Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise

Umut Şimşekli [*1 2] **Lingjiong Zhu** [*3] **Yee Whye Teh** [2] **Mert Gürbüzbalaban** [4]

Abstract

Stochastic gradient descent with momentum (SGDm) is one of the most popular optimization algorithms in deep learning. While there is a rich theory of SGDm for convex problems, the theory is considerably less developed in the context of deep learning where the problem is non-convex and the gradient noise might exhibit a heavy-tailed behavior, as empirically observed in recent studies. In this study, we consider a *continuous-time* variant of SGDm, known as the underdamped Langevin dynamics (ULD), and investigate its asymptotic properties under heavy-tailed perturbations. Supported by recent studies from statistical physics, we argue both theoretically and empirically that the heavy-tails of such perturbations can result in a bias even when the step-size is small, in the sense that *the optima of stationary distribution* of the dynamics might not match *the optima of the cost function to be optimized*. As a remedy, we develop a novel framework, which we coin as *fractional* ULD (FULD), and prove that FULD targets the so-called Gibbs distribution, whose optima exactly match the optima of the original cost. We observe that the Euler discretization of FULD has noteworthy algorithmic similarities with *natural gradient* methods and *gradient clipping*, bringing a new perspective on understanding their role in deep learning. We support our theory with experiments conducted on a synthetic model and neural networks.

[*]Equal contribution [1]LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France [2]Department of Statistics, University of Oxford, Oxford, UK [3]Department of Mathematics, Florida State University, Tallahassee, USA [4]Department of Management Science and Information Systems, Rutgers Business School, Piscataway, USA. Correspondence to: Umut Şimşekli <umut.simsekli@telecom-paris.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

Gradient-based optimization algorithms have been the de facto choice in deep learning for solving the optimization problems of the form:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \Big\{ f(\mathbf{x}) \triangleq (1/n) \sum\nolimits_{i=1}^{n} f^{(i)}(\mathbf{x}) \Big\}, \quad (1)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ denotes the non-convex loss function, $f^{(i)}$ denotes the loss contributed by an individual data point $i \in \{1, \dots, n\}$, $\mathbf{x} \in \mathbb{R}^d$ denotes the collection of all the parameters of the neural network. Among others, stochastic gradient descent with momentum (SGDm) is one of the most popular algorithms for solving such optimization tasks (see e.g., Sutskever et al. (2013); Smith et al. (2018)), and is based on the following iterative scheme:

$$\tilde{\mathbf{v}}^{k+1} = \tilde{\gamma}\tilde{\mathbf{v}}^{k} - \tilde{\eta}\nabla\tilde{f}_{k+1}(\mathbf{x}^k), \quad \mathbf{x}^{k+1} = \mathbf{x}^k + \tilde{\mathbf{v}}^{k+1}, \quad (2)$$

where k denotes the iteration number, $\tilde{\eta}$ is the step-size, $\tilde{\gamma}$ is the friction, and $\tilde{\mathbf{v}}$ denotes the *velocity* (also referred to as momentum). Here, $\nabla\tilde{f}_k$ denotes the stochastic gradients defined as follows:

$$\nabla\tilde{f}_k(\mathbf{x}) \triangleq (1/b) \sum\nolimits_{i\in\Omega_k} f^{(i)}(\mathbf{x}), \quad (3)$$

where $\Omega_k \subset \{1, \dots, n\}$ denotes a random subset drawn from the set of data points with $|\Omega_k| = b \ll n$ for all k.

When the gradients are computed on all the data points (i.e., $\nabla\tilde{f}_k = \nabla f$), SGDm becomes *deterministic* and can be viewed as a discretization of the following *continuous-time* system (Gao et al., 2018a; Maddison et al., 2018):

$$\mathrm{d}\mathbf{v}_t = -(\gamma\mathbf{v}_t + \nabla f(\mathbf{x}_t))\mathrm{d}t, \qquad \mathrm{d}\mathbf{x}_t = \mathbf{v}_t\mathrm{d}t, \quad (4)$$

where $\mathbf{v}_t$ is still called the velocity. The connection between this system and (2) becomes clearer, if we discretize this system by using the Euler scheme with step-size η:

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \eta(\gamma\mathbf{v}^{k+1} + \nabla f(\mathbf{x}^k)), \\ \mathbf{x}^{k+1} = \mathbf{x}^k + \eta\mathbf{v}^{k+1}, \quad (5)$$

and make the change of variables $\tilde{\mathbf{v}}^k \triangleq \eta\mathbf{v}^k$, $\tilde{\gamma} \triangleq (1 - \eta\gamma)$, and $\tilde{\eta} \triangleq \eta^2$. However, due to the presence of the

stochastic gradient noise $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f(\mathbf{x})$, the sequence $\{\mathbf{x}_k, \mathbf{v}_k\}_{k\in\mathbb{N}_+}$ will be a *stochastic process* and the deterministic system (4) would not be an appropriate proxy.

Understanding the statistical properties of $\{\mathbf{x}_k, \mathbf{v}_k\}_{k\in\mathbb{N}_+}$ would be of crucial importance as it might reveal the peculiar properties that lie behind the performance of SGDm for learning with neural networks. A popular approach for understanding the dynamics of stochastic optimization algorithms in deep learning is to impose some structure on the noise U_k and relate the process (2) to a stochastic differential equation (SDE) (Mandt et al., 2016; Jastrzebski et al., 2017; Hu et al., 2019; Chaudhari & Soatto, 2018; Zhu et al., 2019; Şimşekli et al., 2019b). For instance, by assuming that the second-order moments of the stochastic gradient noise are bounded (i.e., $\mathbb{E}\|U_k(\mathbf{x})\|^2 < \infty$ for all admissible k, $\mathbf{x}$), one might argue that U_k can be approximated by a Gaussian random vector due to the central limit theorem (CLT) (Fischer, 2011). Under this assumption, we might view (2) as a discretization of the following SDE, which is also known as the *underdamped* or *kinetic* Langevin dynamics:

$$\begin{aligned} \mathrm{d}\mathbf{v}_t &= -(\gamma \mathbf{v}_t + \nabla f(\mathbf{x}_t))\mathrm{d}t + \sqrt{2\gamma/\beta}\mathrm{d}\mathrm{B}_t \\ \mathrm{d}\mathbf{x}_t &= \mathbf{v}_t \mathrm{d}t, \end{aligned} \tag{6}$$

where B_t denotes the d-dimensional Brownian motion and $\beta > 0$ is called the inverse temperature variable, measuring the noise intensity along with γ. It is easy to check that, under very mild assumptions, the solution process $\{\mathbf{x}_t, \mathbf{v}_t\}_{t\geq 0}$ admits an invariant distribution whose density is proportional to $\exp(-\beta(f(\mathbf{x}) + \|\mathbf{v}\|^2/2))$, where the function $\|\mathbf{v}\|^2/2$ is often called the *Gaussian kinetic energy* (see e.g. (Betancourt et al., 2017)) and the distribution itself is called the Boltzmann-Gibbs measure (Pavliotis, 2014; Gao et al., 2018a; Hérau & Nier, 2004; Dalalyan & Riou-Durand, 2020). We then observe that the marginal distribution $\mathbf{x}$ in the stationarity has a density proportional to $\exp(-\beta f(\mathbf{x}))$, which indicated that any local minimum of f appears as a local maximum of this density. This is a desirable property since it implies that, when the gradient noise U_k has light tails, the process will spend more time near the local minima of f. Furthermore, it has been shown that as β goes to infinity, the marginal distribution of $\mathbf{x}$ concentrates around the global optimum $\mathbf{x}^\star$. This observation has yielded interesting results for understanding the dynamics of SGDm in the contexts of both sampling and optimization with convex and non-convex potentials f (Gao et al., 2018a;b; Zou et al., 2018; Lu et al., 2017; Şimşekli et al., 2018).

While the Gaussianity assumption can be accurate in certain settings such as small networks (Martin & Mahoney, 2019; Panigrahi et al., 2019), recently it has been empirically demonstrated that in several deep learning setups, the stochastic gradient noise can exhibit a *heavy-tailed* behavior

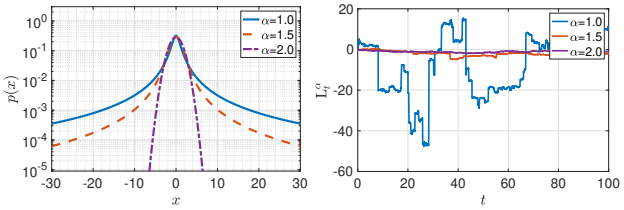


Figure 1. $\mathcal{S}\alpha\mathcal{S}$ densities and L_t^α.

(Şimşekli et al., 2019a; Zhang et al., 2019b)[1]. While the Gaussianity assumption would not be appropriate in this case since the conventional CLT would not hold anymore, nevertheless we can invoke the generalized CLT, which states that the asymptotic distribution of U_k will be a symmetric α-stable distribution ($\mathcal{S}\alpha\mathcal{S}$); a class of distributions that are commonly used in the statistical physics literature as an approximation to heavy-tailed random variables (Sliusarenko et al., 2013; Dubkov et al., 2008). As we will define in more detail in the next section, in the core of $\mathcal{S}\alpha\mathcal{S}$, lies the parameter $\alpha \in (0, 2]$, which determines the heaviness of the tail of the distribution. The tails get heavier as α gets smaller, the case $\alpha = 2$ reduces to the Gaussian random variables. This is illustrated in Figure 1.

Şimşekli et al. (2019b;a) empirically illustrated that, in deep neural networks, the statistical structure of U_k can be better captured by using an α-stable distribution. With the assumption of U_k being $\mathcal{S}\alpha\mathcal{S}$ distributed, the choice of Brownian motion will be no longer appropriate and should be replaced with an α-stable Lévy motion, which motivates the following Lévy-driven SDE:

$$\begin{aligned} \mathrm{d}\mathbf{v}_t &= -(\gamma \mathbf{v}_{t-} + \nabla f(\mathbf{x}_t))\mathrm{d}t + \sqrt{2\gamma/\beta}\mathrm{d}\mathrm{L}_t^\alpha, \\ \mathrm{d}\mathbf{x}_t &= \mathbf{v}_t \mathrm{d}t, \end{aligned} \tag{7}$$

where $\mathbf{v}_{t-}$ denotes the left limit of $\mathbf{v}_t$ and L_t^α denotes the α-stable Lévy process with independent components, which coincides with $\sqrt{2}\mathrm{B}_t$ when $\alpha = 2$. Unfortunately, when $\alpha < 2$, as opposed to its Brownian counterpart, the invariant measures of such SDEs do not admit an analytical form in general; yet, one can still show that the invariant measure cannot be in the form of the Boltzmann-Gibbs measure (Eliazar & Klafter, 2003).

A more striking property of (7) was very recently revealed in a statistical physics study (Capała & Dybiec, 2019), where the authors numerically illustrated that, even when f has a single minimum, the invariant measure of (7) can exhibit

[1] In two recent studies, Gürbüzbalaban et al. (2020) and Hodgkinson & Mahoney (2020) have shown that the stationary distribution of the stochastic gradient descent (SGD) algorithm can be indeed a heavy-tailed distribution depending on the choice of the step-size and the batch-size. On the other hand, in another recent study, Şimşekli et al. (2020) have provided generalization bounds for a general class of SDEs, including heavy- and light-tailed ones.

multiple maxima, none of which coincides with the minimum of f. A similar property has been formally proven in the overdamped dynamics with Cauchy noise (i.e., $\alpha = 1$ and $\gamma \to \infty$) by Sliusarenko et al. (2013). Since the process (7) would spend more time around the modes of its invariant measure (i.e., the high probability region), in an optimization context (i.e., for larger β) the sample paths would concentrate around these modes, which might be arbitrarily distant from the optima of f. In other words, the heavy-tails of the gradient noise could result in an undesirable bias, which would be still present even when the step-size is taken to be arbitrarily small. As we will detail in Section 3, informally, this phenomenon stems from the fact that the heavy-tailed noise leads to aggressive updates on $\mathbf{v}$, which are then directly transmitted to $\mathbf{x}$ due to the dynamics. Unless 'tamed', these updates create an hurling effect on $\mathbf{x}$ and drift it away from the modes of the "potential" f that is sought to be minimized.

Contributions: In this study, we develop a *fractional* underdamped Langevin dynamics whose invariant distribution is guaranteed to be in the form of the Boltzmann-Gibbs measure, hence its optima exactly match the optima of f. We first prove a general theorem which holds for any kinetic energy function, which is not necessarily the Gaussian kinetic energy. However, it turns out that some components of the dynamics might not admit an analytical form for an arbitrary choice of the kinetic energy. Then we identify two choices of kinetic energies, where all the terms in dynamics can be written in an analytical form or accurately computable. We also analyze the Euler discretization of (14) and identify sufficient conditions for ensuring weak convergence of the ergodic averages computed over the iterates.

We observe that the discretization of the proposed dynamics has interesting algorithmic similarities with natural gradient descent (Amari, 1998) and gradient clipping (Pascanu et al., 2013), which we believe bring further theoretical understanding for their role in deep learning. Finally, we support our theory with experiments conducted on both synthetic settings and neural networks.

2. Technical Background & Related Work

The stable distributions are heavy-tailed distributions that appear as the limiting distribution of the generalized CLT for a sum of i.i.d. random variables with infinite variance (Lévy, 1937). In this paper, we are interested in centered *symmetric α-stable distribution*. A scalar random variable X follows a symmetric α-stable distribution denoted as $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ if its characteristic function takes the form: $\mathbb{E}\left[e^{i\omega X}\right] = \exp\left(-\sigma^\alpha|\omega|^\alpha\right)$, $\omega \in \mathbb{R}$, where $\alpha \in (0, 2]$ and $\sigma > 0$. Here, $\alpha \in (0, 2]$ is known as the tail-index, which determines the tail thickness of the distribution. $\mathcal{S}\alpha\mathcal{S}$ becomes heavier-tailed as α gets smaller. $\sigma > 0$ is known as the scale parameter that measures the spread of X around 0. The probability density function of a symmetric α-stable distribution, $\alpha \in (0, 2]$, does not yield closed-form expression in general except for a few special cases. When $\alpha = 1$ and $\alpha = 2$, $\mathcal{S}\alpha\mathcal{S}$ reduces to the Cauchy and the Gaussian distributions, respectively. When $0 < \alpha < 2$, α-stable distributions have heavy-tails so that their moments are finite only up to the order α in the sense that $\mathbb{E}[|X|^p] < \infty$ if and only if $p < \alpha$, which implies infinite variance.

Lévy motions are stochastic processes with independent and stationary increments. Their successive displacements are random and independent, and statistically identical over different time intervals of the same length, and can be viewed as the continuous-time analogue of random walks. The best known and most important examples are the Poisson process, Brownian motion, the Cauchy process and more generally stable processes. Lévy motions are prototypes of Markov processes and of semimartingales, and concern many aspects of probability theory. We refer to (Bertoin, 1996) for a survey on the theory of Lévy motions.

In general, Lévy motions are heavy-tailed, which make it appropriate to model natural phenomena with possibly large variations, that often occurs in statistical physics (Eliazar & Klafter, 2003), signal processing (Kuruoglu, 1999), and finance (Mandelbrot, 1997).

We define L_t^α, a d-dimensional symmetric α-stable Lévy motion with independent components as follows. Each component of L_t^α is an independent scalar α-stable Lévy process, which is defined as follows: (cf. Figure 1)

(i) $\mathrm{L}_0^\alpha = 0$ almost surely.
(ii) For any $t_0 < t_1 < \cdots < t_N$, the increments $\mathrm{L}_{t_n}^\alpha - \mathrm{L}_{t_{n-1}}^\alpha$ are independent, $n = 1, 2, \ldots, N$.
(iii) The difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ and L_{t-s}^α have the same distribution: $\mathcal{S}\alpha\mathcal{S}((t-s)^{1/\alpha})$ for $s < t$.
(iv) L_t^α has stochastically continuous sample paths, i.e. for any $\delta > 0$ and $s \geq 0$, $\mathbb{P}(|\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha| > \delta) \to 0$ as $t \to s$.

When $\alpha = 2$, we obtain a scaled Brownian motion $\sqrt{2}\mathrm{B}_t$ as a special case so that the difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ follows a Gaussian distribution $\mathcal{N}(0, 2(t-s))$ and L_t^α is almost surely continuous. When $0 < \alpha < 2$, due to the stochastic continuity property, symmetric α-stable Lévy motions can have have a countable number of discontinuities, which are often known as *jumps*. The sample paths are continuous from the right and they have left limits, a property known as càdlàg (Duan, 2015).

Recently, Şimşekli (2017) extended the *overdamped* Langevin dynamics to an SDE driven by L_t^α, given as:[2]

[2] In Şimşekli (2017), (8) does not contain an inverse temperature β, which was later on introduced in Nguyen et al. (2019).

$$\mathrm{d}\mathbf{x}_t = b(\mathbf{x}_{t-}, \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{dL}_t^\alpha, \tag{8}$$

where the drift $b(\mathbf{x}, \alpha) = ((b(\mathbf{x}, \alpha))_i, 1 \leq i \leq d)$ is defined as follows:

$$(b(\mathbf{x}, \alpha))_i = -\mathcal{D}_{x_i}^{\alpha-2}(\phi(\mathbf{x})\partial_{x_i} f(\mathbf{x}))/\phi(\mathbf{x}). \tag{9}$$

Here, $\phi(\mathbf{x}) = \exp(-f(\mathbf{x}))$ and $\mathcal{D}$ denotes the fractional Riesz derivative (Riesz, 1949):

$$\mathcal{D}^\gamma u(x) := \mathcal{F}^{-1}\left\{|\omega|^\gamma (\mathcal{F}(u))(\omega)\right\}(x), \tag{10}$$

where $\mathcal{F}$ denotes the Fourier transform. Briefly, $\mathcal{D}^\gamma$ extends usual differentiation to fractional orders and when $\gamma = 2$ it coincides (up to a sign difference) with the usual second-order derivative $-\mathrm{d}^2 f(x)/\mathrm{d}x^2$.

The important property of the process (8) is that it admits an invariant distribution whose density is proportional to $\exp(-\beta f(\mathbf{x}))$ (Nguyen et al., 2019). It is easy to show that, when $\alpha = 2$, the drift reduces to $b(\mathbf{x}, 2) = -\nabla f(\mathbf{x})$, hence we recover the classical overdamped dynamics:

$$\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2/\beta}\mathrm{dB}_t. \tag{11}$$

Since the fractional Riesz derivative is costly to compute, Şimşekli (2017) proposed an approximation of $b(\mathbf{x}, \alpha)$ based on the alternative definition of $\mathcal{D}$ given in (Ortigueira, 2006), such that:

$$b(\mathbf{x}, \alpha) \approx -c_\alpha \nabla f(\mathbf{x}), \tag{12}$$

where $c_\alpha := \Gamma(\alpha - 1)/\Gamma(\alpha/2)^2$. This approximation essentially results in replacing B_t with L_t^α in (11) in a rather straightforward manner. While avoiding the computational issues originated from the Riesz derivatives, as shown in (Nguyen et al., 2019), this approximation can induce an arbitrary bias in a non-convex optimization context. Besides, the stationary distribution of this approximated dynamics was analytically derived in (Sliusarenko et al., 2013) under the choice of $\alpha = 1$ and $f(x) = x^4/4 - ax^2/2$ for $x \in \mathbb{R}^1$ and $a > 0$. These results show that, in the presence of heavy-tailed perturbations, the drift should be modified, otherwise an inaccurate approximation of the Riesz derivatives can result in an explicit bias, which moves the modes of the distribution away from the modes of f.

From a pure Monte Carlo perspective, Ye & Zhu (2018) extended the fractional overdamped dynamics (8) to higher-order dynamics and proposed the so-called fractional Hamiltonian dynamics (FHD), given as follows:

$$\begin{aligned}
\mathrm{d}\mathbf{x}_t =& \mathcal{D}^{\alpha-2}\{\phi(\mathbf{z}_t)\mathbf{v}_t\}/\phi(\mathbf{z}_t)\mathrm{d}t, \\
\mathrm{d}\mathbf{v}_t =& -\mathcal{D}^{\alpha-2}\{\phi(\mathbf{z}_t)\nabla f(\mathbf{x}_t)\}/\phi(\mathbf{z}_t)\mathrm{d}t \\
& - \gamma\mathcal{D}^{\alpha-2}\{\phi(\mathbf{z}_t)\mathbf{v}_t\}/\phi(\mathbf{z}_t)\mathrm{d}t + \gamma^{1/\alpha}\mathrm{dL}_t^\alpha,
\end{aligned} \tag{13}$$

where $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{v}_t)$, and $\phi(\mathbf{z}) = e^{-f(\mathbf{x}) - \frac{1}{2}\|\mathbf{v}\|^2}$. They showed that the invariant measure of the process has a density proportional to $\phi(\mathbf{z})$, i.e., the Boltzmann-Gibbs measure. Similar to the overdamped case (8), the Riesz derivatives do not admit an analytical form in general. Hence they approximated them by using the same approximation given in (12), which yields the SDE given in (7) (up to a scaling factor). This observation also confirms that the heavy-tailed noise requires an adjustment in the dynamics, otherwise the induced bias might drive the dynamics away from the minima of f (Capała & Dybiec, 2019).

3. Fractional Underdamped Langevin Dynamics

In this section, we develop the fractional underdamped Langevin dynamics (FULD), which is expressed by the following SDE:

$$\begin{aligned}
\mathrm{d}\mathbf{v}_t &= -(\gamma c(\mathbf{v}_{t-}, \alpha) + \nabla f(\mathbf{x}_t))\mathrm{d}t + (\gamma/\beta)^{1/\alpha}\mathrm{dL}_t^\alpha, \\
\mathrm{d}\mathbf{x}_t &= \nabla g(\mathbf{v}_t)\mathrm{d}t,
\end{aligned} \tag{14}$$

where $c : \mathbb{R}^d \times (0, 2] \mapsto \mathbb{R}^d$ is the *drift function* for the velocity and $g : \mathbb{R}^d \mapsto \mathbb{R}$ denotes a general notion of *kinetic energy*. In the next theorem, which is the main theoretical result of this paper, we will identify the relation between these two functions such that the solution process will keep the *generalized* Boltzmann-Gibbs measure, $\exp(-\beta(f(\mathbf{x}) + g(\mathbf{v})))\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}$ invariant. All the proofs are given in the supplementary document.

Theorem 1. *Let $c(\mathbf{v}, \alpha) = ((c(\mathbf{v}, \alpha))_i, 1 \leq i \leq d)$ has the following form:*

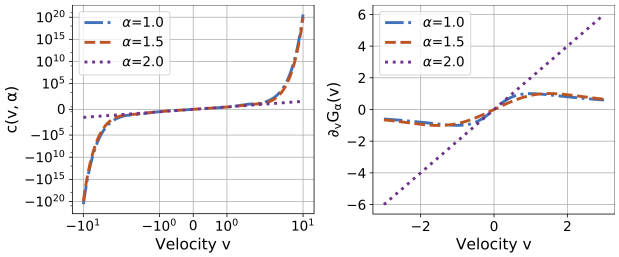
$$(c(\mathbf{v}, \alpha))_i := \frac{\mathcal{D}_{v_i}^{\alpha-2}(\psi(\mathbf{v})\partial_{v_i} g(\mathbf{v}))}{\psi(\mathbf{v})}, \qquad \psi(\mathbf{v}) := e^{-g(\mathbf{v})}. \tag{15}$$

The measure $\pi(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{v}) \propto e^{-\beta(f(\mathbf{x})+g(\mathbf{v}))}\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}$ on $\mathbb{R}^d \times \mathbb{R}^d$ is an invariant probability measure for the Markov process $(\mathbf{x}_t, \mathbf{v}_t)$.

One of the main features of FULD is that the fractional Riesz derivatives only appears in the drift c, which *only* depends on $\mathbf{v}$. This is highly in contrast with FHD (13), where the Riesz derivatives are taken over both $\mathbf{x}$ and $\mathbf{v}$, which is the source of intractability. Moreover, FULD enjoys the freedom to choose different kinetic energy functions $g(\mathbf{v})$. In the sequel, we will investigate two options for g, such that the drift c can be analytically obtained.

3.1. Gaussian kinetic energy

In classical overdamped Langevin dynamics and Hamiltonian dynamics, the default choice of kinetic energy is the Gaussian kinetic energy, which corresponds to taking $g(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|^2$ (Neal, 2010; Livingstone et al., 2019;



(a) (b)

Figure 2. Illustration of one dimensional a) drift function c for the Gaussian kinetic energy, b) ∇G_α for the $\mathcal{S}\alpha\mathcal{S}$ kinetic energy.

Dalalyan & Riou-Durand, 2020). With this choice, the fractional dynamics becomes:

$$\begin{aligned} \mathrm{d}\mathbf{v}_t &= -\gamma c(\mathbf{v}_{t-}, \alpha)\mathrm{d}t - \nabla f(\mathbf{x}_t)\mathrm{d}t + (\gamma/\beta)^{1/\alpha}\mathrm{dL}_t^\alpha, \\ \mathrm{d}\mathbf{x}_t &= \mathbf{v}_t \mathrm{d}t. \end{aligned} \tag{16}$$

In the next result, we will show that in this case, the drift c admits an analytical solution.

Theorem 2. *Let $g(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|^2$. Then, for any $1 \leq i \leq d$,*

$$(c(\mathbf{v}, \alpha))_i = \frac{2^{\frac{\alpha}{2}} v_i}{\sqrt{\pi}} \Gamma\left(\frac{\alpha+1}{2}\right) {}_1F_1\left(\frac{2-\alpha}{2}; \frac{3}{2}; \frac{v_i^2}{2}\right), \tag{17}$$

where Γ is the gamma function and ${}_1F_1$ is the Kummer confluent hypergeometric function. In particular, when $\alpha = 2$, we have $(c(\mathbf{v}, \alpha))_i = v_i$.

We observe that the fractional dynamics (16) strictly extends the underdamped Langevin dynamics (6) as $c(\mathbf{v}, 2) = \mathbf{v}$.

Let us now investigate the form of the new drift c and its implications. In Figure 2(a), we illustrate c for the $d = 1$ dimensional case (note that for $d > 1$, each component of c still behaves like Figure 2(a)). We observe that due to the hypergeometric function ${}_1F_1$, the drift grows exponentially fast with $|v|$ whenever $\alpha < 2$. Semantically, this means that, in order to be able to compensate the large jumps incurred by L_t^α, the drift has to react very strongly and hence prevent v to take large values. To illustrate this behavior, we provide more visual illustrations in the supplementary document.

Even though this aggressive behavior of c can be beneficial for the continuous-time system, it is unfortunately clear that its Euler-Maruyama discretization will not yield a practical algorithm due to the same behavior. Indeed, we would need the function c to be Lipschitz continuous in order to guarantee the algorithmic stability of its discretization (Kloeden & Platen, 1999); however, if we consider the integral form of ${}_1F_1$ (cf. (Abramowitz & Stegun, 1972)), we observe that the function

$$(c(\mathbf{v}, \alpha))_i = \frac{2^{\frac{\alpha}{2}} v_i}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{2-\alpha}{2})} \int_0^1 e^{\frac{v_i^2}{2}t} t^{-\frac{\alpha}{2}} (1-t)^{\frac{\alpha-1}{2}} \mathrm{d}t$$

is clearly not Lipschitz continuous in v_i. Therefore, we conclude that FULD with the Gaussian kinetic energy is mostly of theoretical interest.

3.2. Alpha-stable kinetic energy

The dynamics with the Gaussian kinetic energy requires a very strong drift c mainly because we force the dynamics to make sure that the invariant distribution of $\mathbf{v}$ to be a Gaussian. Since the Gaussian distribution has light-tails, it cannot tolerate samples with large magnitudes, hence requires a large dissipation to make sure $\mathbf{v}$ does not take large values.

In order to avoid such an explosive drift that potentially degrades practicality, next we explore *heavy-tailed* kinetic energies, which would allow the components of $\mathbf{v}$ to take large values, while still making sure that the drift c in (15) admits an analytical form.

In our next result, we show that, when we choose an $\mathcal{S}\alpha\mathcal{S}$ kinetic energy, such that the tail-index α of this kinetic energy matches the one of the driving process L_t^α, the drift c simplifies and becomes the identity function.

Theorem 3. *Let $e^{-g_\alpha(v)}$ be the probability density function of $\mathcal{S}\alpha\mathcal{S}(\frac{1}{\alpha^{1/\alpha}})$. Choose $\psi(\mathbf{v}) = e^{-G_\alpha(\mathbf{v})}$ in (15), where $G_\alpha(\mathbf{v}) = \sum_{i=1}^d g_\alpha(v_i)$ for any $\mathbf{v} = (v_1, \ldots, v_d)$. Then,*

$$(c(\mathbf{v}, \alpha))_i = v_i, \qquad 1 \leq i \leq d. \tag{18}$$

This result hints that, perhaps $G_\alpha(v)$ is the natural choice of kinetic energy for the systems driven by L_t^α.

It now follows from Theorem 3 that the FULD with α-stable kinetic energy reduces to the following SDE:

$$\begin{aligned} \mathrm{d}\mathbf{v}_t &= -\gamma \mathbf{v}_{t-}\mathrm{d}t - \nabla f(\mathbf{x}_t)\mathrm{d}t + (\gamma/\beta)^{1/\alpha}\mathrm{dL}_t^\alpha, \\ \mathrm{d}\mathbf{x}_t &= \nabla G_\alpha(\mathbf{v}_t)\mathrm{d}t. \end{aligned} \tag{19}$$

It can be easily verified that $\nabla G_\alpha(\mathbf{v}_t) = \mathbf{v}_t$ for $\alpha = 2$, as $g_2(v) = \frac{1}{2}\log 2\pi + \frac{1}{2}v^2$, hence, the SDE (19) also reduces to the classical underdamped Langevin dynamics (6).

While this choice of g results in an analytically available c, unfortunately the function ∇G_α itself admits a closed-form analytical formula only when $\alpha = 1$ or $\alpha = 2$, due to the properties of the $\mathcal{S}\alpha\mathcal{S}$ densities. Nevertheless, as ∇G_α is based on one-dimensional $\mathcal{S}\alpha\mathcal{S}$ densities, it can be very accurately computed by using the recent methods developed in (Ament & O'Neil, 2018). On the other hand, in the next section, we will show that ∇G_α is Lipschitz continuous for all $\alpha \in (0, 2]$, which implies that under standard regularity conditions on f, the Boltzmann-Gibbs measure is the unique invariant measure of (19).

We visually inspect the behavior of ∇G_α in Figure 2(b) for dimension one. We observe that, as soon as $\alpha < 2$, ∇G_α

takes a very smooth form. Besides, for small $|v|$ the function behaves like a linear function and when $|v|$ goes to infinity, it vanishes. This behavior can be interpreted as follows: since $\mathbf{v}$ can take larger values due to the heavy tails of the kinetic energy, in order to be able target the correct distribution, the dynamics compensates the potential bursts in $\mathbf{v}$ by passing it through the asymptotically vanishing ∇G_α.

3.3. Euler discretization and weak convergence analysis

As visually hinted in Figure 2(b), the function ∇G_α has strong regularity, which makes (19) to be potentially beneficial for practical implementations. Indeed, it is easy to verify that ∇G_α is Lipschitz continuous for $\alpha = 1$ and 2, and in our next result, we show that this observation is true for any admissible α, which is a desired property when discretizing continuous-time dynamics.

Proposition 1. *For $0 < \alpha \leq 2$, the map $v \mapsto g'_\alpha(v)$ is Lipschitz continuous, hence $\mathbf{v} \mapsto \nabla G_\alpha(\mathbf{v})$ is also Lipschitz continuous.*

Accordingly we consider the following Euler-Maruyama discretization for (19):

$$\begin{aligned} \mathbf{v}^{k+1} &= \tilde{\gamma}_k \mathbf{v}^k - \eta_k \nabla f(\mathbf{x}^k) + (\eta_k \gamma/\beta)^{1/\alpha} \mathbf{s}^{k+1}, \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + \eta_k \nabla G_\alpha(\mathbf{v}^{k+1}), \end{aligned} \tag{20}$$

where $\tilde{\gamma}_k = 1 - \gamma\eta_k$, $\mathbf{s}^k$ is a random vector whose components are independently $\mathcal{S}\alpha\mathcal{S}(1)$ distributed, and $(\eta_k)_{k\in\mathbb{N}_+}$ is a sequence of step-sizes.

In this section, we analyze the weak convergence of the ergodic averages computed by using (20). Given a test function h, consider its expectation with respect to the target measure π, i.e. $\pi(h) := \mathbb{E}_{X\sim\pi}[h(X)] = \int h(\mathbf{x})\pi(\mathrm{d}\mathbf{x})$ with $\pi(\mathrm{d}\mathbf{x}) \propto \exp(-\beta f(\mathbf{x}))\mathrm{d}\mathbf{x}$. We will discuss next how this expectation can be approximated through the sample averages

$$\bar{\pi}_K(h) := (1/S_K) \sum_{k=1}^{K} \eta_k h(\mathbf{x}^k), \tag{21}$$

where $S_K := \sum_{k=1}^{K} \eta_k$ is the cumulative sum of the step-size sequence.

We note that Langevin-based algorithms have been used in the literature to obtain global convergence guarantees for non-convex optimization, see e.g. (Raginsky et al., 2017; Xu et al., 2018; Gao et al., 2018b; Zou et al., 2019; Nguyen et al., 2019). In particular, Nguyen et al. (2019) used an overdamped fractional Langevin dynamics for non-convex optimizations. The proposed model in our paper can also be used to study the non-convex optimizations and we expect that our underdamped dynamics may have improved theoretical guarantees compared to (Nguyen et al., 2019).

We now present the assumptions that imply our results.

Assumption 1. *The step-size sequence $\{\eta_k\}$ is non-increasing and satisfies $\lim_{k\to\infty} \eta_k = 0$ and $\lim_{K\to\infty} S_K = \infty$.*

Assumption 2. *Let $V : \mathbb{R}^{2d} \to \mathbb{R}_+$ be a twice continuously differentiable function, satisfying $\lim_{\|\mathbf{z}\|\to\infty} V(\mathbf{z}) = \infty$, $\|\nabla V\| \leq C\sqrt{V}$ for some $C > 0$ and has a bounded Hessian $\nabla^2 V$. Given $p \in (0, \frac{1}{2}]$, there exists $a \in (1 - \frac{p}{2}, 1]$, $\beta_1 \in \mathbb{R}$, $\beta_2 > 0$ such that $\|b\|^2 \leq CV^a$ and $\langle \nabla V, b\rangle \leq \beta_1 - \beta_2 V^a$ where $b(\mathbf{v}, x) = (-\gamma\mathbf{v} - \nabla f(\mathbf{x}), \nabla G_\alpha(\mathbf{v}))$ is the drift of the $(\mathbf{v}_t, \mathbf{x}_t)$ process defined in (19).*

These are common assumptions ensuring that the SDE is simulated with infinite time-horizon and the process is not explosive (Panloup, 2008; Şimşekli, 2017). We can now establish the weak convergence of (21) and present it as a corollary to Theorem 1, Proposition 1, and (Panloup, 2008) (Theorem 2).

Corollary 1. *Assume that the gradient ∇f is Lipschitz continuous and has linear growth i.e., there exists $C > 0$ such that $\|\nabla f(\mathbf{x})\| \leq C(1 + \|\mathbf{x}\|)$ for all $\mathbf{x}$. Furthermore, assume that Assumptions 1 and 2 hold for some $p \in (0, 1/2]$. If the test function $h = o(V^{\frac{p}{2}+a-1})$ then*

$$\bar{\pi}_K(h) \to \pi(h) \quad \textit{almost surely as } K \to \infty.$$

3.4. Connections to existing approaches

We now point out interesting algorithmic connections between (20) and two methods that are commonly used in practice. We first roll back our initial hypothesis that the gradient noise is $\mathcal{S}\alpha\mathcal{S}$ distributed, i.e., $\nabla \tilde{f}_k(\mathbf{x}) = \nabla f(\mathbf{x}) + (\eta_k\gamma/\beta)^{1/\alpha}\mathbf{s}^k$, and modify (20) as follows:

$$\begin{aligned} \mathbf{v}^{k+1} &= \tilde{\gamma}_k \mathbf{v}^k - \eta_k \nabla \tilde{f}_{k+1}(\mathbf{x}^k), \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + \eta_k \nabla G_\alpha(\mathbf{v}^{k+1}). \end{aligned} \tag{22}$$

As a special case when $\tilde{\gamma}_k = 0$, we obtain a stochastic gradient descent-type recursion:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \eta_k \nabla G_\alpha(-\eta_k \nabla \tilde{f}_{k+1}(\mathbf{x}^k)). \tag{23}$$

Let us now consider *gradient-clipping*, a heuristic approach for eliminating the problem of 'exploding gradients', which often appear in training neural networks (Pascanu et al., 2013; Zhang et al., 2019a). Very recently, Zhang et al. (2019b) empirically illustrated that such explosions stem from heavy-tailed gradients and formally proved that gradient clipping indeed improves convergence rates under heavy-tailed perturbations. We notice that, the behavior of (22) is reminiscent of gradient clipping: due to the vanishing behavior of ∇G_α for $\alpha < 2$, as the components of $\mathbf{v}^k$ gets larger in magnitude, the update applied on $\mathbf{x}^k$ gets smaller. The behavior becomes more prominent in (23). On the other hand, (22) is more aggressive in the sense that the

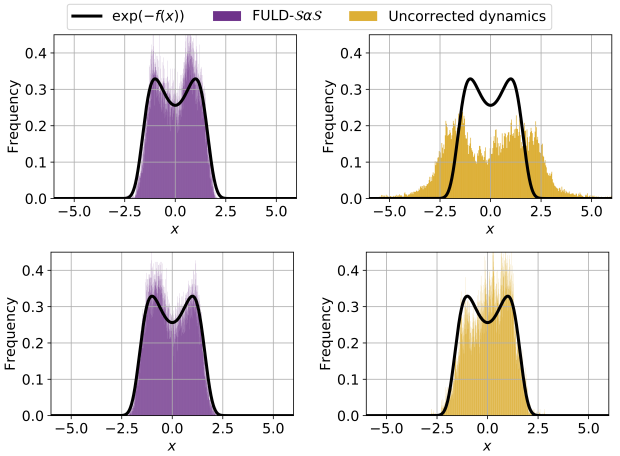


Figure 3. Estimated invariant measures for the quartic potential: top $\alpha = 1$, bottom $\alpha = 1.9$.

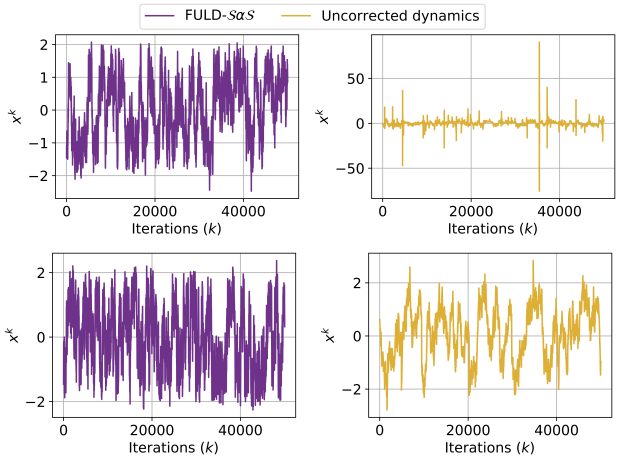


Figure 4. Illustration of the iterates for the quartic potential: top $\alpha = 1$, bottom $\alpha = 1.9$.

updates can get arbitrarily small as the value of α decreases as opposed to being 'clipped' with a threshold.

The second connection is with the natural gradient descent algorithm, where the stochastic gradients are preconditioned with the inverse Fisher information matrix (FIM) (Amari, 1998). Here FIM is defined as $\mathbb{E}[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top]$, where the expectation is taken over the data. Notice that when $\alpha = 1$ (i.e., Cauchy distribution), we have the following form: $\nabla G_1(\mathbf{v}) = \left(\frac{2v_1}{v_1^2+1}, \dots, \frac{2v_d}{v_d^2+1}\right)$. Therefore, we observe that, in (23), $\nabla G_1(\nabla \tilde{f}_k(\mathbf{x}))$ can be equivalently written as $\mathbf{M}_k(\mathbf{x})^{-1} \nabla \tilde{f}_k(\mathbf{x})$, where $\mathbf{M}_k(\mathbf{x})$ is a diagonal matrix with entries $m_{ii} = ((\nabla \tilde{f}_k(\mathbf{x}))_i^2 + 1)/2$. Therefore, we can see $\mathbf{M}_k$ as an estimator of the diagonal part of FIM, as they will be in the same order when $|(\nabla \tilde{f}_k(\mathbf{x}))_i|$ is large. Besides, (22) then appears as its momentum extension. However, $\mathbf{M}_k$ will be biased mainly due to the fact that FIM is the average of the squared gradients, whereas $\mathbf{M}_k$ is based on the square of the average gradients. This connection is rather surprising, since a seemingly unrelated, differential geometric approach turns out to have strong algorithmic similarities with a method that naturally arises when the gradient noise is Cauchy distributed.

4. Numerical Study

In this section, we will illustrate our theory on several experiments which are conducted in both synthetic and real-data settings[3]. We note that, as expected, FULD with Gaussian kinetic energy did not yield a numerically stable discretization due to the explosive behavior of c. Hence, in this section, we only focus on FULD with $\mathcal{S}\alpha\mathcal{S}$ kinetic energy and from now on we will simply refer to FULD with $\mathcal{S}\alpha\mathcal{S}$ kinetic energy as FULD.

[3]We provide our implementation in https://github.com/umutsimsekli/fuld.

4.1. Synthetic setting

We first consider a one-dimensional synthetic setting, similar to the one considered in (Capała & Dybiec, 2019). We consider a quartic potential function with a quadratic component, $f(x) = x^4/4 - x^2/2$. We then simulate the 'uncorrected dynamics' (UD) given in (7) and FULD (19) by using the Euler-Maruyama discretization to compare their behavior for different α. For $\alpha \notin \{1, 2\}$, we used the software given in (Ament & O'Neil, 2018) for computing ∇G_α.

Figure 3 illustrates the distribution of the samples generated by simulating the two dynamics. In this setup, we set $\beta = 1$, $\eta = 0.01$, $\gamma = 10$ with number of iterations $K = 50000$. We observe that, for $\alpha = 1.9$, FULD very accurately captures the form of the distribution, whereas UD exhibits a visible bias and the shape of its resulting distribution is slightly distorted. Nevertheless, since the perturbations are close to a Gaussian in this case (i.e., α is close to 2), the difference is not substantial and can be tolerable in an optimization context. However, this behavior becomes much more emphasized when we use a heavier-tailed driving process: when $\alpha = 1$, we observe that the target distribution of UD becomes distant from the Gibbs measure $\exp(-f(x))$, and more importantly its modes no longer match the minima of f; agreeing with the observations presented in (Capała & Dybiec, 2019)[4]. On the other hand, thanks to the correction brought by ∇G_α, FULD still captures the target distribution very accurately, even when the driving force is Cauchy.

On the other hand, in our experiments we observed that, for small values of α, UD can quickly become numerically unstable and even diverge for slightly larger step-sizes, whereas this problem never occurred for FULD. This out-

[4]We note that the overdamped dynamics with the uncorrected drift exhibits a similar behavior to the one of the uncorrected underdamped dynamics with sufficiently large γ.

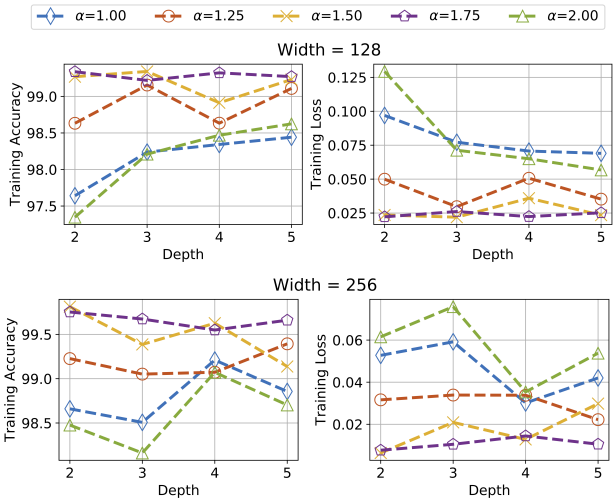


Figure 5. Neural network results on MNIST (training).

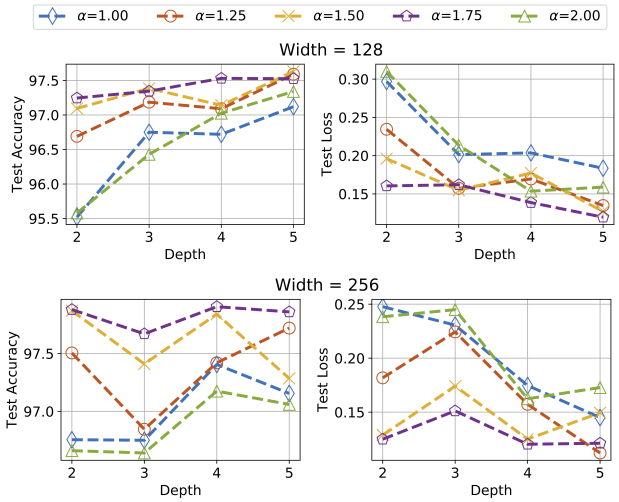


Figure 6. Neural network results on MNIST (test).

come also stems from the fact that UD does not have any mechanism to compensate the potential large updates originating from the heavy-tailed perturbations. To illustrate this observation more clearly, in Figure 4 we illustrate the iterates $(\mathbf{x}^k)_{k=1}^K$ which were used for producing Figure 3. We observe that, while the iterates of UD are well-behaved for $\alpha = 1.9$, the magnitude range of the iterates gets quite large when α is set to 1. On the other hand, for both values of α, FULD iterates are always kept in a reasonable range, thanks to the clipping-like effect of ∇G_α.

4.2. Neural networks

In our next set of experiments, we evaluate our theory on neural networks. In particular, we apply the iterative scheme given in (22) as an optimization algorithm for training neural networks, and compare its behavior with classical SGDm defined in (2). In this setting, we do not add any explicit noise, all the stochasticity comes from the potentially heavy-tailed stochastic gradient noise (3) under the assumption that the noise can be well-modeled by using an $\mathcal{S}\alpha\mathcal{S}$ vector (see Section 3.4 for the explicit assumption).

We consider a fully-connected network for a classification task on the MNIST and CIFAR10 datasets, with different depths (i.e. number of layers) and widths (i.e. number of neurons per layer). For each depth-width pair, we train two neural networks by using SGDm (2) and our modified version (22), and compare their final train/test accuracies and loss values. We use the conventional train-test split of the datasets: for MNIST we have 60K training and 10K test samples, and for CIFAR10 these numbers are 50K and 10K, respectively. We use the cross entropy loss (also referred to as the 'negative-log-likelihood').

We note that the modified scheme (22) reduces to (2) when $\alpha = 2$, since $\nabla G_2(\mathbf{v}) = \mathbf{v}$. Hence in this section, we will refer to SGDm as the special case of (22) with $\alpha = 2$. On the other hand, in these experiments, directly computing ∇G_α becomes impractical for $\alpha \notin \{1, 2\}$, since the algorithms given in (Ament & O'Neil, 2018) become prohibitively slow with the increased dimension d. However, since ∇G_α is based on the derivatives of the *one-dimensional* $\mathcal{S}\alpha\mathcal{S}$ densities $g_\alpha(v)$ (see Theorem 3), for $\alpha \in (1, 2)$, we first precomputed the values of $g_\alpha(v)$ over a fine grid of $v \in [-100, 100]$; then, during the SGDm recursion, we approximated ∇G_α by linearly interpolating the values of g_α that are precomputed over this grid. We expect that, if the stochastic gradient noise can be well-approximated by using an $\mathcal{S}\alpha\mathcal{S}$ distribution, then the modified dynamics should exhibit an improved performance since it would eliminate the potential bias brought by the heavy-tailed noise.

In these experiments, we set $\eta = 0.1$, $\gamma = 0.1$ for MNIST, and $\gamma = 0.9$ for CIFAR10. We run the algorithms for $K = 10000$ iterations [5]. We measure the accuracy and the loss at every 100th iteration and we report the average of the last two measurements. Figures 5 and 6 show the results obtained on the MNIST dataset. We observe that, in most of the cases, setting $\alpha = 1.75$ yields a better performance in terms both training and testing accuracies/losses. This difference becomes more visible when the width is set to 256: the accuracy difference between the algorithms reaches $\approx 2\%$. We obtain a similar result on the CIFAR10 dataset, as illustrated in Figures 7 and 8. In most of the cases $\alpha = 1.75$ performs better, with the maximum accuracy difference being $\approx 4.5\%$, implying the gradient noise can be approximated by an $\mathcal{S}\alpha\mathcal{S}$ random variable.

[5] Since the scale of the gradient noise is proportional to $(\gamma/\beta)^{\frac{1}{\alpha}}$ (see (20)), in this setup, a fixed γ implicitly determines β.

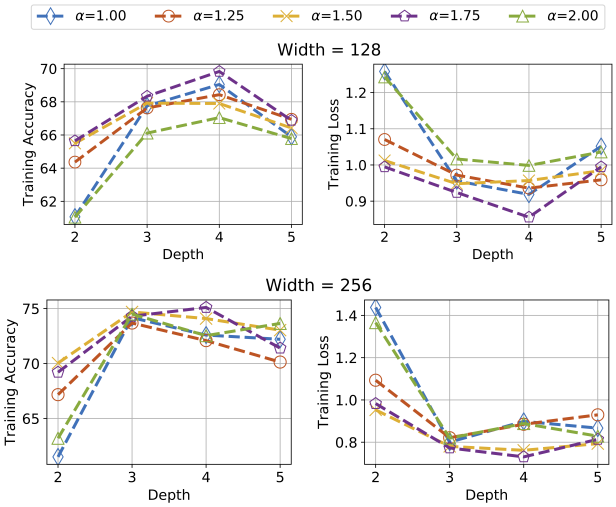


Figure 7. Neural network results on CIFAR10 (training).

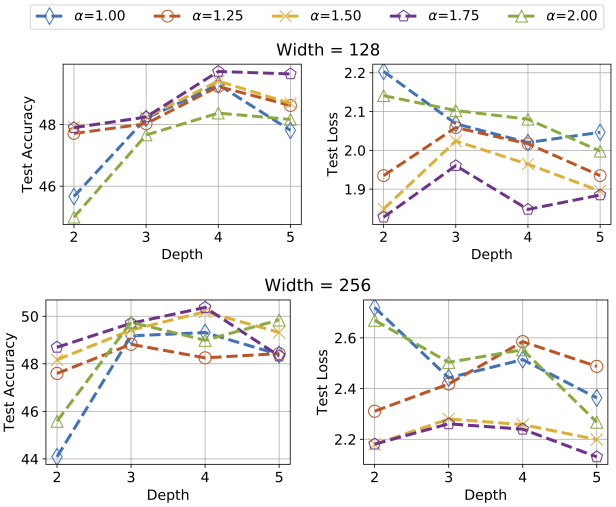


Figure 8. Neural network results on CIFAR10 (test).

We observed a similar behavior when the width was set to 64. However, when we set the width to 32 we did not perceive a significant difference in terms of the performance of the algorithms. On the other hand, when the width was set to 512, $\alpha = 2$ resulted in a slightly better performance, which would be an indication that the Gaussian approximation is closer. The corresponding figures are provided in the supplementary document.

5. Conclusion and Future Directions

We considered the continuous-time variant of SGDm, known as the underdamped Langevin dynamics (ULD), and developed theory for the case where the gradient noise can be well-approximated by a heavy-tailed α-stable random vector. As opposed to naïvely replacing the driving stochastic force in ULD, which correspondonds to running SGDm with heavy-tailed gradient noise, the dynamics that we developed exactly target the Boltzmann-Gibbs distribution, and hence do not introduce an implicit bias. We further established the weak convergence of the Euler-Maruyama discretization and illustrated interesting connections between the discretized algorithm and existing approaches commonly used in practice. We supported our theory with experiments on a synthetic setting and fully connected neural networks.

Our framework opens up interesting future directions. Our current modeling strategy requires a state-independent, isotropic noise assumption, which would not accurately reflect the reality. While anisotropic noise can be incorporated to our framework by using the approach of Ye & Zhu (2018), state-dependent noise introduces challenging technical difficulties. Similarly, it has been illustrated that the tail-index α can depend on the state and different components of the noise can have a different α (Şimşekli et al., 2019a). Incorporating such state dependencies would be an important direction of future research. Finally, it has been shown that the heavy-tailed perturbations yield shorter escape times (Nguyen et al., 2019) in the overdamped dynamics, and extending such results to the underdamped case is still an open problem.

Acknowledgments

We thank Jingzhao Zhang for fruitful discussions. The contribution of Umut Şimşekli to this work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014) project, and by the industrial chair Data science & Artificial Intelligence from Télécom Paris. Lingjiong Zhu is grateful to the support from Simons Foundation Collaboration Grant. Mert Gürbüzbalaban acknowledges support from the grants NSF DMS-1723085 and NSF CCF-1814888.

References

Abramowitz, M. and Stegun, I. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover, New York, 1972.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Ament, S. and O'Neil, M. Accurate and efficient numerical calculation of stable densities via optimized quadrature and asymptotics. *Statistics and Computing*, 28:171–185, 2018.

Bertoin, J. *Lévy Processes*. Cambridge University Press, Cambridge, UK, 1996.

Betancourt, M., Byrne, S., Livingstone, S., and Girolami, M. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, 2017.

Capała, K. and Dybiec, B. Stationary states for underdamped anharmonic oscillators driven by Cauchy noise. *arXiv preprint arXiv:1905.12078*, 2019.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Şimşekli, U. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov Chain Monte Carlo. In *International Conference on Machine Learning*, pp. 3200–3209, 2017.

Şimşekli, U., Yıldız, Ç., Nguyen, T. H., Richard, G., and Cemgil, A. T. Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. *arXiv preprint arXiv:1806.02617*, 2018.

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019a.

Şimşekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837, 2019b.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. *arXiv preprint arXiv:2006.09313*, 2020.

Dalalyan, A. S. and Riou-Durand, L. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Duan, J. *An Introduction to Stochastic Dynamics*. Cambridge University Press, New York, 2015.

Dubkov, A. A., Spagnolo, B., and Uchaikin, V. V. Lévy flight superdiffusion: An introduction. *International Journal of Bifurcation and Chaos*, 18(09):2649–2672, 2008.

Eliazar, I. and Klafter, J. Lévy-driven Langevin systems: Targeted stochasticity. *Journal of Statistical Physics*, 111 (3-4):739–768, 2003.

Fischer, H. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer Science & Business Media, New York, 2011.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Global convergence of Stochastic Gradient Hamiltonian Monte Carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration. *arXiv:1809.04618*, 2018a.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. *arXiv:1812.07725*, 2018b.

Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. *arXiv preprint arXiv:2006.04740*, 2020.

Hérau, F. and Nier, F. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Archive for Rational Mechanics and Analysis*, 171(2):151–218, 2004.

Hodgkinson, L. and Mahoney, M. W. Multiplicative noise and heavy tails in stochastic optimization. *arXiv preprint arXiv:2006.06293*, 2020.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Science and Applications*, 4(1): 3–32, 2019.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

Kloeden, P. E. and Platen, E. *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer Verlag, Berlin, 1999.

Kuruoglu, E. E. *Signal Processing in α-Stable Noise Environments: A Least ℓ_p-Norm Approach*. PhD Thesis, University of Cambridge, 1999.

Lévy, P. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1937.

Livingstone, S., Faulkner, M. F., and Roberts, G. O. Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. *Biometrika*, 106(2):303–319, 2019.

Lu, X., Perrone, V., Hasenclever, L., Teh, Y. W., and Vollmer, S. J. Relativistic Monte Carlo. In *Artificial Intelligence and Statistics*, pp. 1236–1245, 2017.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

Mandelbrot, B. B. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*. Springer Science & Business Media, New York, 1997.

Mandt, S., Hoffman, M., and Blei, D. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 354–363, 2016.

Martin, C. H. and Mahoney, M. W. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. *arXiv preprint arXiv:1901.08278*, 2019.

Neal, R. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds.), 2010.

Nguyen, T. H., Şimşekli, U., Gurbuzbalaban, M., and Richard, G. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, pp. 273–283, 2019.

Nguyen, T. H., Şimşekli, U., and Richard, G. Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization. In *International Conference on Machine Learning*, pp. 4810–4819, 2019.

Ortigueira, M. D. Riesz potential operators and inverses via fractional centred derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-Gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

Panloup, F. Recursive computation of the invariant measure of a stochastic differential equation driven by a Lévy process. *Annals of Applied Probability*, 18(2):379–426, 2008.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318, 2013.

Pavliotis, G. A. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer Science & Business Media, New York, 2014.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703, 2017.

Riesz, M. L'intégrale de Riemann-Liouville et le problème de Cauchy. *Acta Mathematica*, 81(1):1–222, 1949.

Sliusarenko, O. Y., Surkov, D., Gonchar, V. Y., and Chechkin, A. V. Stationary states in bistable system driven by Lévy noise. *The European Physical Journal Special Topics*, 216(1):133–138, 2013.

Smith, S. L., Kindermans, P., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pp. 1139–1147, 2013.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3122–3133, 2018.

Ye, N. and Zhu, Z. Stochastic fractional Hamiltonian Monte Carlo. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Analysis of gradient clipping and adaptive scaling with a relaxed smoothness condition. *arXiv preprint arXiv:1905.11881*, 2019a.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why ADAM beats SGD for attention models. *arXiv preprint arXiv:1912.03194*, 2019b.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pp. 7654–7663, 2019.

Zou, D., Xu, P., and Gu, Q. Stochastic variance-reduced Hamilton Monte Carlo methods. In *International Conference on Machine Learning*, pp. 6028–6037, 2018.

Zou, D., Xu, P., and Gu, Q. Stochastic gradient Hamiltonian Monte Carlo methods with recursive variance reduction. In *Advances in Neural Information Processing Systems*, volume 32, 2019.