Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization

Thanh Huy Nguyen, Umut Şimşekli, Gael Richard

Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization

Thanh Huy Nguyen [1] **Umut Şimşekli** [1] **Gaël Richard** [1]

Abstract

Recent studies on diffusion-based sampling methods have shown that Langevin Monte Carlo (LMC) algorithms can be beneficial for non-convex optimization, and rigorous theoretical guarantees have been proven for both asymptotic and finite-time regimes. Algorithmically, LMC-based algorithms resemble the well-known gradient descent (GD) algorithm, where the GD recursion is perturbed by an additive Gaussian noise whose variance has a particular form. Fractional Langevin Monte Carlo (FLMC) is a recently proposed extension of LMC, where the Gaussian noise is replaced by a heavy-tailed α-stable noise. As opposed to its Gaussian counterpart, these heavy-tailed perturbations can incur large jumps and it has been empirically demonstrated that the choice of α-stable noise can provide several advantages in modern machine learning problems, both in optimization and sampling contexts. However, as opposed to LMC, only asymptotic convergence properties of FLMC have been yet established. In this study, we analyze the non-asymptotic behavior of FLMC for non-convex optimization and prove finite-time bounds for its expected suboptimality. Our results show that the weak-error of FLMC increases faster than LMC, which suggests using smaller step-sizes in FLMC. We finally extend our results to the case where the exact gradients are replaced by stochastic gradients and show that similar results hold in this setting as well.

[1]LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France. Correspondence to: Thanh Huy Nguyen <thanh.nguyen@telecom-paristech.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1. Introduction

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 & 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):

$$\mathrm{d}X(t) = -\nabla f(X(t))\mathrm{d}t + \sqrt{2/\beta}\,\mathrm{dB}(t), \;\; t \geq 0, \quad (1)$$

where $X(t) \in \mathbb{R}^d$, f is a smooth function which is often non-convex, $\beta \in \mathbb{R}_+$ is called the 'inverse temperature' parameter, and $\mathrm{B}(t)$ is the standard Brownian motion in $\mathbb{R}^d$.

Under some regularity conditions on f, one can show that the Markov process $(X_t)_{t\geq 0}$, i.e. the solution of the SDE (1), is ergodic with its unique invariant measure π, whose density is proportional to $\exp(-\beta f(x))$ (Roberts & Stramer, 2002). An important feature of this measure is that, when β goes to infinity, its density concentrates around the global minimum $x^\star \triangleq \arg\min_{x\in\mathbb{R}^d} f(x)$ (Hwang, 1980; Gelfand & Mitter, 1991). This property implies that, if we could simulate (1) for large enough β and t, the simulated state $X(t)$ would be close to $x^\star$.

This connection between diffusions and optimization, motivates simulating (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 & Stramer, 2002):

$$W_{\mathrm{ULA}}^{k+1} = W_{\mathrm{ULA}}^{k} - \eta\nabla f(W_{\mathrm{ULA}}^{k}) + \sqrt{\frac{2\eta}{\beta}}\Delta B_{k+1}, \quad (2)$$

where $k \in \mathbb{N}_+$ denotes the iterations, η denotes the step-size, and $(\Delta B_n)_n$ 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 & Stramer, 2002; Lamberton & Pages, 2003; Durmus & 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 ΔB_k in (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., 2017) 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).

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 1) (Şimşekli, 2017; Ye & 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 (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 (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 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):

$$W_{\text{FLA}}^{k+1} = W_{\text{FLA}}^{k} - \eta c_\alpha \nabla f(W_{\text{FLA}}^{k}) + \left(\frac{\eta}{\beta}\right)^{\frac{1}{\alpha}} \Delta L_{k+1}^{\alpha}, \quad (3)$$

where $\alpha \in (1, 2]$ is called the characteristic index, c_α is a known constant, and $\{\Delta L_k^\alpha\}_{k \in \mathbb{N}_+}$ is a sequence of α-stable distributed random variables. As we will detail in Section 2, FLA coincides with ULA when $\alpha = 2$. Recently, Ye & Zhu (2018) extended FLA to Hamiltonian dynamics. The experimental results in (Şimşekli, 2017) and (Ye & Zhu, 2018) showed that the use of the heavy-tailed increments can provide advantages in multi-modal settings, robustness to algorithm parameters. Ye & Zhu (2018) further illustrated that in an optimization context their algorithm achieves better generalization in deep neural networks. In another recent study, Şimşekli et al. (2019) illustrated that FLA can also be used as a proxy for understanding the dynamics of stochastic gradient descent in deep learning.

Even though asymptotic convergence properties of FLMC were established for decreasing step-sizes in (Şimşekli, 2017; Panloup, 2008), these results do not explain the behavior of the 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.

1.1. Overview of the main result

In this study, we analyze the non-asymptotic behavior of FLA for non-convex optimization. In particular, we analyze the expected suboptimality $\mathbb{E}[f(W_{\text{FLA}}^k) - f^\star]$, where $f^\star \triangleq f(x^\star)$. As we will describe in detail in Section 4, 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 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 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 $\mathbb{E}[f(W_{\text{FLA}}^k) - f^\star]$ is bounded by a sum of four terms, sum-

[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 & Teh, 2011), which is obtained by replacing the gradients in (2) with stochastic gradients.

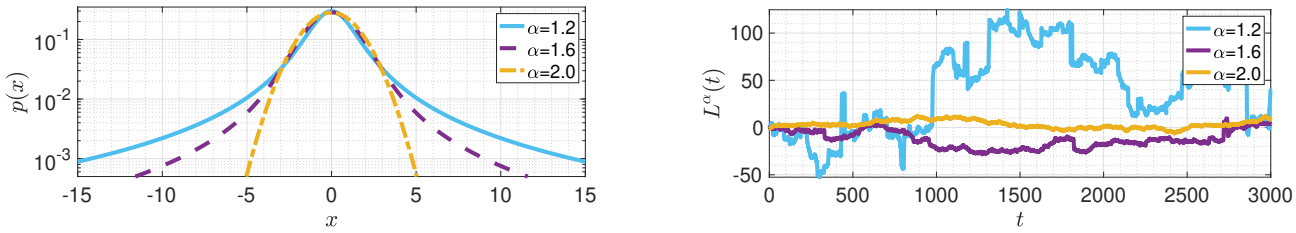

Figure 1. Illustration of the density function of the symmetric α-stable ($\mathcal{S}\alpha\mathcal{S}$) distribution (left) and the α-stable Lévy motion (right). As α gets smaller, $\mathcal{S}\alpha\mathcal{S}$ becomes heavier-tailed and consequently, $L^\alpha(t)$ incurs larger jumps.

marized as follows:

$$\mathbb{E}[f(W^k_{\text{FLA}}) - f^\star] \le \mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4,$$

where

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

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 $\frac{1}{q} \le \gamma + \frac{\gamma}{q}$ 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 $\text{Poly}(\ldots)$ 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 6, 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 FLA has a worse dependency on k and η, as compared to ULA. This suggests that one should use smaller step-sizes in FLA.

2. Technical Background and Preliminaries

2.1. Notations and basic definitions

In this section, we will define the basic quantities that will be used throughout the paper. 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 $\lambda \ge 1$, 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).

From now on, we will denote W^k_{FLA} as W^k for notational simplicity. All the proofs are given in the supplementary document.

2.2. α-Stable Distributions and α-Stable Lévy Motion

Definition 2 (Symmetric α-stable random variables)**.** *The α-stable distribution appears as the limiting distribution in the generalized CLT (Samorodnitsky & 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 $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$.

The parameter α is called the *characteristic index* or the *tail index*, since it determines the tail behavior of the distribution. Perhaps the most important special case of symmetric α-stable distributions is the Gaussian distribution: $\mathcal{S}\alpha\mathcal{S}(\sigma) = \mathcal{N}(0, 2\sigma^2)$ when $\alpha = 2$. As we decrease α, the distribution becomes *heavier-tailed*. Moreover, when $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$, the moment $\mathbb{E}[|X|^p]$ is finite if and only if $p < \alpha$. This implies that the distribution has infinite variance (i.e. the variance diverges) whenever $\alpha \ne 2$. It is easy to draw random samples from $\mathcal{S}\alpha\mathcal{S}$ by using (Chambers et al., 1976).

Definition 3 (Symmetric α-stable Lévy motion)**.** *A scalar symmetric α-stable Lévy motion $L^\alpha(t)$, with $0 < \alpha \le 2$, 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* $t \to s$.

We illustrate $\mathcal{S}\alpha\mathcal{S}$ and $L^\alpha(t)$ in Figure 1. In the rest of the paper, $L^\alpha(t)$ will denote a d-dimensional Lévy process whose components are independent scalar symmetric α-stable Lévy motions as defined in Definition 3.

2.3. Fractional Langevin Monte Carlo

The FLMC framework is based on a Lévy-driven SDE, that is defined as follows:

$$\mathrm{d}X(t) = \Psi(X(t-), \alpha)\mathrm{d}t + (1/\beta)^{1/\alpha}\mathrm{d}L^\alpha(t) \tag{4}$$

where $X(t-)$ denotes the *left limit* of the process at time t, $L^\alpha(t)$ denotes the d-dimensional Lévy motion as described in Section 2.2. FLMC is built up on the following result:

Theorem 1 (Şimşekli (2017)). *Consider the SDE* (4) *in the case* $d = 1$, $\beta = 1$, *and* $\alpha \in (1, 2]$, *where the drift* Ψ *is defined as follows:*

$$\Psi(x, \alpha) \triangleq -\frac{\mathcal{D}^{\alpha-2}\left(\phi(x)\frac{\partial f(x)}{\partial x}\right)}{\phi(x)}. \tag{5}$$

where $\mathcal{D}$ *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* $\hat{u} \triangleq \mathcal{F}(u)$. *Then,* π *is an invariant measure of the Markov process* $(X(t))_{t\geq 0}$ *that is a solution of the SDE given by* (4).

This theorem states that if the drift (5) can be computed, then the sample paths of (4) can be considered as samples drawn from π. However, computing (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 & 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 $\phi(x) \triangleq \exp(-\beta f(x))$, $c_\alpha \triangleq \Gamma(\alpha-1)/\Gamma(\alpha/2)^2$ and Γ denotes the Gamma function. With this choice of approximation, in the d-dimensional case we obtain FLA, as given in (3). We can observe that, when $\alpha = 2$, (4) becomes the Langevin equation (1) and FLA becomes ULA.

3. Assumptions and the Main Result

We start by defining three different stochastic processes $X_1(t)$, $X_2(t)$, and $X_3(t)$, which will be the main constructs in our analysis. We first informally define these processes as follows: X_2 is a continuous-time process that interpolates W^k in time and it will let us avoid dealing with the discrete-time process W^k directly. X_1 is the limiting process of X_2 when the step-size goes to zero. Finally, X_3 is a process whose law converges to the Gibbs measure π.

In our approach, we will first relate X_2 to its limiting process X_1. Since it is more challenging to relate X_1 to $x^\star$, we will then relate X_1 to X_3, and 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:

$$\begin{aligned}\mathbb{E}f(W^k) - f^* =& \mathbb{E}f(X_2(k\eta)) - \mathbb{E}f(X_1(k\eta)) \\ &+ \mathbb{E}f(X_1(k\eta)) - \mathbb{E}f(X_3(k\eta)) \\ &+ \mathbb{E}f(X_3(k\eta)) - \mathbb{E}f(\hat{W}) \\ &+ \mathbb{E}f(\hat{W}) - f^*, \end{aligned} \tag{6}$$

where $X_i(k\eta)$ with $i = 1, 2, 3$ denotes the state reached by the three stochastic processes at time $k\eta$, and $\hat{W}$ is a random variable drawn from π. We will now formally define the processes X_1, X_2, and X_3.

The first SDE is the continuous-time limit of the FLA algorithm given in (3) and defined as follows for $t \geq 0$:

$$\mathrm{d}X_1(t) = b_1(X_1(t-), \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t), \tag{7}$$

where the drift function has the following form:

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

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

$$\mathrm{d}X_2(t) = b_2(X_2, \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t), \tag{8}$$

where $X_2 \equiv \{X_2(t)\}_{t\geq 0}$ denotes the whole process and the drift function is chosen as follows:

$$b_2(X_2, \alpha) \triangleq -c_\alpha \sum_{k=0}^{\infty} \nabla f(X_2(k\eta))\mathbb{I}_{[k\eta,(k+1)\eta[}(t).$$

Here, $\mathbb{I}$ denotes the indicator function, i.e. $\mathbb{I}_A(x) = 1$ if $x \in A$ and $\mathbb{I}_A(x) = 0$ if $x \notin A$. It is easy to verify that $X_2(k\eta) = W^k$ 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:

$$\mathrm{d}X_3(t) = b(X_3(t-), \alpha)\mathrm{d}t + \beta^{-1/\alpha}\mathrm{d}L^\alpha(t), \tag{9}$$

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

$$(b(x,\alpha))_i \triangleq -\frac{\mathcal{D}_{x_i}^{\alpha-2}\left(\phi(x)\frac{\partial f(x)}{\partial x_i}\right)}{\phi(x)}. \tag{10}$$

Here, $\mathcal{D}_{x_i}$ denotes the Riesz derivative along the direction x_i (Ortigueira et al., 2014). With this definition for the drift, we have the following result for the invariant measure of X_3, which is an extension of Theorem 1 to general d and β.

Lemma 1. *The SDE (9) with drift b defined by (10) admits π as an invariant distribution of its solution $(X_3(t))_{t\geq 0}$.*

The process $\{X_3(t)\}_t$ will play an important role in our analysis, since it will enable us to relate W^k to the Gibbs measure π, whose samples will be close to the global optimum $x^\star$ with high probability (Pavlyukevich, 2007).

We now state our assumptions that will imply our main result.

H1. *There exists a constant $B \geq 0$ such that*

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

H2. *The gradient of f is Hölder continuous with constants $M > 0$, $0 \leq \gamma < 1$:*

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

H3. *For some $m > 0$ and $b \geq 0$, f is (m, b, γ)-dissipative:*

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

The assumptions **H**1-**H**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 **H**3 weaker and **H**2 more restrictive than the case where $\gamma = 1$. Nevertheless, **H**2 can be replaced by *local* Hölder continuity. A more detailed discussion is given in the supplementary document.

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

H4. *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 **H**2 and 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.

Next, we require b to be dissipative for *large distances* and we assume a bounded moment condition, which will be used for establishing the ergodicity of X_3.

H5. *1) For all $x, y \in \mathbb{R}^d$ and for some constants $\bar{\gamma} \in [0, 1]$, $l_0 \geq 0$, $K_1 > 0$ and $K_2 > 0$, the following holds:*

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

2) For any $t > 0$, $\hat{\gamma} \in (0, \alpha)$, and for any coupling P_t of $X_3(t)$ and $\hat{W} \sim \pi$, we have:

$$\int \|X_3(t) - \hat{W}\|^{\hat{\gamma}} \mathrm{d}P_t < C_*,$$

for some constant $C_* > 0$.

Proposition 1. *Under assumptions **H**1-**H**3 and **H**5, the distribution of $X_3(t)$ 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*

$$\mathcal{W}_\lambda(\mu_{3t}, \pi) \leq Ce^{-C_1 t}, \tag{11}$$

where μ_{3t} denotes the probability measure of $X_3(t)$.

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 (11) 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.

H6. *There exists a constant $L > 0$ such that $L < m$ and*

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

where the function b is defined in (10).

In Corollary 2 of (Şimşekli, 2017), it has been shown that **H**6 holds if the tails of π vanish sufficiently quickly (cf. Assumption **H**4 in (Şimşekli, 2017)). On the other hand,

the gap between 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 & Duman, 2012) (cf. Theorem 2 of (Şimşekli, 2017)).

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

H7. *There exists a constant $C > 0$ such that*

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

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

Now, we are ready to state our main result.

Theorem 2. *Under conditions **H1**-**H7** 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:*

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

More explicit constants can be found in the supplementary document. 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.

4. Proof Overview

Our proof strategy consists of bounding each of the four terms in (6) 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 & Wu, 2016) to Wasserstein distance with fractional orders.

Lemma 2. *Let V and 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 $0 \leq \gamma < 1$,*

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

and $\max\left\{ \left(\mathbb{E}\|W\|^{\gamma p}\right)^{\frac{1}{p}}, \left(\mathbb{E}\|V\|^{\gamma p}\right)^{\frac{1}{p}} \right\} < \infty$. 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 (6) 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 $X_i(t)$.

The term $\mathcal{W}_q(\mu_{3t}, \pi)$ is related to the ergodicity of the process (9) and it has been shown that this distance diminishes exponentially for a considerably large class of Lévy diffusions (Masuda, 2007; Xie & 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 **H**5 and **H**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 (7). 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 & 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:*

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

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

$$\begin{aligned}\Delta X_{ij}(s) &\triangleq X_i(s) - X_j(s) \\ \Delta b_{ij}(s-) &\triangleq b_i(X_i(s-), \alpha) - b_j(X_j(s-), \alpha).\end{aligned}$$

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 3. *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 the supplementary document. By only considering the leading terms of the bound provided in Theorem 3, 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* $X_1(t)$ *and* $X_2(t)$ *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 (6).

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

$$\left|\mathbb{E}[f(X_1(k\eta))] - \mathbb{E}[f(X_2(k\eta))]\right| \leq C\left(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}} + k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}}d\right).$$

Remark 1. *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 4. *We have the following estimate:*

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

By considering the leading terms of the bound in Theorem 4 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:*

$$|\mathbb{E}[f(X_1(k\eta))] - \mathbb{E}[f(X_3(k\eta))]| \leq C\left(k^{\gamma+\frac{\gamma+q}{q}}\eta^{\gamma+\frac{1}{q}}\beta^{-\frac{\gamma}{\alpha}}d + k^{\gamma+\frac{\gamma+q}{q}}\eta^{\frac{1}{q}}\right).$$

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

We now pass to the term $\mathbb{E}f(X_3(k\eta)) - \mathbb{E}f(\hat{W})$ of (6). Since we already assumed that μ_{3t} exponentially converges to π in Wasserstein distance (cf. **H**5), as a direct application of Lemma 2, we obtain the following result.

Lemma 4. *Let* $\hat{W}$ *be a random variable drawn from the invariant measure* $\pi \propto \exp(-\beta f)$ *of* (9). *There exists a constant* $C \geq 0$ *such that the following bound holds:*

$$|\mathbb{E}[f(X_3(t))] - \mathbb{E}[f(\hat{W})]| \leq C\frac{b\beta + d}{m}\exp(-\lambda_*\beta^{-1}t).$$

Remark 3. *For any* $\varepsilon > 0$, *if we take* $k\eta > \frac{\beta}{\lambda_*}\log\left(\frac{1}{\varepsilon}\right)$, *then the bound in Lemma 4 can be scaled as* $\varepsilon\mathrm{Poly}(\beta, d)$.

We finally bound the term $\mathbb{E}f(\hat{W}) - f^*$, 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*

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

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

5. Additional Remarks

5.1. 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 2 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 FLA has a worse dependency both on k and η, which is not surprising and indeed in-line with the existing literature (Mikulevičius & Zhang, 2011).

5.2. Discussion on smoothness assumptions

In this section we will discuss Assumption **H**4 and provide a condition on γ such that **H**4 holds. Let us recall the four

constraints given in **H**4:

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

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.

By solving these four conditions (the details are given in the supplementary document), we obtain $1 < q < (1+\sqrt{5})/2$, $p > (3+\sqrt{5})/2$, and $\gamma < 1/p < (3-\sqrt{5})/2$.

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

Under these observations, Theorem 2 is restated in Corollary S1 in the supplementary document. 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.

6. Extensions

6.1. Guarantees for Posterior Sampling

In this section, we will discuss the implications of our results in the classical Monte Carlo sampling context. 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 W^k and π. The result is stated as follows:

Corollary 5. *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 $\mathrm{P}(X|Y)$.

6.2. Extension to Stochastic Gradients

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

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

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:

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

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

Theorem 5. *Assume that for each i, the function $x \mapsto f^{(i)}(x)$ satisfies the conditions **H**1-**H**7. Let us replace ∇f by ∇f_k in (3). If, in addition, there exists $\delta \in [0,1)$ for any k, such that*

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

for $x \in \mathbb{R}^d$, 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 5 for obtaining a bound for the expected discrepancy, given as follows:

Corollary 6. *Under the same assumptions as in Theorem 5, we have the following bound:*

$$\big|\mathbb{E}[f(X_1(k\eta))] - \mathbb{E}[f(X_2(k\eta))]\big| \leq C(1+\delta)\Big(k^{1+\frac{1}{q}}\eta^{\frac{1}{q}} + k^{1+\frac{1}{q}}\eta^{\frac{1}{q}+\frac{\gamma}{\alpha q}}\beta^{-\frac{(q-1)\gamma}{\alpha q}} d\Big).$$

These results show that the guarantees for FLA will still hold even under the presence of stochastic gradients.

7. Conclusion

In this study, we focused on FLA, which is a recent extension of ULA, and can be seen as a perturbed version of the gradient descent algorithm with heavy-tailed α-stable noise. We analyzed the non-asymptotic behavior of FLMC 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 FLA increases faster than ULA, which suggests using smaller step-sizes in FLA. We finally extended our results to the case where exact gradients are replaced by stochastic gradients and showed that similar results hold in this setting as well. A clear future direction implied by our results is the investigation of the local behavior of FLA.

Acknowledgments

This work is supported by the French National Research Agency (ANR) as a part of the FBIMATRIX project (ANR-16-CE23-0014). We thank Reviewer 3 for her/his insightful comments.

References

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. sgmcmc: An R package for stochastic gradient Markov chain Monte Carlo. *arXiv preprint arXiv:1710.00578*, 2017.

Birdal, T. and Şimşekli. Probabilistic permutation synchronization using the Riemannian structure of the Birkhoff polytope. In *CVPR*, 2019.

Birdal, T., Şimşekli, U., Eken, M. O., and Ilic, S. Bayesian pose graph optimization via bingham distributions and tempered geodesic MCMC. In *NeurIPS*, pp. 308–319, 2018.

Çelik, C. and Duman, M. Crank–Nicolson method for the fractional diffusion equation with the Riesz fractional derivative. *Journal of Computational Physics*, 231(4): 1743–1750, 2012.

Chambers, J. M., Mallows, C. L., and Stuck, B. W. A method for simulating stable random variables. *Journal of the american statistical association*, 71(354):340–344, 1976.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pp. 2269–2277, 2015.

Şimşekli, U., Badeau, R., Cemgil, A. T., and Richard, G. Stochastic quasi-Newton Langevin Monte Carlo. In *ICML*, 2016.

Şimşekli, U., Yildiz, C., Nguyen, T. H., Cemgil, A. T., and Richard, G. Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. In *ICML*, pp. 4674–4683, 2018.

Şimşekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *ICML*, 2019.

Dalalyan, A. S. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *Proceedings of the 2017 Conference on Learning Theory*, 2017a.

Dalalyan, A. S. 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.

Durmus, A. and Moulines, E. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *arXiv preprint arXiv:1507.05021*, 2015.

Durmus, A. and Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Durmus, A., Şimşekli, U., Moulines, E., Badeau, R., and Richard, G. Stochastic gradient Richardson-Romberg Markov Chain Monte Carlo. In *NIPS*, 2016.

Erdogdu, M. A., Mackey, L., and Shamir, O. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pp. 9693–9702, 2018.

Gairing, J., Högele, M., and Kosenkova, T. Transportation distances and noise sensitivity of multiplicative Lévy sde with applications. *Stochastic Processes and their Applications*, 128(7):2153–2178, 2018.

Gao, X., Gurbuzbalaban, M., and Zhu, L. Breaking reversibility accelerates Langevin dynamics for global nonconvex optimization. *arXiv preprint arXiv:1812.07725*, 2018a.

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 preprint arXiv:1809.04618*, 2018b.

Gelfand, S. B. and Mitter, S. K. Recursive stochastic algorithms for global optimization in R^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

Hwang, C. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pp. 1177–1182, 1980.

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.

Lamberton, D. and Pages, G. 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.

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

Ma, Y. A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2899–2907, 2015.

Masuda, H. Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stochastic processes and their applications*, 117(1):35–56, 2007.

Mikulevičius, R. and Zhang, C. 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.

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

Ortigueira, M. D., Laleg-Kirati, T. M., and Machado, J. A. T. Riesz potential versus fractional Laplacian. *Journal of Statistical Mechanics*, (09), 2014.

Panloup, F. 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.

Pavlyukevich, I. Cooling down lévy flights. *Journal of Physics A: Mathematical and Theoretical*, 40(41):12299, 2007.

Polyanskiy, Y. and Wu, Y. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1674–1703, 2017.

Roberts, G. O. and Stramer, O. Langevin Diffusions and Metropolis-Hastings Algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, December 2002. ISSN 13875841.

Samorodnitsky, G. and Taqqu, M. S. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC press, 1994.

Şimşekli, U. Fractional Langevin Monte carlo: Exploring Levy driven stochastic differential equations for Markov chain Monte Carlo. In *ICML*, pp. 3200–3209, 2017.

Tzen, B., Liang, T., and Raginsky, M. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In *Proceedings of the 2018 Conference on Learning Theory*, 2018.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pp. 681–688, 2011.

Xie, L. and Zhang, X. Ergodicity of stochastic differential equations with jumps and singular coefficients. *arXiv preprint arXiv:1705.07402*, 2017.

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. 3125–3136, 2018.

Yanovsky, V. V., Chechkin, A. V., Schertzer, D., and Tur, A. V. Lévy anomalous diffusion and fractional Fokker–Planck equation. *Physica A: Statistical Mechanics and its Applications*, 282(1):13–34, 2000.

Ye, N. and Zhu, Z. Stochastic fractional Hamiltonian Monte Carlo. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3019–3025, 7 2018.

Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1980–2022, 2017.