D. Arpit, C. Tsirigotis, and Y. Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3125–3136, 2018.

S. Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

X. Yang et al. Multifractality of jump diffusion processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 54, pages 2042–2074. Institut Henri Poincaré, 2018.

V. V. Yanovsky, A. V. Chechkin, D. Schertzer, and A. V. Tur. Lévy anomalous diffusion and fractional Fokker–Planck equation. *Physica A: Statistical Mechanics and its Applications*, 282(1):13–34, 2000.

N. Ye and Z. Zhu. Stochastic fractional Hamiltonian Monte Carlo. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3019–3025, 7 2018.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017a.

Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1980–2022, 2017b.

Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

ECOLE DOCTORALE

Titre : Heavy-tailed Nature of Stochastic Gradient Descent in Deep Learning: Theoretical and Empirical Analysis

Mots clés : Algorithme du gradient stochastique, apprentissage profond, distribution α-stable

Résumé : Dans cette thèse, nous nous intéressons à l'algorithme du gradient stochastique (SGD). Plus précisément, nous effectuons une analyse théorique et empirique du comportement du bruit de gradient stochastique (GN), qui est défini comme la différence entre le gradient réel et le gradient stochastique, dans les réseaux de neurones profonds. Sur la base de ces résultats, nous apportons une perspective alternative aux approches existantes pour étudier SGD. Le GN dans SGD est souvent considéré comme gaussien pour des raisons mathématiques. Cette hypothèse permet d'étudier SGD comme une équation différentielle stochastique (SDE) pilotée par un mouvement brownien. Nous soutenons que l'hypothèse de la gaussianité pourrait ne pas tenir dans les contextes d'apprentissage profond et donc rendre inappropriées les analyses basées sur le mouvement brownien. Inspiré de phénomènes naturels non gaussiens, nous considérons le GN dans un contexte plus général qui suggère que le GN est mieux approché par un vecteur aléatoire *à queue lourde* α-stable. En conséquence, nous proposons d'analyser SGD comme une discrétisation d'une SDE pilotée par un mouvement Lévy. Premièrement, pour justifier l'hypothèse α-stable, nous menons des expériences sur des scénarios communs d'apprentissage en profondeur et montrons que dans tous les contextes, le GN est hautement non gaussien et présente des queues lourdes. Deuxièmement, sous l'hypothèse du GN à queue lourde, nous fournissons une analyse non asymptotique pour que la dynamique en temps discret SGD converge vers le minimum global en termes de sous-optimalité. Enfin, nous étudions la nature de métastabilité de la SDE pilotée par le mouvement de Lévy qui peut ensuite être exploitée pour clarifier le comportement de SGD, notamment en termes de "préférence de larges minima". Plus précisément, nous fournissons une analyse théorique formelle où nous dérivons des conditions explicites pour la taille de pas de sorte que le comportement de métastabilité de SGD, considéré comme une SDE en temps discret, est similaire à sa limite de temps continu. Nos résultats ouvrent une perspective différente et éclairent davantage l'idée selon laquelle SGD préfère les minima larges.

Title : Heavy-tailed Nature of Stochastic Gradient Descent in Deep Learning: Theoretical and Empirical Analysis

Keywords : Stochastic Gradient Descent, deep learning, α-stable distribution

Abstract : In this thesis, we are concerned with the Stochastic Gradient Descent (SGD) algorithm. Specifically, we perform theoretical and empirical analysis of the behavior of the stochastic gradient noise (GN), which is defined as the difference between the true gradient and the stochastic gradient, in deep neural networks. Based on these results, we bring an alternative perspective to the existing approaches for investigating SGD. The GN in SGD is often considered to be Gaussian for mathematical convenience. This assumption enables SGD to be studied as a stochastic differential equation (SDE) driven by a Brownian motion. We argue that the Gaussianity assumption might fail to hold in deep learning settings and hence render the Brownian motion-based analyses inappropriate. Inspired by non-Gaussian natural phenomena, we consider the GN in a more general context that suggests that the GN is better approximated by a *heavy-tailed* α-stable random vector. Accordingly, we propose to analyze SGD as a discretization of an SDE driven by a Lévy motion. Firstly, to justify the α-stable assumption, we conduct experiments on common deep learning scenarios and show that in all settings, the GN is highly non-Gaussian and exhibits heavy-tails. Secondly, under the heavy-tailed GN assumption, we provide a non-asymptotic analysis for the discrete-time dynamics SGD to converge to the global minimum in terms of suboptimality. Finally, we investigate the metastability nature of the SDE driven by Lévy motion that can then be exploited for clarifying the behavior of SGD, especially in terms of 'preferring wide minima'. More precisely, we provide formal theoretical analysis where we derive explicit conditions for the step-size such that the metastability behavior of SGD, viewed as a discrete-time SDE, is similar to its continuous-time limit. We show that the behaviors of the two systems are indeed similar for small step-sizes and we describe how the error depends on the algorithm and problem parameters. We illustrate our metastability results with simulations on a synthetic model and neural networks. Our results open up a different perspective and shed more light on the view that SGD prefers wide minima.

Institut Polytechnique de Paris
91120 Palaiseau, France