On the Theoretical Properties of Noise Correlation in Stochastic Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Studying the properties of stochastic noise to optimize complex non-convex functions has been an active area of research in the field of machine learning. Prior work [51, 46] has shown that the noise of stochastic gradient descent improves optimization by overcoming undesirable obstacles in the landscape. Moreover, injecting artificial Gaussian noise has become a popular idea to quickly escape saddle points. Indeed, in the absence of reliable gradient information, the noise is used to explore the landscape, but it is unclear what type of noise is optimal in terms of exploration ability. In order to narrow this gap in our knowledge, we study a general type of continuous-time non-Markovian process, based on fractional Brownian motion, that allows for the increments of the process to be correlated. This generalizes processes based on Brownian motion, such as the Ornstein-Uhlenbeck process. We demonstrate how to discretize such processes which gives rise to the new algorithm "fPGD". This method is a generalization of the known algorithms PGD and Anti-PGD [33]. We study the properties of fPGD both theoretically and empirically, demonstrating that it possesses exploration abilities that, in some cases, are favorable over PGD and Anti-PGD. These results open the field to novel ways to exploit noise for training machine learning models.

1 Introduction

Injecting random noise to the parameters of a function has been a commonly used technique in optimization, dating back to at least [28, 2]. More recently, [29] demonstrated the empirical benefits of injecting Gaussian noise (a.k.a. Brownian motion noise) to the gradients while training deep networks. Injecting Gaussian noise has also been a common technique used to show faster escape from saddle points [14]. Importantly, the noise of the optimizer also has an effect on the type of minima that can be reached, raising some important connections to flatness and generalization properties [21]. A distinctive property of noise is its exploration ability which is crucial when the gradient of the objective function is not reliable, either due to saddles, valleys, plateaus, or local minima. Despite a large body of work debating the distribution of the intrinsic noise of stochastic gradient descent (SGD) [40, 50, 42], there is a large gap in our knowledge regarding the role of noise to escape undesirable regions.

In order to address this question, we investigate the behavior of a general type of non-Markovian stochastic process known as fractional Brownian motion (fBM). The latter is a generalization of Brownian motion where the increments need not be independent but can instead be positively or negatively correlated. Interestingly, the processes that can be constructed from fBM recover known instances in the literature, including perturbed gradient descent (PGD) [51, 20], and anticorrelated PGD (Anti-PGD) [33]. As we will see, the correlation between increments has a drastic effect of the behavior of the optimizer, and neither PGD or Anti-PGD are optimal in all cases.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Contribution. The first question we address is *what is the effect of the noise correlation on the overall magnitude of the process and its ability to explore the landscape?* In many applications, this is typically captured by ensuring a uniform control on the evolution of the random process. This can be achieved by bounding its expected supremum, which is connected to the concept of exit time, see Remark 2. To conduct such an analysis, we will consider a well-known type of continuous-time stochastic process with drift known as the fractional Ornstein-Uhlenbeck (fOU) process, which is connected to SGD (see Section 3). We will also demonstrate how to discretize such a process to obtain a practical optimization algorithm that we name fPGD (for fractional PGD). Concretely, our main result will be to provide upper and lower bounds on the expected supremum of the fOU process. These bounds will give a precise characterization of the effect of the noise correlation on the magnitude of the process. Second, we experimentally investigate the practical behavior of fBM noise injection: In Section 5.1, we demonstrate on a bi-stable potential that fBM perturbations indeed help optimizers to move faster from one local minimum to another – as suggested by our theory. In Section 5.2, we demonstrate the existence of an optimization problem where non-Markovian versions of fPGD indeed outperform the known Markovian special cases (PGD and Anti-PGD) – a first proof-of-concept for the practical utility of non-Markovian noise injection.

2 Related work

Non-Markovian stochastic processes. The vast majority of the work done in stochastic optimization for machine learning focuses on Markovian processes. Yet, non-Markovian processes are known to appear naturally in many situations. They for instance describe the widths of consecutive annual rings of a tree or the values of the log return of a stock [6]. In this paper, we are especially interested in fractional Brownian Motion (fBM) as an example of a non-Markovian process. fBM is a generalization of Brownian motion where the increments need not be independent but can instead be positively or negatively correlated. The amount of correlation is captured by the Hurst parameter $H \in (0, 1)$ (a formal introduction to fBM is discussed in Section 3).

Mathematical techniques for non-Markovian processes are largely underdeveloped compared to their Markovian counterparts. There are two sets of results that are of relevance to the problem discussed in this work. The first set of results are escape times for fBM which are studied by [25, 4, 37, 41, 43], but they are restricted to simple 1-dimensional processes, and they are often applicable to specific problems such as Kramer's problem, instead of more general problems of interest. Extending these results to multiple dimensions is difficult, and we are not aware of any existing result.

The second results of relevance are bounds on the expected supremum of the stochastic process, sometimes referred to as maximal inequalities. There is a large body of work focusing on the expected supremum of "pure" fBM (i.e., without any drift term), including [31, 27, 36]. For instance, [8] establish both upper and lower bounds on the expected supremum of fBM scaling as $\mathcal{O}\left(1/\sqrt{H}\right)$. Some limited results exist with time-dependent drift terms in the 1-dimensional case, e.g. [35] gives an asymptotic result for $H \in \left(\frac{1}{2}, 1\right)$. In the multi-dimensional case, [23] derive lower and upper bound for the distribution of fBM in the multi-dimensional case, see Remark 4. In this work, our goal will be to provide sharp bounds for stochastic processes with more general drift terms than prior work and in the multi-dimensional case.

Brownian motion and Ornstein-Uhlenbeck process. A well-studied type of stochastic process is Brownian motion (BM), which is a special case of fBM (corresponding to $H = \frac{1}{2}$) where the increments are not correlated. When also considering a stochastic process with a drift, we obtain the well-known Ornstein-Uhlenbeck process, which has been the subject of a large body of work, including both results on the exit time [1, 9] as well as on the expected supremum [15, 19]. Finally, [10] also derived a maximal inequality for a multi-dimensional Ornstein-Uhlenbeck process, although their result is restricted to the radial case.

SGD & Noise Injection. The distribution of the noise of SGD is often a subject of debate in the literature. In some cases, the noise of SGD appears to be heavy-tailed [34, 40, 16], which yields faster exit from sharp to flat minima [30]. Noise is also often added to (stochastic) gradient descent. For instance, Perturbed Gradient Descent (PGD) is a version of GD where artificial noise is added to the parameters after every step. Multiple PGD methods have been shown to help quickly escape spurious local minima [51] and saddle points [20]. Another way to add perturbations to SGD is to add noise to the labels of the data used for training. Recent work has demonstrated that such perturbations are indeed beneficial for generalization by implicitly regularizing the loss [7, 17, 12].

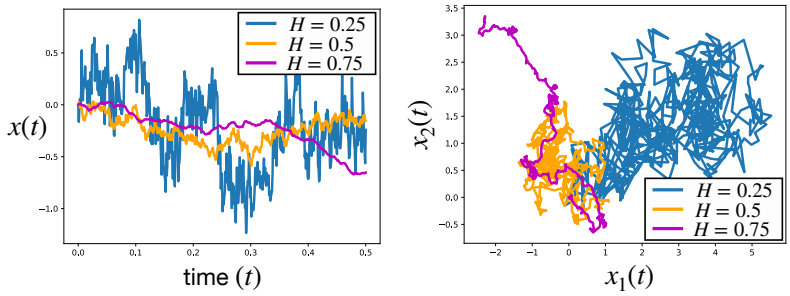

Figure 1: Discretized path of fBM in one dimension (left) and two dimensions (right) for different Hurst parameters. We observe that low values of H yield stochastic processes that explore the space in a denser way that is less "biased" by the direction of the first few perturbations. This can in fact be formalized through the concept of Hausdorff dimension which is given by $d = 2 - H$ for all $H \in (0, 1)$, see Remark 2.

3 Background

Fractional Brownian Motion. Fractional Brownian motion (fBM) is a generalization of Brownian motion where the increments of the stochastic process need not be independent. This family of Gaussian processes was developed for some hydrological modelling by [22]. [18] later studied the long term water flow characteristics of the Nile River that was affected by long periods of drought. In such a scenario, the self similarity of the process is related to a long term dependence in the water levels. Since then, fBM has also found applications in financial modeling [32]. We now give a formal definition of fBM, and discuss its main properties.

Definition 1. *Fractional Brownian motion (fBM) is a centered Gaussian process* $(B^H(t))_{t\in[0,T]}$ *for* $T > 0$ *with covariance function*

$$\mathbb{E}[B^H(t)B^H(s)] = \tfrac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

where $H \in (0, 1)$ *is called the* ***Hurst parameter****.*

Hence, fBM is a generalization of Brownian motion, which is recovered for $H = \frac{1}{2}$. If $H > \frac{1}{2}$, then the increments are positively correlated, while they are negatively correlated if $H < \frac{1}{2}$. For $H \neq \frac{1}{2}$, the increments of fBM are not independent, although they are stationary, with variance

$$\mathbb{E}(|B^H(t) - B^H(s)|^2) = |t-s|^{2H}. \tag{1}$$

A crucial property of negatively correlated increments is their ability to fill in the space, as illustrated in Figure 1; see Remark 2.

Fractional Ornstein-Uhlenbeck process. We conduct our analysis on the fractional Ornstein-Uhlenbeck process, which can be thought of as a generalization of the well-known Ornstein-Uhlenbeck process that has already found many applications in the field of machine learning [7, 24, 44].

Formally, this process is a d-dimensional stochastic process $X_t^H \in \mathbb{R}^d$ defined by

$$dX_t^H = AX_t dt + dB_t^H, \quad X_0 = \xi, \tag{2}$$

where $A \in \mathbb{R}^{d\times d}$ and $B_t^H = (B_t^{H,1}, \dots B_t^{H,d})$ is a fBM with $H \in (0, 1)$. For further information regarding the fractional Ornstein-Uhlenbeck process, we refer the reader to [11, 26, 3]. We will for instance later make use of the covariance of this process that has been derived by [26, 3].

The fractional Ornstein-Uhlenbeck process is relevant for optimization purposes as it corresponds to optimizing a quadratic function of the form $f(x) = \frac{1}{2}x^\top Ax$. Indeed, the gradient of the function $\nabla f(x) = Ax$ then corresponds to the drift term in Eq. (2). We also note that when the stochastic process reaches stationarity, the stationary distribution of the iterates is known to be approximated by a quadratic function [24, 39]. This implies that such processes can be used to analyze the behavior of complex functions at stationarity, but also close to saddle points where quadratic functions often play a crucial role to analyze the escape behavior of gradient descent methods [20].

Discretization by fractional Perturbed Gradient Descent (fPGD). For standard Brownian motion B_t (i.e. for Hurst parameter $H = 1/2$), the discretization of the continuous-time SDE $dX_t = -\nabla f(X)dt + \sigma dB_t$ of any optimization problem $\operatorname{argmin}_x f(x)$, is written

$$x_{n+1} = x_n - \eta \nabla f(x_n) + \sigma[B_{\eta(n+1)} - B_{\eta(n))}], \tag{3}$$

with step size $\eta > 0$. This method is called *perturbed gradient descent* (PGD) [20]. Here, by instead discretizing the fractional SDE $dX_t = -\nabla f(X)dt + \sigma dB_t^H$ for any $H \in (0,1)$, we obtain a generalized version of PGD, called *fractional PGD* (fPGD):

$$x_{n+1} = x_n - \eta \nabla f(x_n) + \sigma[B^H_{\eta(n+1)} - B^H_{\eta(n))}], \tag{4}$$

which is equal to PGD if $H = 1/2$. The parameter H determines the correlation of noise in fPGD: anticorrelated noise for $H \in (0, 1/2)$, positively correlated noise for $H \in (1/2, 1)$, and uncorrelated (independent) noise for $H = 1/2$. Importantly – since B_t^H converges to a white-noise process (in distribution) as $H \to 0$; see Lemma 4.1 by [8] – the method Anti-PGD [33, Section 2] can be considered as the limit case of fPGD as $H \downarrow 0$. Hence, fPGD with $H \in (0, 1/2)$ interpolates between PGD ($H = 1/2$) and Anti-PGD ($H = 0$). Positive correlations ($H \in (1/2, 1)$) have not been shown to be useful yet, and indeed the below theorems and experiments will indicate that a small H yields quicker escapes from difficult regions such as saddle points and local minima. While the next section will analyze the continuous-time model based on fBM, we will return to the optimizer fPGD in the experiments (Section 5) where we will illustrate how the theoretical properties relate to optimization, and why H might, in some settings, be a useful tuning parameter for perturbed-gradient methods.

4 Analysis

In this section, we provide upper and lower bounds on the expected supremum of the process X_t, i.e. $\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H\|\right]$. Since the analysis is rather technical, we first present the main ideas in the 1-dimensional case, before presenting the general result (which builds on the 1-dimensional case).

4.1 Preliminaries

Our analysis will make use of various bounds based on a technique known as chaining that provides a general framework that can be used to prove uniform bounds on a random process. The main tools we use are Dudley's inequality and Sudakov's inequality, which give upper and lower bounds on $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ in terms of the metric entropy of T denoted by $\log \mathcal{N}(T, d, \epsilon)$. Both inequalities are stated in Appendix A.1. We will first be interested in bounding the quantity $\mathbb{E}[\sup_{t\in T} X_t]$ in the 1-dimensional case where $T \subseteq \mathbb{R}$ is a given interval. The proof for the general d-dimensional case builds on the 1-dimensional case.

We will consider processes that have sub-Gaussian increments as defined below. For that, we recall the definition of the sub-Gaussian norm of a random variable X, denoted by $\|X\|_{\psi_2}$, as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$.

Definition 2. *A random process $(X_t)_{t\in T}$ defined on a metric space (T, d) has sub-Gaussian increments if there exists $K \geq 0$ such that $\|X_t - X_s\|_{\psi_2} \leq K d(t, s) \quad \forall t, s \in T$.*

4.2 One-dimensional case

In this section, we derive both upper and lower bounds on the expected supremum of the stochastic process defined in Eq. (2). We start with the upper bound and then with a lower bound.

Theorem 1 (Upper bound, 1-dimensional case)**.** *Consider the stochastic process X_t^H defined by*

$$dX_t^H = \beta X_t dt + \sigma dB_t^H,$$

with $X_0 = 0, \sigma \in \mathbb{R}$ and $|\beta| < 1$. Then, for $t \in [0, 1]$,

$$\mathbb{E}\left[\sup_{t\in[0,1]} X_t^H\right] \leq C\left\{\sigma\sqrt{C_\beta} + \sqrt{\frac{1}{H}}\int_0^{2\sigma\sqrt{C_\beta}} \sqrt{\log\left(\frac{2\sigma\sqrt{C_\beta}}{\varepsilon}\right)} d\varepsilon\right\},$$

where $C_\beta = \frac{e^{2\beta}}{1-\beta}$ and C are positive constants (independent of H).

Theorem 2 (Lower bound, 1-dimensional case)**.** *Consider the stochastic process X_t^H defined by*

$$dX_t^H = -\beta X_t dt + \sigma dB_t^H,$$

with $X_0 = 0, \sigma \in \mathbb{R}$ and $|\beta| < 1$. Then

$$\mathbb{E}\left[\sup_{t\in[0,1]} X_t^H\right] \geq c\left\{e^{-\frac{1}{2}}\sqrt{\frac{1}{H}\log\left(\frac{\sigma C_\beta}{e^{-\frac{1}{2}}}\right)}\right\},$$

where $C_\beta = e^{-2|\beta|}$ and c are positive constants (independent of H).

We note that both bounds scale with $\mathcal{O}\left(\frac{1}{\sqrt{H}}\right)$, which implies the bound is tight and can not be improved in terms of the Hurst parameter.

Remark 1 (Extension to any finite interval)**.** *Although the theorems are stated for the interval $T = [0, 1]$, one can extend the results to an arbitrary finite interval using the self-similarity property of fBM, namely $B_{at}^H \overset{d}{=} a^H B_t^H$ for $a > 0$. This implies that $\mathbb{E}\left[\sup_{s\in[0,t]} B_s^H\right] = t^H \mathbb{E}\left[\sup_{s\in[0,1]} B_s^H\right]$.*

4.3 Multi-dimensional case

We now generalize the bounds to the multi-dimensional case. The quantity of interest becomes the expected supremum of the norm, or more precisely $\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H - \mathbb{E}[X_t^H]\|\right]$. Similarly to the 1-dimensional case, we apply Sudakov's inequality to the Gaussian component of the multi-dimensional fOU process X_t^H [1].

Theorem 3 (Upper bound, d-dimensional case)**.** *Consider the d-dimensional stochastic process $X_t^H \in \mathbb{R}^d$ defined by*

$$dX_t^H = AX_t dt + dB_t^H, X_0 = \xi,$$

where $A \in \mathbb{R}^{d\times d}$ and $B_t^H = (B_t^{H,1}, \dots B_t^{H,d})$ is a fBM with $H \in (0, 1)$. Denote by $a_{ij}(t)$ the matrix entries of $\exp(tA)A$. Then

$$\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H - \mathbb{E}[X_t^H]\|\right] \leq Cd^2C_2 + Cd\frac{1}{\sqrt{H}}\int_0^{dC_2}\sqrt{\log\left(\frac{dC_2}{\varepsilon}\right)}d\varepsilon,$$

where $C_2 = \max_{i,j=1}^d \sup_{t\in[0,1]} |a_{ij}(t)|^2 + \max_{i,j=1}^d \sup_{t\in[0,1]} |a_{ij}(t)| + 1$ and $C > 0$ is a constant independent of H.

Theorem 4 (Lower bound, d-dimensional case)**.** *Consider the d-dimensional stochastic process $X_t^H \in \mathbb{R}^d$ defined by*

$$dX_t^H = AX_t dt + dB_t^H, X_0 = \xi,$$

where $A \in \mathbb{R}^{d\times d}$ and $B_t^H = (B_t^{H,1}, \dots B_t^{H,d})$ is a fBM with $H \in (0, \frac{1}{2})$. Denote by $a_{ij}(t)$ the matrix entries of $\exp(tA)A$ and assume that there exists $i_0 \in \{1, \dots d\}$ such that $\min_{t\in[0,1]} a_{i_0 i_0}(t) \geq -1 + C_1$, where $0 < C_1 < 1$. Then

$$\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H - \mathbb{E}[X_t^H]\|\right] \geq ce^{-\frac{1}{2}}\sqrt{\frac{1}{H}\log\left(\frac{2C_1}{e^{-\frac{1}{2}}}\right)},$$

where c is a constant independent of H.

[1] An alternative option, which we do not pursue in this paper, would be to generalize the result of Sudakov's inequality to the case of sub-Gaussian or sub-exponential processes. However, the proof technique presented by [45] strongly relies on Gaussian comparison inequalities and might be difficult to generalize.

We obtain an upper bound of the form $\mathcal{O}\left(\frac{1}{\sqrt{H}}\right)$ with a matching lower bound, although we do require a condition on the entries of the matrix A, which we discuss further in Example 1 below.

Example 1 (Assumption lower bound)**.** *Let $A = diag(\lambda_1, ..., \lambda_d)$. Then $e^{tA}A = diag(e^{t\lambda_1}\lambda_1, ..., e^{t\lambda_d}\lambda_d)$. If some diagonal element of A is non-negative, then the assumptions of the above theorem are satisfied. Suppose that all entries $\lambda_1, ..., \lambda_d$ are negative and that $\lambda_{i_0} > -1$, for some i_0. Then $\min_{t\in[0,1]} a_{i_0 i_0}(t) = \min_{t\in[0,1]}(e^{t\lambda_{i_0}}\lambda_{i_0}) \geq \lambda_{i_0} > -1 + C_1$ for some $C_1 \in (0,1)$. Hence, if some diagonal element $\lambda_{i_0} > -1$ exists, then the conditions of the previous theorem are fulfilled.*

We make a few comments regarding the meaning of the bound, as well as some extensions.

First of all, the theorems discussed above bound the expected supremum of the process. We now explain how these are connected to what we call "exploration ability".

Remark 2 (Exploration ability)**.** *We give three different justifications for the concept of exploration. 1) Expected behavior: Based on Theorem 4, we have*

$$\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H\|\right] + \sup_{t\in[0,1]} \mathbb{E}\|X_t^H\| \geq \mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H - \mathbb{E}[X_t^H]\|\right] \geq ce^{-\frac{1}{2}}\sqrt{\frac{1}{H}\log\left(\frac{2C_1}{e^{-\frac{1}{2}}}\right)},$$

where c, C_1 are constant independent of H.
Since $\mathbb{E}X_t^H$ solves the ODE $\phi(t) = \xi + \int_0^t A\phi(s)ds$, the quantity $a := \sup_{t\in[0,1]} \|\mathbb{E}[X_t^H]\| < \infty$ is independent of H. Thus $\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H\|\right] \geq ce^{-\frac{1}{2}}\sqrt{\frac{1}{H}\log\left(\frac{2C_1}{e^{-\frac{1}{2}}}\right)} - a$. Hence, for a given $H_1 \in \left(0, \frac{1}{2}\right)$, there is an $H < H_1$ such that $\boxed{\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H\|\right] > \mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^{H_1}\|\right]}$.

*We conclude that the **average maximal distance of X_t^H from the origin is larger than that of $X_t^{H_1}$**. This is what we mean by the process having a better exploration ability. Better exploration, in this sense, can help to escape suboptimal local minima more quickly (experiments in Section 5.1) or to explore the space around a saddle point for an exit (experiments in Section 5.2).*

2) Exit time: First, note that $\mathbb{E}\left[\sup_{t\in[0,1]} \|X_t^H\|\right] = \int_0^\infty \mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^H\| > s\right) ds$, therefore $\int_0^\infty \left\{\mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^H\| > s\right) - \mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^{H_1}\| > s\right)\right\} ds > 0$.

*Thus, we can find a $s_0 > 0$ such that $\mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^H\| > s_0\right) - \mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^{H_1}\| > s_0\right) > 0$ and, since probability distributions are right-continuous, there exists $\epsilon > 0$ such that for all $s \in [s_0, s_0 + \epsilon)$, $\mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^H\| > s\right) > \mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^{H_1}\| > s\right) > 0$. On the other hand, $\mathbb{P}\left(\sup_{t\in[0,1]} \|X_t^H\| > s\right) = \mathbb{P}(\tau_s^H < 1)$ where τ_s^H denotes the first exit time of X^H from the ball $B_s(0)$ of radius s. We conclude that $\boxed{\mathbb{P}(\tau_s^H < 1) > \mathbb{P}(\tau_s^{H_1} < 1)\ \forall s \in [s_0, s_0 + \epsilon)}$, i.e. **the probability of an exit before time s of H is larger than that of H_1**.*

3) Space-filling property: We note that fractional Brownian motion with low Hurst parameter H fills the space more densely. This behavior comes from the fact that the graph of the fBM has a Hausdorff dimension and box-counting dimension given by $d = 2 - H$. The latter implies that $N(1/n) \approx Cn^d$ for large n, where $N(\varepsilon)$ denotes the number of boxes with side length ε, which is needed to cover the graph of the fBM. Therefore, the graph of a (1-dimensional) fBM "behaves" like a $2-$dimensional smooth manifold for small H. See [38]. The above claim about the Hausdorff dimension and the box-filling dimension of fBM is proved by [47]. Further related articles are [5] and [48].

Remark 3 (More general version)**.** *Let us mention that the results in Theorems 4 and 3 can be also generalized to the case of linear SDE's of the form*

$$X_t = x + \int_0^t A(s)X_s ds + \int_0^t \Sigma(s)dB_s^H,$$

where $A : [0,T] \longrightarrow \mathbb{R}^{d\times d}$ is locally bounded and measurable, $\Sigma : [0,T] \longrightarrow \mathbb{R}^{d\times d}$ is continuous differentiable and $H < \frac{1}{2}$. Here the stochastic integral with respect to $B_\cdot^H$ is given by the Young

integral. In fact, one can show that $\int_0^t \Sigma(s)dB_s^H = \Sigma(t)B_t^H - \int_0^t \frac{\partial}{\partial s}\Sigma(s)B_s^H ds$ *(see e.g. [6]). By using the linear transformation* $Y_t := X_t - \int_0^t \Sigma(s)dB_s^H$*, we observe that* $Y_t = x + \int_0^t (A(s)Y_s + a(s))ds$*, where* $a(t) := A(t)\int_0^t \Sigma(s)dB_s^H$*. Denote by* $\Phi(t), 0 \leq t \leq T$ *the fundamental solution associated with* $A(t)$*, that is the unique (absolutely continuous) solution to the equation* $\dot{\Phi}(t) = A(t)\Phi(t), \Phi(0) = I_{d\times d}$*, the* $d \times d$ *identity matrix. Hence,*

$$X_t = \Phi(t)x + \Phi(t)\int_0^t \Phi^{-1}(s)A(s)\int_0^s \Sigma(u)dB_u^H ds + \int_0^t \Sigma(s)dB_s^H, 0 \leq t \leq T.$$

Although the solution $X_t, 0 \leq t \leq T$ *is in general not stationary any longer, we can use similar arguments as in the proofs of Theorems 4 and 3 and the latter representation of the process to derive a lower and upper bound for* $\mathbb{E}\left[\sup_{0\leq t\leq 1}\|X_t\|\right]$.

Remark 4 (Known bounds for distribution of fBM in multi-dimensional case)**.** *In Theorems 4 and 3 we obtained a lower and upper bound for the expected supremum of the norm of a (centered) multi-dimensional fOU process. In comparison to the latter we also mention the following corresponding result for distributions with respect to the fBm (see Theorem 4.6 by [23]): Let* $B_t^H, 0 \leq t \leq 1$ *be a fractional Brownian motion in* $\mathbb{R}^d$ *for* $H \in (0,1)$*. Then the following lower and upper bound of the distribution of* $\sup_{0\leq t\leq 1}\left\|B_t^H\right\|$ *(*$\|\cdot\|$ *maximum norm) hold: There exist constants* $0 < K_1(H) \leq K_2(H) < \infty$ *such that for all* $0 < \varepsilon \leq 1$*:*

$$e^{-dK_2(H)\varepsilon^{-\frac{1}{H}}} \leq \Pr\left(\sup_{0\leq t\leq 1}\left\|B_t^H\right\| \leq \varepsilon\right) \leq e^{-dK_1(H)\varepsilon^{-\frac{1}{H}}}.$$

Remark 5 (Tightness of the bound)**.** *As in the 1-D case, the bound is right in terms of* H*, which is the most important quantity for the argument made in this paper. Regarding the dimension* d*, note that the lower bound does not depend on it (and thus it is not improvable) while the upper bound might not be tight.*

5 Experiments

By bounding the expected supremum of the stochastic process X_t^H, our analysis reveals some insights into its exploration capability. We have seen that the correlation between increments directly affects the supremum. Concretely, lower values of the Hurst parameters are expected to yield processes with better exploration abilities, as we detailed in Remark 2. To demonstrate the relevance of our theoretical findings for optimization, we run a series of experiments with the optimizer fPGD from Eq. (4). In Section 5.1, we will illustrate how fPGD behaves on bi-stable optimization landscapes, where there are two adjacent minima of different desirability (also studied by [51, 49]). In Section 5.2, we will demonstrate on an embedded saddle point that different choices of H might be beneficial in different optimization phases – thus justifying the introduction of fPGD as a compromise between PGD and Anti-PGD.

5.1 Moving to better local minima in non-convex optimization

We consider the behavior of fPGD in a bi-stable optimization landscape, given by

$$f(x) = \left[v_0 + \frac{k_0}{2}x^2\right]\chi_{x\leq a} + \left[v_1 - \frac{k_1}{2}(x - \frac{a+c}{2})^2\right]\chi_{x\leq a} + \left[v_2 + \frac{k_2}{2}(x-m)^2\right]\chi_{x>c}, \quad (5)$$

where χ is the indicator function and the parameters (v_0, v_1, v_2, a, c, m) determine the position and shape of two adjacent local minima. Similar landscapes have been studied in the context of escaping spurious local minima [51] and in the context of generalization in machine learning by selecting flat minima [30, 49]. We will here study the former case, but we obtain analogous results in the latter (presented in Appendix B). For the former case, we set $(v_0, v_1, v_2, a, c, m) = (0.0, 45.0, -12.5, 3.5, 13.9, 18.0)$; for the corresponding landscape see the upper left subplot of Figure 2.

On this landscape, we run fPGD with different choices of H. Note that, in the two minima (and the saddle in between at $(a+c)/2$) fPGD is indeed a discretization of a fractional Ornstein–Uhlenbeck process, like the one considered in Theorems 1 and 2. We thus expect that a small Hurst parameter leads to a faster escape from the shallow minimum (left) to the deep minimum (right). We run fPGD

for $N = 1000$ iterations with a fixed step size of $\eta = 1.0$, initialised at $x = 0$ (the bottom of the shallow minimum). To stay in the setting of Theorems 1 and 2, we set $\sigma = 80.0 \cdot N^{-H}$. Thus, for all $H \in (0, 1)$, the employed fBM B_t^H, defined on the time interval $t \in [0, N]$, is by Remark 1 distributed like $80.0 \cdot B_t^H$ on $t \in [0, 1]$ – in the sense that the discrete steps $\{B_k^H; k = 0, \dots, N\}$ are distributed like $\{80.0 \cdot B_{k/N}^H; k = 0, \dots, N\}$. By this H-dependent choice of σ, we give all fBMs the same final variance of 80.0^2 which lets us zoom in on the impact of correlation. (Otherwise the perturbations would have different variances; this way we disentangle the correlation from the variance.) We note that here the experimental outcomes are not sensitive to the precise value of σ.

Figure 2 depicts our findings. We say that fPGD is in the shallow minimum at iteration k if its iterate x_k is smaller than a, and in the deep minimum if it is larger than c; both a and c are plotted as dashed lines in the optimization landscape (top row) and in the trajectories (bottom row). In the bottom row, for all H, we start below the lower dashed line a (shallow minimum) and then cross above the higher dashed line c (deep minimum). It is evident that a smaller H enables faster crossings. We thus escape faster from the spurious shallow minimum to the desirable deep one. (Note, however, if the noise is not cooled down, one might oscillate back.) The upper right plot depicts the cumulative distribution function of the first-exit time (i.e. the first time fPGD reaches the deep minimum, above c). Again, we can observe that for a small Hurst parameter the probability of a fast exit is significantly higher.

Overall, these findings for fPGD match the continuous-time result of Theorems 1 and 2. Another bi-stable landscape (with a sharp and a flat minimum) is studied in Appendix B with the same findings. Note that physicists [41] have also observed faster escapes for smaller Hurst parameters on a related Kramer-like escape problem.

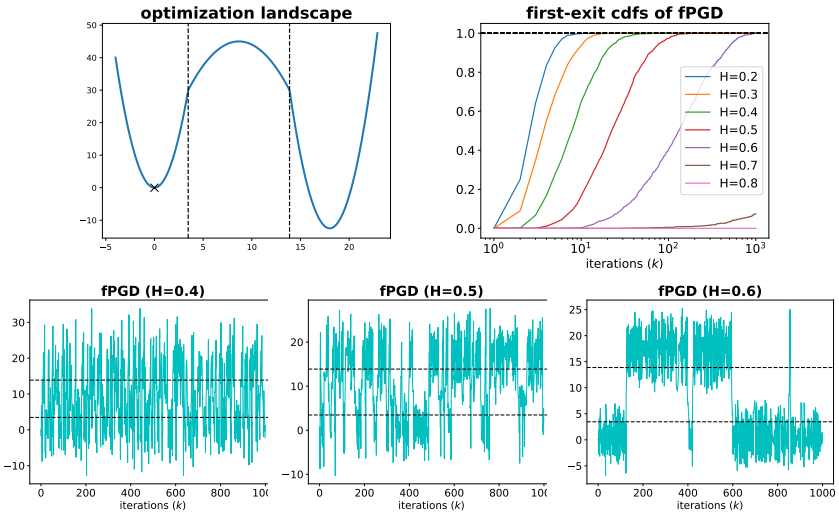

Figure 2: Experiments with fPGD (4) on the bi-stable objective (5) initialised at shallow minimum $x = 0$. Objective depicted on the upper right. Behavior of fPGD for $H \in \{0.4, 0.5, 0.6\}$ in bottom row. Smaller H gives faster switching between minima. Faster first-exit times from the shallow to the deep minimum are demonstrated by the cumulative density functions (cdf), depicted on the upper right. The cdf is computed by averaging over 1000 simulations. Details in text.

5.2 Non-Markovian fPGD as a trade-off between PGD and Anti-PGD

As discussed earlier, fPGD interpolates between two known Markovian methods: PGD ($H = 1/2$) and Anti-PGD ($H = 0$). Both PGD and Anti-PGD have unique advantages; see [51] or [20] for PGD and [33] for Anti-PGD. It is thus natural to ask whether there are cases where non-Markovian fPGD ($H \in (0, 1/2)$) is beneficial. While we do not have a complete answer to this question, we here present a setting that demonstrates that $H \in (0, 1/2)$ can indeed outperform both PGD ($H = 1/2$) and Anti-PGD ($H = 0$) simultaneously. We consider the following regularized quadratic: $f(x) = \frac{1}{2} x^\intercal M x + \lambda \sum_{i=1}^d x_i^4$, where x_i is the i-th coordinate of $x \in \mathbb{R}^d$, $M \in \mathbb{R}^{d \times d}$, $d = 400$, and $\lambda = 0.0001$. If M has negative eigenvalues, this landscape has a d-dim. saddle point at $x = 0$. If a stochastic optimizer can find the negative eigenvectors, it will reach a global minimum away from zero – while gradient descent remains stuck at zero. The problem gets increasingly hard as the number of negative eigendirections decreases [20]. Indeed, [13] showed that vanilla gradient

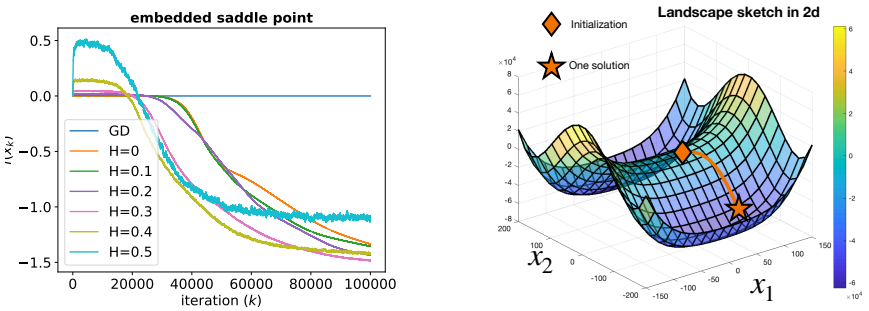

Figure 3: In the embedded saddle landscape, we see a trade-off between first exiting the saddle quickly (with a larger H) and then converging to the bottom of the minimum (with a smaller H). We plot the average over 5 runs. Non-Markovian versions of fPGD (with $H = 0.3$ or $H = 0.4$) perform better than both Markovian versions, PGD ($H = 0.5$) and Anti-PGD ($H = 0$). Details in text.

descent takes in the worst case exponential (in the number of dimensions) time to escape such saddle points. We choose M as a diagonal matrix whose diagonal entries are uniformly sampled from $[0, 1]$, and – to make it a challenging saddle – we multiply the 10 lowest of these entries by -1.

We test fPGD on this problem. We initialise at the saddle $x = 0$ and run fPGD with $\sigma = 0.005$ and $\eta = 0.005$ for 10^6 iterations with different Hurst parameters. (While in general σ and H are tuning parameters that can be jointly tuned, here $\sigma = 0.005$ is a good choice for all values of H which we demonstrate in Appendix B.2.) The results are presented in Figure 3. In this problem, there is indeed a trade-off between uncorrelated ($H = 0.5$) and anticorrelated perturbations ($H < 0.5$): While a large Hurst parameter ($H = 0.5$) moves away from the saddle quickly, the non-Markovian fPGD ($H = 0.4$) finds the direction with negative curvature most quickly (in the sense that it gets below zero). Once they have exited, all H perform similar for a while. Eventually, however, smaller Hurst parameters converge better to the bottom of the final minimum – thereby eventually catching up and surpassing the at-first faster higher Hurst parameters. Which H is best, will therefore depend on how many iterations the computational budget allows for. Here, the non-Markovian $H = 0.3$ and $H = 0.4$ yield the best final accuracy.

This suggests that, if all other parameters are the same, different choices of Hurst parameters are beneficial in different situations. It can thus be advantageous to tune and adapt the correlation of noise by choosing H. It is especially interesting that both PGD ($H = 1/2$) and Anti-PGD ($H = 0$) have shortcomings in this experiment: PGD has to increase the loss significantly in the beginning before exiting the saddle and does not converge to the bottom of the minimum in the end. Anti-PGD takes a long time to exit the saddle in the beginning. In contrast, non-Markovian fPGD (with $H = 0.3$ or $H = 0.4$) stays clear of these two undesirable extremes.

We stress, however, that further work is needed to better understand the interplay between the scale σ and the correlation parameter H, which interact according to Eq. (1). Future work could analyze the dependency between variance and correlation. Moreover, one could design an adaptive method that jointly tunes σ and H to maximize the efficiency of perturbed-gradient methods. For the above experiments, we again emphasize that our constant choice of σ is suitable for all H, as justified in Appendix B.2.

6 Conclusion

We analysed the behavior of a new type of stochastic noise with correlated increments. We found that the amount of correlation, captured by the Hurst parameter H, has a direct effect on the overall magnitude of the process and its ability to explore. This insight opens the doors to a new area of research where the noise of stochastic processes could be manipulated to speed up the convergence of an optimization method, as well as target specific minima (e.g., depending on their flatness properties). Our experimental results validate our theoretical analysis. We showed that, in bi-stable potentials, injection of fractional noise with small H speeds up the escape from a suboptimal minimum to a better one. Also, we showed that there are indeed settings where non-Markovian (somewhat anti-correlated) noise outperforms the standard Markovian (uncorrelated or perfectly anti-correlated) noise.

References

[1] Larbi Alili, Pierre Patie, and Jesper Lund Pedersen. Representations of the first hitting time density of an ornstein-uhlenbeck process. *Stochastic Models*, 21(4):967–980, 2005.

[2] Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.

[3] Giacomo Ascione, Yuliya Mishura, and Enrica Pirozzi. Fractional ornstein-uhlenbeck process with stochastic forcing, and its applications. *Methodology and Computing in Applied Probability*, 23(1):53–84, 2021.

[4] Frank Aurzada et al. On the one-sided exit problem for fractional brownian motion. *Electronic Communications in Probability*, 16:392–404, 2011.

[5] Antoine Ayache and Yimin Xiao. Asymptotic properties and Hausdorff dimensions of fractional Brownian sheets. *Journal of Fourier Analysis and Applications*, 11:407–439, 2005.

[6] Francesca Biagini, Yaozhong Hu, Bernt Øksendal, and Tusheng Zhang. *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media, 2008.

[7] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[8] Konstantin Borovkov, Yuliya Mishura, Alexander Novikov, and Mikhail Zhitlukhin. Bounds for expected maxima of gaussian processes and their discrete approximations. *Stochastics*, 89(1):21–37, 2017.

[9] Konstantin Borovkov and Alexander Novikov. On exit times of lévy-driven ornstein–uhlenbeck processes. *Statistics & Probability Letters*, 78(12):1517–1525, 2008.

[10] Yu L Botnikov. Davis-type inequalities for some diffusion processes. *Journal of Mathematical Sciences*, 137(1):4502–4509, 2006.

[11] Patrick Cheridito, Hideyuki Kawaguchi, and Makoto Maejima. Fractional ornstein-uhlenbeck processes. *Electronic Journal of probability*, 8:1–14, 2003.

[12] Alex Damian, Tengyu Ma, and Jason Lee. Label noise SGD provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

[13] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30, 2017.

[14] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[15] S Graversen and Goran Peskir. Maximal inequalities for the ornstein-uhlenbeck process. *Proceedings of the American Mathematical Society*, 128(10):3035–3041, 2000.

[16] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pages 3964–3975, 2021.

[17] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357, 2021.

[18] Harold Edwin Hurst. Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers*, 5(5):519–543, 1956.

[19] Chen Jia and Guohuan Zhao. Moderate maximal inequalities for the ornstein-uhlenbeck process. *Proceedings of the American Mathematical Society*, 148(8):3607–3615, 2020.

[20] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *J. ACM*, 68(2), 2021.

[21] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[22] Andrei N Kolmogorov. Wienersche spiralen und einige andere interessante kurven in hilbertscen raum, cr (doklady). *Acad. Sci. URSS (NS)*, 26:115–118, 1940.

[23] Wenbo V Li and Q-M Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19:533–597, 2001.

[24] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

[25] Zbigniew Michna. On tail probabilities and first passage times for fractional brownian motion. *Mathematical Methods of Operations Research*, 49(2):335–354, 1999.

[26] Yuliya Mishura, Vladimir I Piterbarg, Kostiantyn Ralchenko, and Anton Yurchenko-Tytarenko. Stochastic representation and pathwise properties of fractional cox-ingersoll-ross process. *arXiv preprint arXiv:1708.02712*, 2017.

[27] George M Molchan. Maximum of a fractional brownian motion: probabilities of small values. *Communications in mathematical physics*, 205(1):97–111, 1999.

[28] Alan F Murray and Peter J Edwards. Synaptic weight noise during mlp learning enhances fault-tolerance, generalization and learning trajectory. *Advances in neural information processing systems*, pages 491–491, 1993.

[29] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

[30] Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, 2019.

[31] Alexander Novikov and Esko Valkeila. On some maximal inequalities for fractional brownian motions. *Statistics & probability letters*, 44(1):47–54, 1999.

[32] Bernt Øksendal. Fractional brownian motion in finance. *Preprint series. Pure mathematics http://urn. nb. no/URN: NBN: no-8076*, 2003.

[33] Antonio Orvieto, Hans Kersting, Francis Bach, Frank Proske, and Aurelien Lucchi. Anticorrelated noise injection for improved generalization. *arXiv preprint arXiv:2202.02831*, 2022.

[34] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-Gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

[35] BLS Prakasa Rao. Some maximal inequalities for fractional brownian motion with polynomial drift. *Stochastic Analysis and Applications*, 31(5):785–799, 2013.

[36] BLS Prakasa Rao. Maximal inequalities for fractional brownian motion: An overview. *Stochastic Analysis and Applications*, 32(3):450–479, 2014.

[37] Lloyd P Sanders and Tobias Ambjörnsson. First passage times for a tracer particle in single file diffusion and fractional brownian motion. *The Journal of chemical physics*, 136(17):05B605, 2012.

[38] M. Schroeder. Fractals, chaos, power laws: Minutes from an infinite paradise. *New York. W. H. Freeman*, pages 41–45, 1991.

[39] Zeev Schuss. *Theory and applications of stochastic processes: an analytical approach*, volume 170. Springer Science & Business Media, 2009.

[40] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

[41] Oleksii Yu Sliusarenko, Vsevolod Yu Gonchar, Aleksei V Chechkin, Igor M Sokolov, and Ralf Metzler. Kramers-like escape driven by fractional gaussian noise. *Physical Review E*, 81(4):041119, 2010.

[42] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, pages 9058–9067, 2020.

[43] Chengli Tan, Jiangshe Zhang, and Junmin Liu. Understanding long range memory effects in deep neural networks. *arXiv preprint arXiv:2105.02062*, 2021.

[44] Prakhar Verma, ADAM Vincent, and Arno Solin. Sparse gaussian processes for stochastic differential equations. In *The Symbiosis of Deep Learning and Differential Equations*, 2021.

[45] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[46] Mingwei Wei and David J Schwab. How noise affects the Hessian spectrum in overparameterized neural networks. *arXiv preprint arXiv:1910.00195*, 2019.

[47] Dong Sheng Wu and Yi Min Xiao. Dimensional properties of fractional Brownian motion. *Acta Mathematica Sinica, English Series*, 23(4):613–622, 2007.

[48] Dongsheng Wu and Yimin Xiao. Geometric properties of fractional Brownian sheets. *Journal of Fourier Analysis and Applications*, 13(1):1–37, 2007.

[49] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

[50] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? Insights from a noisy quadratic model. *Advances in neural information processing systems*, pages 8196–8207, 2019.

[51] Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pages 7594–7602, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See discussion at the end of the experiment section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is more of theoretical nature.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Detailed proofs are available in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code for experiments are provided as supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All code were run on a Macbook 2020 with 2,3 GHz Quad-Core Intel Core i7. We did not consider this info interesting to readers, but are happy to include it, if desired.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]