Stochastic Gradient and Langevin Processes

Xiang Cheng [1] **Dong Yin** [1] **Peter Bartlett** [1] **Michael Jordan** [1]

Abstract

We prove quantitative convergence rates at which discrete Langevin-like processes converge to the invariant distribution of a related stochastic differential equation. We study the setup where the additive noise can be non-Gaussian and state-dependent and the potential function can be non-convex. We show that the key properties of these processes depend on the potential function and the second moment of the additive noise. We apply our theoretical findings to studying the convergence of Stochastic Gradient Descent (SGD) for non-convex problems and corroborate them with experiments using SGD to train deep neural networks on the CIFAR-10 dataset.

1. Introduction

Stochastic Gradient Descent (SGD) is one of the workhorses of modern machine learning. In many nonconvex optimization problems, such as training deep neural networks, SGD is able to produce solutions with good generalization error; indeed, there is evidence that the generalization error of an SGD solution can be significantly better than that of Gradient Descent (GD) (Keskar et al., 2016; Jastrzębski et al., 2017; He et al., 2019). This suggests that, to understand the behavior of SGD, it is not enough to consider the limiting cases such as small step size or large batch size where it degenerates to GD. In this paper, we take an alternate view of SGD as a sampling algorithm, and aim to understand its convergence to an appropriate stationary distribution.

There has been rapid recent progress in understanding the finite-time behavior of MCMC methods, by comparing them to stochastic differential equations (SDEs), such as the Langevin diffusion. It is natural in this context to think of SGD as a discrete-time approximation of an SDE. There are, however, two significant barriers to extending previous analyses to the case of SGD. First, these analysis are often restricted to isotropic Gaussian noise, whereas the noise in SGD can be far from Gaussian. Second, the noise depends significantly on the current state (the optimization variable). For instance, if the objective is an average over training data with a nonnegative loss, as the objective approaches zero the variance of minibatch SGD goes to zero. Any attempt to cast SGD as an SDE must be able to handle this kind of noise.

This motivates the study of Langevin MCMC-like methods that have a state-dependent noise term:

$$w_{(k+1)\delta} = w_{k\delta} - \delta \nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_{k\delta}, \eta_k), \qquad (1)$$

where $w_t \in \mathbb{R}^d$ is the state variable at time t, δ is the step size, $U : \mathbb{R}^d \to \mathbb{R}$ is a (possibly nonconvex) potential, $\xi : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is the *noise function*, and η_k are sampled i.i.d. according to some distribution over Ω (for example, in minibatch SGD, Ω is the set of subsets of indices in the training sample).

Throughout this paper, we assume that $\mathbb{E}_\eta [\xi(x, \eta)] = 0$ for all x. We define a matrix-valued function $M(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ to be the square root of the covariance matrix of ξ; i.e., for all x, $M(x) := \sqrt{\mathbb{E}_\eta [\xi(x, \eta)\xi(x, \eta)^T]}$, where for a positive semidefinite matrix G, $A = \sqrt{G}$ is the unique positive semidefinite matrix such that $A^2 = G$.

In studying the generalization behavior of SGD, earlier work (Jastrzębski et al., 2017; He et al., 2019) propose that (1) be approximated by the stochastic process $y_{(k+1)\delta} = y_{k\delta} - \delta \nabla U(y_{k\delta}) + \sqrt{\delta} M(y_{k\delta})\theta_k$ where $\theta_k \sim \mathcal{N}(0, I)$, or, equivalently:

$$dy_t = -\nabla U(y_{k\delta})dt + M(y_{k\delta})dB_t \quad \text{for } t \in [k\delta, (k+1)\delta], \qquad (2)$$

with B_t denoting standard Brownian motion (Karatzas & Shreve, 1998). Specifically, the non-Gaussian noise $\xi(\cdot, \eta)$ is approximated by a Gaussian variable $M(\cdot)\theta$ with the same covariance, via an assumption that the minibatch size is large and an appeal to the central limit theorem.

The process in (2) can be seen as the Euler-Murayama discretization of the following SDE:

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t. \qquad (3)$$

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley. Correspondence to: Xiang Cheng <x.cheng@berkeley.edu>.

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

We let p^* denote the invariant distribution of (3).

We prove quantitative bounds on the discretization error between (2), (1) and (3), as well as convergence rates of (2) and (1) to p^*. Our bounds are in Wasserstein-1 distance (denoted by $W_1(\cdot,\cdot)$ in the following). We present the full theorem statements in Section 5, and summarize our contributions below:

1. In Theorem 1, we bound the discretization error between (2) and (3). Informally, Theorem 1 states:

   1. *If* $x_0 = y_0$, *then for all* k, $W_1(x_{k\delta}, y_{k\delta}) = O(\sqrt{\delta})$;

   2. *For* $n \geq \tilde{O}\left(\frac{1}{\delta}\right)$, $W_1(p^*, \mathsf{Law}(y_{n\delta})) = O(\sqrt{\delta})$,

   where $\mathsf{Law}(\cdot)$ denotes the distribution of a random vector. This is a crucial intermediate result that allows us to prove the convergence of (1) to (3). We highlight that the variable diffusion matrix: 1) leads to a very large discretization error, due to the scaling factor of $\sqrt{\delta}$ in the $M(y_{k\delta})\theta_k$ noise term, and 2) makes the stochastic process non-contractive (this is further compounded by the non-convex drift). Our convergence proof relies on a carefully constructed Lyapunov function together with a specific coupling. Remarkably, the ϵ dependence in our iteration complexity is the same as that in Langevin MCMC with constant isotropic diffusion (Durmus & Moulines, 2016).

2. In Theorem 2, we bound the discretization error between (1) and (3). Informally, Theorem 2 states:

   1. *If* $x_0 = w_0$, *then for all* k, $W_1(x_{k\delta}, w_{k\delta}) = O(\delta^{1/8})$;

   2. *For* $n \geq \tilde{O}\left(\frac{1}{\delta}\right)$, $W_1(p^*, \mathsf{Law}(w_{n\delta})) = O(\delta^{1/8})$.

   Notably, the noise in each step of (1) may be far from Gaussian, but for sufficiently small step size, (1) is nonetheless able to approximate (3). This is a weaker condition than earlier work, which must assume that the batch size is sufficiently large so that CLT ensures that the per-step noise is approximately Gaussian.

3. Based on Theorem 2, we predict that for sufficiently small δ, two different processes of the form (1) will have similar distributions if their noise terms ξ have the same covariance matrix, as that leads to the same limiting SDE (3). In Section 6, we evaluate this claim empirically: we design a family of SGD-like algorithms and evaluate their test error at convergence. We observe that the noise covariance alone is a very strong predictor for the test error, regardless of higher moments of the noise. This corroborates our theoretical prediction that the noise covariance approximately determines the distribution of the solution. This is also in line with, and extends upon, observations in earlier work that the ratio of batch size to learning rate correlates with test error (Jastrzębski et al., 2017; He et al., 2019).

2. Related Work

Previous work has drawn connections between SGD noise and generalization (Mandt et al., 2016; Jastrzębski et al., 2017; He et al., 2019; Hoffer et al., 2017; Keskar et al., 2016). Notably, Mandt et al. (2016); He et al. (2019); Jastrzębski et al. (2017) analyze favorable properties of SGD noise by arguing that in the neighborhood of a local minimum, (2) is roughly the discretization of an Ornstein-Uhlenbeck (OU) process, and so the distribution of $y_{k\delta}$ approximates is approximately Gaussian. However, empirical results (Keskar et al., 2016; Hoffer et al., 2017) suggest that SGD generalizes better by finding better local minima, which may require us to look beyond the "OU near local minimum" assumption to understand the global distributional properties of SGD. Indeed, Hoffer et al. (2017) suggest that SGD performs a random walk on a random loss landscape, Kleinberg et al. (2018) propose that SGD noise helps smoooth out "sharp minima." Jastrzębski et al. (2017) further note the similarity between (1) and an Euler-Murayama approximation of (3). Chaudhari & Soatto (2018) also made connections between SGD and SDE. Our work tries to make these connections rigorous, by quantifying the error between (3), (2) and (1), without any assumptions about (3) being close to an OU process or being close to a local minimum.

Our work builds on a long line of work establishing the convergence rate of Langevin MCMC in different settings (Dalalyan, 2017; Durmus & Moulines, 2016; Ma et al., 2018; Gorham et al., 2016; Cheng et al., 2018; Erdogdu et al., 2018; Li et al., 2019). We will discuss our rates in relation to some of this work in detail following our presentation of Theorem 1. We note here that some of the techniques used in this paper were first used by Eberle (2011); Gorham et al. (2016), who analyzed the convergence of (3) to p^* without log-concavity assumptions. Erdogdu et al. (2018) studied processes of the form (2) as an approximation to (3) under a distant-dissipativity assumption, which is similar to the assumptions made in this paper. For the sequence (2), they prove an $O(1/\epsilon^2)$ iteration complexity to achieve ϵ integration error for any pseudo-Lipschitz loss f with polynomial growth derivatives up to fourth order. In comparison, we prove W_1 convergence between $\mathsf{Law}(y_{k\delta})$ and p^*, which is equivalent to $\sup_{\|\nabla f\|_\infty \leq 1} |\mathbb{E}\left[f(y_{k\delta})\right] - \mathbb{E}_{y\sim p^*}\left[f(y)\right]|$, also with rate $\tilde{O}(1/\epsilon^2)$. By smoothing the W_1 test function, we believe that the results by Erdogdu et al. (2018) can imply a qualitatively similar result to Theorem 1, but with a worse dimension and ϵ dependence.

In concurrent work by Li et al. (2019), the authors study a process based on a stochastic Runge-Kutta discretization scheme of (3). They prove an $\tilde{O}\left(\frac{d}{\epsilon^{-2/3}}\right)$ iteration complexity to achieve ϵ error in W_2 for an algorithm based on Runge-Kutta discretization of (3). They make a strong assumption of *uniform dissipativity* (essentially assuming that the process (3) is uniformly contractive), which is much stronger

than the assumptions in this paper, and may be violated in the settings of interest considered in this paper.

There has been a number of work (Chen et al., 2016; Li et al., 2018; Anastasiou et al., 2019) which establish CLT results for SGD with very small step size (rescaled to have constant variance). These work generally focus on the setting of "OU process near a local minimum", in which the diffusion matrix is constant.

Finally, a number of authors have studied the setting of heavy-tailed gradient noise in neural network training. (Zhang et al., 2019) showed that in some cases, the heavy-tailed noise can be detrimental to training, and a clipped version of SGD performs much better. (Simsekli et al., 2019) argue that when the SGD noise is heavy-tailed, it should not be modelled as a Gaussian random variable, but instead as an α-stable random variable, and propose a Generalized Central Limit Theorem to analyze the convergence in distribution. Our paper does not handle the setting of heavy-tailed noise; our theorems require that the norm of the noise term uniformly bounded, which will be satisfied, for example, if gradients are explicitly clipped at a threshold, or if the optimization objective has Lipschitz gradients and the SGD iterates stay within a bounded region.

3. Motivating Example

It is generally difficult to write down the invariant distribution of (3). In this section, we consider a very simple one-dimensional setting which does admit an explicit expression for p^*, and serves to illustrate some remarkable properties of anisotropic diffusion matrices.

Let us define $D(x) := M^2(x)$. Our analysis will be based on the Fokker-Planck equation, which states that p^* is the invariant distribution of (3) if

$$0 = \mathbf{div}(p^*(x)\nabla U(x)) + \mathbf{div}(p^*(x)\Gamma(x) + D(x)\nabla p^*(x)), \tag{4}$$

where $\Gamma(x)$ is a vector whose i^{th} coordinate equals $\sum_{j=1}^d \frac{\partial}{\partial x_j}[D(x)]_{i,j}$. In the one-dimensional setting, we can explicitly write down the density of $p^*(x)$. Note that in this case, $\Gamma(x) = \nabla D(x)$. Let $V(x) := \int_0^x \left(\frac{\nabla U(x)}{D(x)} + \frac{\nabla D(X)}{D(x)}\right)dx = \int_0^x \left(\frac{\nabla U(x)}{D(x)}\right)dx + \log D(x) - \log D(0)$. We can verify that $p^*(x) \propto e^{-V(x)}$ satisfies (4).

For a concrete example, let the potential $U(x)$ and the diffusion function $M(x)$ be defined as

$$U(x) := \begin{cases} \frac{1}{2}x^2, & \text{for } x \in [-1, 4] \\ \frac{1}{2}(x+2)^2 - 1, & \text{for } x \leq -1 \\ \frac{1}{2}(x-8)^2 - 16, & \text{for } x \geq 4 \end{cases}$$

$$M(x) = \begin{cases} \frac{1}{2}(x+2), & \text{for } x \in [-2, 8] \\ 1, & \text{for } x \leq -2 \\ 6, & \text{for } x \geq 8 \end{cases}.$$

We plot $U(x)$ in Figure 1a. Note that $U(x)$ has two local minima: a shallow minimum at $x = -2$ and a deeper minimum at $x = 8$. A plot of $M(x)$ can be found in Figure 1b. $M(x)$ is constructed to have increasing magnitude at larger values of x. This has the effect of biasing the invariant distribution towards smaller values of x.

We plot $V(x)$ in Figure 1c. Remarkably, $V(x)$ has only one local minimum at $x = -2$. The larger minimum of $U(x)$ at $x = 8$ has been smoothed over by the effect of the large diffusion $M(x)$. This is very different from when the noise is homogeneous (e.g., $M(x) = I$), in which case $p^*(x) \propto e^{-U(x)}$. We also simulate (3) (using (2)) for the given $U(x)$ and $M(x)$ for 1000 samples (each simulated for 1000 steps), and plot the histogram in Figure 1d.

4. Assumptions and Definitions

In this section, we state the assumptions and definitions that we need for our main results in Theorem 1 and Theorem 2.

Assumption A *We assume that $U(x)$ satisfies*

1. *The function $U(x)$ is continuously-differentiable on $\mathbb{R}^d$ and has Lipschitz continuous gradients; that is, there exists a positive constant $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\|_2 \leq L\|x - y\|_2$.*
2. *U has a stationary point at zero: $\nabla U(0) = 0$.*
3. *There exists a constant $m > 0, L_R, R$ such that for all $\|x - y\|_2 \geq R$,*

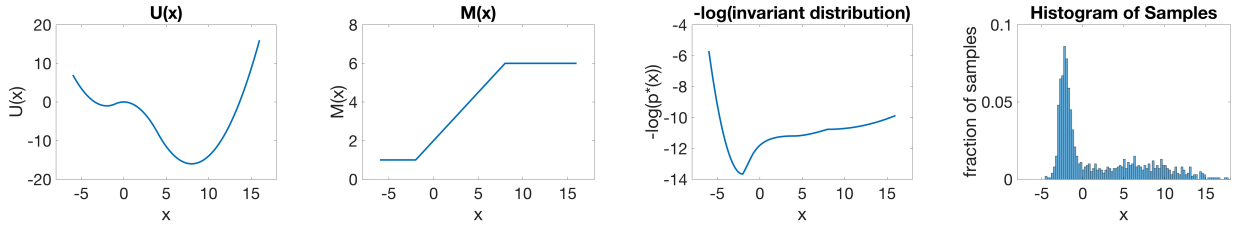
$$\langle \nabla U(x) - \nabla U(y), x - y\rangle \geq m\|x - y\|_2^2. \tag{5}$$

and for all $\|x - y\|_2 \leq R$, $\|\nabla U(x) - \nabla U(y)\|_2 \leq L_R\|x - y\|_2$.

Remark 1 *This assumption, and minor variants, is common in the nonconvex sampling literature (Eberle, 2011; 2016; Cheng et al., 2018; Ma et al., 2018; Erdogdu et al., 2018; Gorham et al., 2016).*

Assumption B *We make the following assumptions on ξ and M:*

1. *For all x, $\mathbb{E}[\xi(x, \eta)] = 0$.*
2. *For all x, $\|\xi(x, \eta)\|_2 \leq \beta$ almost surely.*



(a) $U(x)$ (b) $M(x)$ (c) $V(x)$ (d) Samples

Figure 1. One-dimensional example exhibiting the importance of state-dependent noise: A simple construction showing how $M(x)$ can affect the shape of the invariant distribution. While $U(x)$ has two local minima, $V(x)$ only has the smaller minimum at $x = -2$. Figure 1d represents samples obtained from simulating using the process (2). We can see that most of the samples concentrate around $x = -2$.

3. *For all x, y, $\|\xi(x,\eta) - \xi(y,\eta)\|_2 \leq L_\xi \|x - y\|_2$ almost surely.*
4. *There is a positive constant c_m such that for all x, $2c_m I \prec M(x)$.*

Remark 2 *We discuss these assumptions in a specific setting in Section 6.2.*

For convenience we define a matrix-valued function $N(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$:

$$N(x) := \sqrt{M(x)^2 - c_m^2 I}. \tag{6}$$

Under Assumption A, we can prove that $N(x)$ and $M(x)$ are bounded and Lipschitz (see Lemma 15 and 16 in Appendix D). These properties will be crucial in ensuring convergence.

Given an arbitrary sample space Ω and any two distribution $p \in \mathscr{P}(\Omega)$ and $q \in \mathscr{P}(\Omega)$, a joint distribution $\zeta \in \mathscr{P}(\Omega \times \Omega)$ is a *coupling* between p and q if its marginals are equal to p and q respectively.

For a matrix, we use $\|G\|_2$ to denote the operator norm: $\|G\|_2 = \sup_{v\in\mathbb{R}^d, \|v\|_2=1} \|Gv\|_2$..

Finally, we define a few useful constants which will be used throughout the paper:

$$\begin{aligned} L_N &:= \frac{4\beta L_\xi}{c_m}, \quad \alpha_q := \frac{L_R + L_N^2}{2c_m^2}, \\ \mathcal{R}_q &:= \max\left\{R, \frac{16\beta^2 L_N}{m \cdot c_m}\right\} \\ \lambda &:= \min\left\{\frac{m}{2}, \frac{2c_m^2}{32{\mathcal{R}_q}^2}\right\} \exp\left(-\frac{7}{3}\alpha_q {\mathcal{R}_q}^2\right). \end{aligned} \tag{7}$$

L_N is the smoothness parameter of the matrix $N(x)$, and we show in Lemma 16 that $\mathrm{tr}\Big((N(x) - N(y))^2\Big) \leq L_N^2 \|x - y\|_2^2$. The constants α_q and $\mathcal{R}_q$ are used to define a Lyapunov function q in Appendix E.1. A key step in our proof uses the fact that, under the dynamics (2), q contracts at a rate of $e^{-\lambda}$, plus discretization error.

5. Main Results

In this section, we present our main convergence results beginning with convergence under Gaussian noise and proceeding to the non-Gaussian case.

Theorem 1 *Let x_t and y_t have dynamics as defined in (3) and (2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|x_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \leq R^2 + \beta^2/m$. Let $\hat{\epsilon}$ be a target accuracy satisfying $\hat{\epsilon} \leq \left(\frac{16(L+L_N^2)}{\lambda}\right) \cdot \exp\left(7\alpha_q \mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q {\mathcal{R}_q}^2 + 1}$. Let δ be a step size satisfying*

$$\delta \leq \min\left\{ \frac{\lambda^2 \hat{\epsilon}^2}{512\beta^2\left(L^2 + L_N^4\right)\exp\left(\frac{14\alpha_q {\mathcal{R}_q}^2}{3}\right)}, \frac{2\lambda\hat{\epsilon}}{\left(L^2 + L_N^4\right)\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\sqrt{R^2 + \beta^2/m}} \right\}.$$

If we assume that $x_0 = y_0$, then there exists a coupling between x_t and y_t such that for any k,

$$\mathbb{E}\left[\|x_{k\delta} - y_{k\delta}\|_2\right] \leq \hat{\epsilon}$$

Alternatively, if we assume $n \geq \frac{3\alpha_q {\mathcal{R}_q}^2}{\delta} \log \frac{R^2 + \beta^2/m}{\hat{\epsilon}}$, then

$$W_1(p^*, p^y_{n\delta}) \leq 2\hat{\epsilon}$$

where $p_t^y := \mathsf{Law}(y_t)$.

Remark 3 *Note that m, L, R are from Assumption A, L_N is from (7), c_m, β, L_ξ are from Assumption B).*

Remark 4 *Finding a suitable y_0 can be done very quickly using gradient descent wrt $U(\cdot)$. The convergence rate to the ball of radius R is very fast, due to Assumption A.3.*

After some algebraic simplifications, we see that for a sufficiently small $\hat{\epsilon}$, achieving $W_1(p^y_{n\delta}, p^*) \leq \hat{\epsilon}$ requires number of steps

$$n = \tilde{O}\left(\frac{\beta^2}{\hat{\epsilon}^2} \cdot \exp\left(\frac{14}{3} \cdot \left(\frac{L_R}{c_m^2} + \frac{16\beta^2 L_\xi^2}{c_m^4}\right) \cdot \max\left\{R^2, \frac{2^{12}\beta^6 L_\xi^2}{m^2 c_m^4}\right\}\right)\right).$$

Remark 5 *The convergence rate contains a term e^{R^2}; this term is also present in all of the work cited in the previous section under Remark 1. Given our assumptions, in particular 5, this dependence is unavoidable as it describes the time to transit between two modes of the invariant distribution. It can be verified to be tight by considering a simple double-well potential.*

Remark 6 *As illustrated in Section 6.2, the m from Assumption B.3 should be thought of as a regularization term which can be set arbitrarily large. In the following discussion, we will assume that $\max\left\{R^2, \frac{\beta^6 L_\xi^2}{m^2 c_m^4}\right\}$ is dominated by the R^2 term.*

To gain intuition about this term, let's consider what it looks like under a sequence of increasingly weaker assumptions:

a. Strongly convex, constant noise: $U(x)$ m-strongly convex, L-smooth, $\xi(x, \eta) \sim \mathcal{N}(0, I)$ for all x. (In reality we need to consider a truncated Gaussian so as not to violate Assumption B.2, but this is a minor issue). In this case, $L_\xi = 0$, $c_m = 1$, $R = 0$, $\beta = \tilde{O}(\sqrt{d})$, so $k = O(\frac{d}{\hat{\epsilon}^2})$. This is the same rate as obtained by Durmus & Moulines (2016). We remark that Durmus & Moulines (2016) obtain a W_2 bound which is stronger than our W_1 bound.

b. Non-convex, constant noise: $U(x)$ not strongly convex but satisfies Assumption A, and $\xi(x, \eta) \sim \mathcal{N}(0, I)$. In this case, $L_\xi = 0$, $c_m = 1$, $\beta = \tilde{O}(\sqrt{d})$ This is the setting studied by Cheng et al. (2018) and Ma et al. (2018). The rate we recover is $k = \tilde{O}\left(\frac{d}{\epsilon^2} \cdot \exp\left(\frac{14}{3}LR^2\right)\right)$, which is in line with Cheng et al. (2018), and is the best W_1 rate obtainable from Ma et al. (2018).

c. Non-convex, state-dependent noise: $U(x)$ satisfies Assumption A, and ξ satisfies Assumption B. To simplify matters, suppose the problem is rescaled so that $c_m = 1$. Then the main additional term compared to setting b. above is $\exp\left(\frac{64\beta^2 L_\xi^2 R^2}{c_m^4}\right)$. This suggests that the effect of a L_ξ-Lipschitz noise can play a similar role in hindering mixing as a L_R-Lipschitz nonconvex drift.

When the dimension is high, computing $M(y_k)$ can be difficult, but if for each x, one has access to samples whose covariance is $M(x)$, then one can approximate $M(y_k)\theta_k$ via the central limit theorem by drawing a sufficiently large number of samples. The proof of Theorem 1 can be readily modified to accommodate this (see Appendix A.5).

We now turn to the non-Gaussian case.

Theorem 2 *Let x_t and w_t have dynamics as defined in (3) and (1) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|x_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|w_0\|_2^2\right] \leq R^2 + \beta^2/m$. Let $\hat{\epsilon}$ be a target accuracy satisfying $\hat{\epsilon} \leq \left(\frac{16(L+L_N^2)}{\lambda}\right) \cdot \exp\left(7\alpha_q \mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. Let $\epsilon := \frac{\lambda}{16(L+L_N^2)} \exp\left(-\frac{7\alpha_q \mathcal{R}_q{}^2}{3}\right)\hat{\epsilon}$. Let $T := \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\epsilon}{32\sqrt{L}\beta}, \frac{\epsilon^2}{128\beta^2}, \frac{\epsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$ and let δ be a step size satisfying*

$$\delta \leq \min\left\{\frac{T\epsilon^2 L}{36d\beta^2 \log\left(\frac{36d\beta^2}{\epsilon^2 L}\right)}, \frac{T\epsilon^4 L^2}{2^{14}d\beta^4 \log\left(\frac{2^{14}d\beta^4}{\epsilon^4 L^2}\right)}\right\}.$$

If we assume that $x_0 = w_0$, then there exists a coupling between x_t and w_t such that for any k,

$$\mathbb{E}\left[\|x_{k\delta} - w_{k\delta}\|_2\right] \leq \hat{\epsilon}.$$

Alternatively, if we assume that $n \geq \frac{3\alpha_q \mathcal{R}_q{}^2}{\delta} \cdot \log\frac{R^2+\beta^2/m}{\hat{\epsilon}}$, then

$$W_1(p^*, p^w_{n\delta}) \leq 2\hat{\epsilon},$$

where $p^w_t := \mathsf{Law}(w_t)$.

Remark 7 *To achieve $W_1(p^*, p^w_{n\delta}) \leq \hat{\epsilon}$, the number of steps needed is of order $n = \tilde{O}\left(\frac{1}{\hat{\epsilon}^8} \cdot e^{29\alpha_q \mathcal{R}_q{}^2}\right)$. The $\hat{\epsilon}$ dependency is considerably worse than in Theorem 1. This is because we need to take many steps of (1) in order to approximate a single step of (2). For details, see the coupling construction in equations (27)–(31) of Appendix B.*

6. Application to Stochastic Gradient Descent

In this section, we will cast SGD in the form of (1). We consider an objective of the form

$$U(w) = \frac{1}{n}\sum_{i=1}^{n} U_i(w). \tag{8}$$

We reserve the letter η to denote a random minibatch from $\{1, \ldots, n\}$, sampled with replacement, and define $\zeta(w, \eta)$ as follows:

$$\zeta(w, \eta) := \nabla U(w) - \frac{1}{|\eta|}\sum_{i \in \eta} \nabla U_i(w) \tag{9}$$

For a sample of size one, i.e. $|\eta| = 1$, we define

$$H(w) := \mathbb{E}\left[\zeta(w,\eta)\zeta(w,\eta)^T\right] \tag{10}$$

as the covariance matrix of the difference between the true gradient and a single sampled gradient at w. A standard run of SGD, with minibatch size $b := |\eta_k|$, then has the following form:

$$\begin{aligned} w_{k+1} &= w_k - \delta\frac{1}{b}\sum_{i\in\eta_k}\nabla U_i(w_k) \\ &= w_k - \delta\nabla U(w_k) + \sqrt{\delta}\Big(\sqrt{\delta}\zeta(w_k,\eta_k)\Big). \end{aligned} \tag{11}$$

We refer to an SGD algorithm with step size δ and minibatch size b a (δ, b)-SGD. Notice that (11) is in the form of (1), with $\xi(w,\eta) = \sqrt{\delta}\zeta(w,\eta)$. The covariance matrix of the noise term is

$$\mathbb{E}\left[\xi(w,\eta)\xi(w,\eta)^T\right] = \frac{\delta}{b}H(w). \tag{12}$$

Because the magnitude of the noise covariance scales with $\sqrt{\delta}$, it follows that as $\delta \to 0$, (11) converges to deterministic gradient flow. However, the loss of randomness as $\delta \to 0$ is not desirable as it has been observed that as SGD approaches GD, through either small step size or large batch size, the generalization error goes up (Jastrzębski et al., 2017; He et al., 2019; Keskar et al., 2016; Hoffer et al., 2017); this is also consistent with our experimental observations in Section 6.3.1.

Therefore, a more meaningful way to take the limit of SGD is to hold the noise term constant in (11). More specifically, we define the ***constant-noise limit*** of (11) as

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t, \tag{13}$$

where $M(x) := \sqrt{\frac{\delta}{b}H(x)}$. Note that this is in the form of (3), with noise covariance $M(x_t)^2$ matching that of SGD in (11). Using Theorem 2, we can bound the W_1 distance between the SGD iterates w_k from (11), and the continuous-time SDE x_t from (13).

6.1. Importance of Noise Covariance

We highlight the fact that the limiting SDE of a discrete process,

$$w_{k+1} = w_k - s\nabla U(w_k) + \sqrt{s}\xi(w_k,\eta_k), \tag{14}$$

depends only on the covariance matrix of ξ. More specifically, as long as ξ satisfies $\sqrt{\mathbb{E}\left[\xi(w,\eta)\xi(w,\eta)^T\right]} = M(w)$, (14) will have (13) as its limiting SDE, *regardless of higher moments of* ξ. This fact, combined with Theorem 2, means that in the limit of $\delta \to 0$ and $k \to \infty$, the distribution of w_k will be determined by the covariance of ξ alone. An immediate consequence is the following: *at convergence, the test performance of any Langevin MCMC-like algorithm is almost entirely determined by the covariance of its noise term.*

Returning to the case of SGD algorithms, since the noise covariance is $M(x)^2 = \frac{\delta}{b}H(x)$ (see (12)), we know that the ratio of step size δ to batch size b is an important quantity which can dictate the test error of the algorithm; this observation has been made many times in prior work (Jastrzębski et al., 2017; He et al., 2019), and our results in this paper are in line with these observations. Here, we move one step further, and provide experimental evidence to show that more fundamentally, it is the noise covariance in the constant-noise limit that controls the test error.

To verify this empirically, we propose the following algorithm called *large-noise SGD*.

Definition 1 *An (s, σ, b_1, b_2)-large-noise SGD is an algorithm that aims to minimize (8) using the following updates:*

$$\begin{aligned} w_{k+1} = w_k &- \frac{s}{b_1}\sum_{i\in\eta_k}\nabla U_i(w_k) \\ &+ \frac{\sigma\sqrt{s}}{b_2}\left(\sum_{i\in\eta_k'}\nabla U_i(w_k) - \sum_{i\in\eta_k''}\nabla U_i(w_k)\right), \end{aligned} \tag{15}$$

where η_k, η_k', and η_k'' are minibatches of sizes b_1, b_2, and b_2, sampled uniformly at random from $\{1,\dots,n\}$ with replacement. The three minibatches are sampled independently and are also independent of other iterations.

Intuitively, an (s, σ, b_1, b_2)-large-noise SGD should be considered as an SGD algorithm with step size s and minibatch size b_1 and an additional noise term. The noise term computes the difference of two independent and unbiased estimates of the full gradient $\nabla U(w_k)$, each using a batch of b_2 data points. Using the definition of ζ in (9), we can verify that the update (15) is equivalent to

$$\begin{aligned} w_{k+1} = w_k &- s\nabla U(w_k) + s\zeta(w_k,\eta_k) \\ &+ \sigma\sqrt{s}(\zeta(w_k,\eta_k'') - \zeta(w_k,\eta_k')), \end{aligned} \tag{16}$$

which is in the form of (1), with

$$\xi(w,\tilde{\eta}) = \sqrt{s}\zeta(w,\eta) + \sigma(\zeta(w,\eta'') - \zeta(w,\eta')), \tag{17}$$

where $\tilde{\eta} = (\eta,\eta',\eta'')$, and $|\eta| = b_1$, $|\eta'| = |\eta''| = b_2$. Further, the noise covariance matrix is

$$\mathbb{E}\left[\xi(w,\tilde{\eta})\xi(w,\tilde{\eta})^T\right] = (\frac{s}{b_1} + \frac{2\sigma^2}{b_2})H(w). \tag{18}$$

Therefore, if we have

$$\frac{s}{b_1} + \frac{2\sigma^2}{b_2} = \frac{\delta}{b}, \tag{19}$$

then an (s, σ, b_1, b_2)-large-noise SGD should have the same noise covariance as a (δ, b)-SGD *(but very different higher noise moments due to the injected noise)*, and based on our theory, the large-noise SGD should have similar test error to that of the SGD algorithm, even if the step size and batch size are different. In Section 6.3, we verify this experimentally. We stress that we are not proposing the large-noise SGD as a practical algorithm. The reason that this algorithm is interesting is that it gives us a family of $(w_k)_{k=1,2,\ldots}$ which converges to (13), and is implementable in practice. Thus this algorithm helps us uncover the importance of noise covariance (and the unimportance of higher noise moments) in Langevin MCMC-like algorithms. We also remark that Hoffer et al. (2017) proposed a different way of injecting noise, multiplying the sampled gradient with a suitably scaled Gaussian noise.

6.2. Satisfying the Assumptions

Before presenting the experimental results, we remark on a particular way that a function $U(w)$ defined in (8), along with the stochastic sequence w_k defined in (15), can satisfy the assumptions in Section 4.

Suppose first that we shift the coordinate system so that $\nabla U(0) = 0$. Let us additionally assume that for each i, $U_i(w)$ has the form

$$U_i(w) = U_i'(w) + V(w),$$

where $V(w) := m(\|x\|_2 - R/2)^2$ is a m-strongly convex regularizer outside a ball of radius R, and each $U_i'(w)$ has L_R-Lipschitz gradients. Suppose further that $m \geq 4 \cdot L_R$. These additional assumptions make sense when we are only interested in $U(w)$ over $B_R(0)$, so $V(w)$ plays the role of a barrier function that keeps us within $B_R(0)$. Then, it can immediately be verified that $U(w)$ satisfies Assumption A with $L = m + L_R$.

The noise term ξ in (17) satisfies Assumption B.1 by definition, and satisfies Assumption B.3 with $L_\xi = (\sqrt{s} + 2\sigma)L$. Assumption B.2 is satisfied if $\zeta(w, \eta)$ is bounded for all w, i.e. the sampled gradient does not deviate from the true gradient by more than a constant. We will need to assume directly Assumption B.4, as it is a property of the distribution of $\nabla U_i(w)$ for $i = 1, \ldots, n$.

6.3. Experiments

In this section, we present experimental results that validate the importance of noise covariance in predicting the test error of Langevin MCMC-like algorithms. In all experiments, we use two different neural network architectures on the CIFAR-10 dataset (Krizhevsky & Hinton, 2009) with the standard test-train split. The first architecture is a simple convolutional neural network, which we call CNN in the following,[1] and the other is the VGG19 network (Simonyan & Zisserman, 2014). To make our experiments consistent with the setting of SGD, we do not use batch normalization or dropout, and use constant step size. In all of our experiments, we run SGD algorithm 2000 epochs such that the algorithm converges sufficiently. Since in most of our experiments, the accuracies on the training dataset are almost 100%, we use the test accuracy to measure the generalization performance.

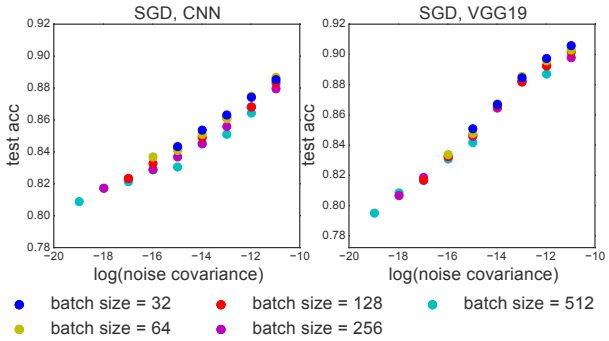


Figure 2. Relationship between test accuracy and the noise covariance of SGD algorithm. In each plot, the dots with the same color correspond to SGD runs with the same batch size but different step sizes.

Recall that according to (12) and (18), for both SGD and large-noise SGD, the noise covariance is a scalar multiple of $H(w)$. For simplicity, in the following, we will slightly abuse our terminology and call this scalar the *noise covariance*; more specifically, for (δ, b)-SGD, the noise covariance is δ/b, and for an (s, σ, b_1, b_2)-large-noise SGD, the noise covariance is $\frac{s}{b_1} + \frac{2\sigma^2}{b_2}$.

6.3.1. Accuracy vs Noise Covariance

In our first experiment, we focus on the SGD algorithm, and show that there is a positive correlation between the noise covariance and the final test accuracy of the trained model. One major purpose of this experiment is to establish baselines for our experiments on large-noise SGD.

We choose constant step size δ from

$$\{0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128\}$$

and minibatch size b from $\{32, 64, 128, 256, 512\}$. For each (step size, batch size) pair, we plot its final test accuracy against its noise covariance in Figure 2. From the plot, we can see that higher noise covariance leads to better final test accuracy, and there is a linear trend between the test accuracy and the logarithm. We also highlight the fact that conditioned on the noise covariance, the test accuracy is not significantly correlated with either the step size or the

[1] We provide details of this CNN architecture in Appendix G.

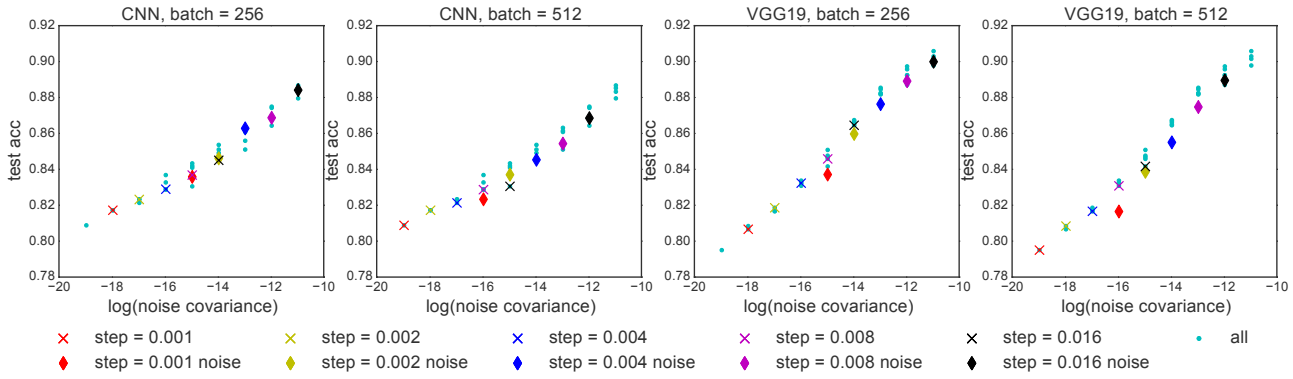


Figure 3. Large-noise SGD. Small dots correspond to all the baseline SGD runs in Figure 2. Each $\times$ corresponds to a baseline SGD run whose step size is specified in the legend and batch size is specified in the title. Each $\diamond$ corresponds to a large-noise SGD run whose noise covariance is 8 times of that of the $\times$ with the same color. As we can see, injecting noise improves test accuracy, and the large-noise SGD runs fall close to the linear trend.

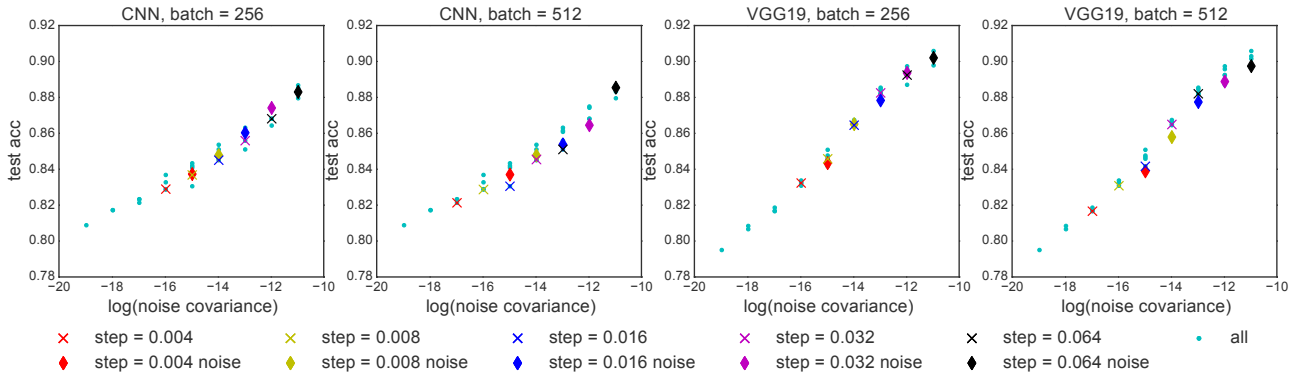


Figure 4. Large-noise SGD. Batch size in the titles represents the batch size of $\times$ runs. Each $\diamond$ corresponds to a large-noise SGD run whose noise covariance matches that of a baseline SGD run whose step size is the same as the $\times$ run with the same color and batch size is 128. Again, large-noise SGD falls close to the linear trend.

minibatch size. In other words, similar to the observations in prior work (Jastrzębski et al., 2017; He et al., 2019), there is a strong correlation between relative variance of an SGD sequence and its test accuracy, regardless of the combination of minibatch size and step size.

6.3.2. Large-Noise SGD

In this section, we implement and examine the performance of the large-noise SGD algorithm proposed in (15). We select a subset of SGD runs with relatively small noise covariance in the experiment in the previous section (we call them *baseline SGD runs*), and implement large-noise SGD by injecting noise. Our goal is to see, for a particular noise covariance, whether large-noise SGD has test accuracy that is similar to SGD, *in spite of significant differences in third-and-higher moments of the noise in large-noise SGD compared to standard SGD*.

Our first experiment is to add noise with the same minibatch size to the (δ, b) baseline SGD run such that the new noise covariance matches that of an $(8\delta, b)$-SGD (an SGD run with larger step size). In other words, we implement $(\delta, \sqrt{7\delta/2}, b, b)$-large-noise SGD, whose noise covariance is 8 times of that of the baseline. Our results are shown in Figure 3. Our second experiment is similar: we add noise with minibatch size 128 to the (δ, b) baseline SGD run with $b \in \{256, 512\}$ such that the new noise covariance matches that of a $(\delta, 128)$-SGD (an SGD run with smaller batch size). More specifically, we implement $(\delta, \sqrt{\frac{1}{2}(1 - \frac{128}{b})\delta}, b, 128)$-large-noise SGD runs. The results are shown in Figure 4. In these figures, each $\times$ denotes a baseline SGD run, with step size specified in the legend and minibatch size specified by plot title. For each baseline SGD run, we have a corresponding large-noise SGD run, denoted by $\diamond$ with the same color. As mentioned, these $\diamond$ runs are designed to match the noise covariance of SGD with larger step size or smaller batch size. In addition to $\times$

and $\diamond$, we also plot using a small teal marker all the other runs from Section 6.3.1. This helps highlight the linear trend between the logarithm of noise covariance and test accuracy that we observed in Section 6.3.1.

As can be seen, the (noise variance, test accuracy) values for the $\diamond$ runs fall close to the linear trend. More specifically, a run of large-noise SGD produces similar test accuracy to vanilla SGD runs with the same noise variance. We highlight two potential implications: First, just like in Section 6.3.1, we observe that the test accuracy strongly correlates with relative variance, even for noise of the form (17), which can have rather different higher moments than ζ (standard SGD noise); Second, since the $\diamond$ points fall close to the linear trend, we hypothesize that the constant-noise limit SDE (13) should also have similar test error. If true, then this implies that we only need to study the potential $U(x)$ and noise covariance $M(x)$ to explain the generalization properties of SGD.

7. Acknowledgements

We wish to acknowledge support by the Army Research Office (ARO) under contract W911NF-17-1-0304 under the Multidisciplinary University Research Initiative (MURI).

References

Anastasiou, A., Balasubramanian, K., and Erdogdu, M. A. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. *arXiv preprint arXiv:1904.02130*, 2019.

Chatzigeorgiou, I. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Cheng, X., Bartlett, P. L., and Jordan, M. I. Quantitative central limit theorems for discrete stochastic processes. *arXiv preprint arXiv:1902.00832*, 2019.

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, 2017.

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

Eberle, A. Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349 (19-20):1101–1104, 2011.

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3-4): 851–886, 2016.

Eldan, R., Mikulincer, D., and Zhai, A. The CLT in high dimensions: quantitative bounds via martingale embedding. *arXiv preprint arXiv:1806.09087*, 2018.

Erdogdu, M. A., Mackey, L., and Shamir, O. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pp. 9671–9680, 2018.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. Measuring sample quality with diffusions. *arXiv preprint arXiv:1611.06972*, 2016.

He, F., Liu, T., and Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, pp. 1141–1150, 2019.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pp. 1731–1741, 2017.

Jastrzębski, 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.

Karatzas, I. and Shreve, S. E. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pp. 47–127. Springer, 1998.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kleinberg, R., Li, Y., and Yuan, Y. An alternative view: When does SGD escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Li, T., Liu, L., Kyrillidis, A., and Caramanis, C. Statistical

inference using sgd. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Li, X., Wu, Y., and Mackey, L. Stochastic Runge-Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, pp. 7746–7758, 2019.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

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

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

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