Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent

Kangqiao Liu *1 Liu Ziyin *1 Masahito Ueda 1,2,3

Abstract

In the vanishing learning rate regime, stochastic gradient descent (SGD) is now relatively well understood. In this work, we propose to study the basic properties of SGD and its variants in the non-vanishing learning rate regime. The focus is on deriving exactly solvable results and discussing their implications. The main contributions of this work are to derive the stationary distribution for discrete-time SGD in a quadratic loss function with and without momentum; in particular, one implication of our result is that the fluctuation caused by discrete-time dynamics takes a distorted shape and is dramatically larger than a continuous-time theory could predict. Examples of applications of the proposed theory considered in this work include the approximation error of variants of SGD, the effect of minibatch noise, the optimal Bayesian inference, the escape rate from a sharp minimum, and the stationary covariance of a few second-order methods including damped Newton’s method, natural gradient descent, and Adam.

1. Introduction

Behind the success of deep learning lies the simple optimization methods such as stochastic gradient descent (SGD) (Bottou, 1998; Sutskever et al., 2013; Dieuleveut et al., 2020; Mori and Ueda, 2020a) and its variants (Duchi et al., 2011; Flammarion and Bach, 2015; Kingma and Ba, 2015), which are used for neural network optimization. Despite the empirical efficiency of SGD, our theoretical understanding of SGD is still limited. Two types of noises for SGD are studied. When the noise is white, the dynamics is governed by the stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011). When the noise is due to minibatch sampling, the noise is called the SGD noise or minibatch noise. So far, nearly all the theoretical attempts at understanding the noise in SGD have adopted the continuous-time approximation by assuming a vanishingly small learning rate (Mandt et al., 2017; Li et al., 2017; Jastrzebski et al., 2017; Chaudhari and Soatto, 2018; Zhu et al., 2019; Xie et al., 2021). This amounts to making an analogy to the diffusion theory in physics (Einstein, 1905; Van Kampen, 1992), and helps understand some properties of SGD and deep learning such as the flatness of the minima selected by training (Jastrzebski et al., 2017; Chaudhari and Soatto, 2018; Smith and Le, 2018; Xie et al., 2021). However, in reality, a large learning rate often leads to qualitatively distinct behavior, including reduced training time and potentially better generalization performance (Keskar et al., 2017; Li et al., 2019). The present work is motivated by the fact that the existing continuous theory is insufficient to describe and predict the properties and phenomena of learning in this large learning rate regime. In fact, we will show that the prediction by continuous theory may deviate arbitrarily from the experimental result.

In this work, we study the stationary-state solutions of discrete-time update rules of SGD. The result can be utilized to analyze SGD without invoking the unrealistic assumption of a vanishingly small learning rate. Specifically, our contributions are:

1. We exactly solve the discrete-time update rules for SGD and its variants with momentum in a local minimum to obtain the analytic form of the covariance matrix of the model parameters at asymptotic time.
2. We apply our results to various settings that have been studied in the continuous-time limit, such as finding the optimal learning rate in a Bayesian setting, understanding the escape from a sharp minimum, and the approximation error of various variants of SGD.
3. Compared with the continuous-time theories, our work requires fewer assumptions and finds significantly improved agreement with experimental results.

In section 2, we present the background of this work. Related works are discussed in section 3. In section 4, we derive our main theoretical results for SGD and its momentum variant. In section 5, we verify our theoretical results.
When noise exists, SGD iteration converges not to a point but a distribution.

In section 6, we apply our solution to some well-known problems that have previously been investigated in the continuous-time limit. A summary of our results is given in Table 1.

2. Background

In the presence of noise, it is difficult to give a simple solution to the discrete dynamics. Under the assumptions of a constant Gaussian noise, a quadratic loss function, and an infinitesimal learning rate, theorists approximate the multidimensional update rule by a continuous-time stochastic differential equation (Mandt et al., 2017; Zhu et al., 2019), which is a multivariate Ornstein–Uhlenbeck process, $\text{d}w_t = -\lambda K w_t \text{d}t + \lambda C w_t \text{d}W_t$, where $C$ is the covariance matrix of the noise, $W_t$ represents a standard multi-dimensional Brownian motion, and $K$ is the Hessian of the local minimum. The stationary covariance matrix, $\Sigma := \lim_{t \to \infty} \text{cov}(w_t^T w_t)$, is found to satisfy the following matrix equation (Van Kampen, 1992):

$$\Sigma K + K \Sigma = \frac{\lambda}{1 - \mu} C,$$  

(1)

where $\mu$ is the momentum hyperparameter; when no momentum is used, $\mu = 0$. Despite the fact that a number of theoretical works has been performed on the basis of the above constant Gaussian noise and the continuous-time approximation (Chaudhari and Soatto, 2018; Jacot et al., 2018; Zhu et al., 2019; Lee et al., 2020; Xie et al., 2021), it is clear that the stationary distribution given by Eq. (1) substantially deviates from the true one obtained in experiments.

The predictions based on these results are qualitatively acceptable only in a small learning rate regime. For a large learning rate, the assumptions simply break down so that the theory becomes invalid.

To intuitively understand how a large learning rate makes a difference, we consider a model with a single parameter $w \in \mathbb{R}$ in a quadratic potential $L = \frac{1}{2}kw^2$ with $k > 0$. SGD obeys the dynamical equations as follows:

$$\begin{align*}
g_t &= kw_{t-1} + \eta_t; \\
w_t &= w_{t-1} - \lambda g_t,
\end{align*}$$

(2)

where $\lambda$ is the learning rate and $\eta_t$ is a normal random variable with zero mean and variance $\sigma^2$. When $\sigma^2 = 0$, the dynamics is deterministic, and the common approach is to assume that $\lambda \ll 1/k$ such that one may take the continuous-time limit of this equation as $\dot{w} = -\lambda kw(t)$, which solves to give $w = w_0 e^{-\lambda t}$. However, this continuous approximation fails when $\lambda$ is large. To see this, we note that the deterministic discrete-time dynamics solves to give

$$w_t = (1 - \lambda k)^t w_0, \quad t \in \mathbb{N},$$

(3)

which is an exponential decay when $\lambda < \frac{1}{k}$, and, in this region, the standard continuous-time dynamics is valid with error $O((1 - \lambda k)^2)$. On the other hand, when $\lambda > \frac{2}{k}$, the learning is so large that the parameter $w$ will diverge; therefore, the interesting region is when $\frac{1}{k} < \lambda < \frac{2}{k}$, where the dynamics is convergent, yet a simple continuous approximation fails. See a schematic illustration in Figure 1. It is therefore urgent to develop a theory that can handle SGD with a large learning rate.

3. Related Works

Large Learning Rate. Although continuous-time theory has been the dominant theoretical approach, Lewkowycz et al. (2020) took a step forward in understanding why a larger learning rate may generalize better. They characterized SGD into three regimes according to the learning rate. They conjectured that a rather large initial learning rate leads to a “catapult phase”, which helps exploration and often leads to better generalization by converging to a flatter minimum. However, their work is mostly empirical and in the noise-free regime. There are more empirical works on large learning rate and generalization. LeCun et al. (2012) found that a large batch size (or a small learning rate) usually leads to reduced generalization performance. Keskar et al. (2017) proposed an empirical measure based on the sharpness of a minimum. They presented numerical evidence that a small learning rate prefers sharp minima that generalize poorly. Goyal et al. (2017) showed that setting the learning rate proportional to the minibatch size ensures good generalization, which is crucial for large scale training. There are also works about how noise, the batch size and the learning rate influence the generalization (Hoffer et al., 2017; Mori and Ueda, 2020a;b). The explanations

![Figure 1](image1.png)
of why a flatter minimum generalizes better are given by some theories such as the minimum description length theory (Rissanen, 1983), a Bayesian view of learning (MacKay, 1992), and the Gibbs free energy (Chaudhari et al., 2019).

### Escape from Sharp Minimum
Theoretically understanding why and how SGD converges to flat minima is crucially important (Hochreiter and Schmidhuber, 1997). In fact, among many complexity measures characterizing generalization (Dziugaite and Roy, 2017; Neyshabur et al., 2018a; Smith and Le, 2018; Chaudhari and Soatto, 2018; Keskar et al., 2017; Liang et al., 2019; Nagarajan and Kolter, 2019), the sharpness-based measures have been shown to be the best up to now (Jiang et al., 2019). The continuous approximation is usually adopted to understand how SGD chooses flat minima. Hu et al. (2019) used diffusion theory to show that escape is easier from a sharp minimum than from a flat one. Wu et al. (2018) studied the relationship among learning rate, batch size and generalization from the perspective of dynamical stability. Iastrzebski et al. (2017) used stochastic differential equations to prove that the higher the ratio of the learning rate to the batch size, the flatter minimum will be selected. Zhu et al. (2019) defined the escape efficiency for a minimum and obtained its explicit expression using diffusion theory. They show that anisotropic noise helps escape from sharp minima effectively. A recent work by Xie et al. (2021) calculated the escape rate from a minimum by adopting the formalism of the Kramers escape rate in physics (Kramers, 1940). It is shown that SGD with minibatch noise favors flatter minima. Our exact results for a large learning rate make it possible to study the selected flatness and the complexity measures more accurately and may enhance our understanding of deep learning.

### Bayesian Inference
SGD has been used for Bayesian inference as well. In Bayesian inference, one assumes a probabilistic model $p(w, x)$ with data $x$ and hidden parameter $w$. The goal is to approximate the posterior $p(w | x)$. Traditionally, stochastic gradient Markov Chain Monte Carlo (MCMC) methods have been used (Welling and Teh, 2011; Ma et al., 2015). A similarity between SGD and MCMC suggests the possibility of SGD being used as approximate Bayesian inference. Mandt et al. (2017) applied SGD to minimize the loss function defined as $- \ln p(w, x)$. They show that one can tune the learning rate such that the Kullback-Leibler (KL) divergence between the learned distribution by SGD and the posterior is minimized. This means that SGD can be regarded as an approximate Bayesian inference. However, for a large learning rate, their assumption is no more valid.

### 4. Theory of Discrete-Time SGD
We propose to deal with the discrete-time SGD directly. We use $w \in \mathbb{R}^D$ to denote the weights of the model viewed as a vector, and the boldness is dropped when $D = 1$. We use capital $K \in \mathbb{R}^{D \times D}$ to denote the Hessian matrix of the quadratic loss function; when $D = 1$, we use the lower-case $k$. We use $\Sigma \in \mathbb{R}^{D \times D}$ to denote the asymptotic covariance matrix of $w$, and $C \in \mathbb{R}^{D \times D}$ to denote the covariance matrix of a general type of noise $\eta \in \mathbb{R}^D$. When the learning rate $\lambda$ is not a scalar but a matrix (sometimes called preconditioning matrix), we use the upper-case letter $\Lambda \in \mathbb{R}^{D \times D}$ instead of $\lambda$. $L$ denotes the training loss function, and $S$ the minibatch size. We use $E[\cdot]$ to denote the expectation over the stationary distribution of the model parameters. The capital $T$ is used as a superscript to denote matrix transpose.
and lower case $t$ is used to denote the time step of optimization. Due to space constraint, we leave derivations to Appendix E.1.

4.1. SGD

Consider a general loss function $L(w')$ for a general differentiable model with parameters $w \in \mathbb{R}^D$; close to a local minimum, we may expand $L(w')$ up to second order in $w'$. Therefore, close to a local minimum, the dynamics of SGD is governed by a general form of a quadratic potential:

$$L(w') \approx \frac{1}{2}(w' - w*)^T K (w' - w*) \approx L(w'),$$

where the Hessian $K$ is a positive definite matrix, and $w*$ is a constant vector. One can redefine $w' - w* \rightarrow w$ to obtain a simplified form: $L(w) = w^T K w/2$. The SGD algorithm with momentum $\mu$ is defined by the update rule

$$\begin{align*}
g_t &= \nabla L(w_{t-1}) + \eta_{t-1} = K w_{t-1} + \eta_{t-1}; \\
m_t &= \mu m_{t-1} + g_t; \\
w_t &= w_{t-1} - \lambda m_t,
\end{align*}$$

(5)

where the noise $\eta_t$ has a finite covariance $C$, and $\mu \in [0, 1)$ is the momentum hyperparameter. We first consider the case without momentum, i.e., $\mu = 0$. Let $k^*$ be the largest eigenvalue of $K$. For any minimum with $\lambda k^* > 2$, the dynamics will diverge, and $w$ will escape from this minimum. Therefore, we focus on the range of $0 < \lambda < \frac{2}{k^*}$. This means that the absolute values of all the eigenvalues of $(I_D - \lambda K)$ are in the range of $(0, 1)$.

**Theorem 1.** (Model fluctuations of discrete SGD in a quadratic potential) Let $w_t$ be updated according to (5) with $\mu = 0$. The stationary covariance of $w$ satisfies

$$\Sigma K + K \Sigma + \lambda K \Sigma K = \lambda C.$$  

(6)

Recall that we have shifted the underlying parameter $w'$ by $w*$, and this result translates to that $\mathbb{E}[w'] = w*$, close to the local minimum $w*$.

**Remark.** The exact condition $\lambda \Sigma K + \lambda K \Sigma - \lambda^2 K \Sigma K = \lambda^2 C$ is different from the classical result obtained from a continuous Ornstein-Uhlenbeck process, which has $\Sigma K + \lambda K \Sigma = \lambda^2 C$. This suggests that the approximation of a discrete-time dynamics with a continuous-time one in Mandt et al. (2017) can be thought of as a $\lambda$-first-order approximation to the true dynamics. This approximation incurs an error of order $O(\lambda^2)$. The error becomes significant or even dominant as $\lambda$ gets large.

**Remark.** This theorem only requires the existence of a finite stationary noise covariance, in contrast to the unrealistic assumption of constant Gaussian noise by Mandt et al. (2017). In Appendix C, we show that the agreement of the theory is as good when the noises are non-gaussian.

4.2. SGD with Momentum

We now consider the case with arbitrary $\mu \in [0, 1)$ in Eq. (5).

**Theorem 2.** (Stationary distribution of discrete SGD with momentum) Let $w_t$ be updated according to (5) with arbitrary $\mu \in [0, 1)$. Then $\Sigma$ is the solution to

$$\begin{align*}
\frac{1 + \mu^2}{1 - \mu^2} \lambda^2 K \Sigma K + \frac{\mu}{1 - \mu^2} (\lambda^2 K \Sigma + \Sigma K^2) + (1 - \mu) \lambda (K \Sigma + \Sigma K) &= \lambda^2 C.
\end{align*}$$

(7)

If the noise is Gaussian, then the stationary state distribution of $w$ is $w \sim N(0, \Sigma)$.

We examine the above result (7) with two limiting examples. First, if there is no momentum, namely $\mu = 0$, we recover the previous result (6) without momentum. Next, if $\lambda K \ll 1$, neglecting the $O((\lambda K)^2)$ terms recovers the result of SGD with momentum under the continuous approximation (1) derived by Mandt et al. (2017). When $C$ commutes with $K$, the above matrix equation can be solved explicitly.

**Corollary 1.** Let $[C, K] = 0$. Then

$$\Sigma = \left[ \frac{\lambda K}{1 + \mu} \left( 2I_D - \frac{\lambda K}{1 + \mu} \right) \right]^{-1} \frac{\lambda^2 C}{1 - \mu^2}.$$  

(8)

We can obtain a more general result when the learning rate is a matrix.

**Theorem 3.** If the learning rate is a positive definite preconditioning matrix $\Lambda$, then $\Sigma$ satisfies

$$\begin{align*}
\frac{1 + \mu^2}{1 - \mu^2} \lambda K \Sigma K \Lambda + \frac{\mu}{1 - \mu^2} (\lambda K \Sigma K + K \Sigma K \Lambda) + (1 - \mu)(\lambda K \Sigma K + K \Sigma K) &= \lambda C \Lambda.
\end{align*}$$

(9)

Note that the matrix $\Lambda$ does not necessarily commute with $K$. We consider an application of this general result to understanding second-order methods in section 6.4.

4.3. Two Typical Kinds of Noise

As mentioned, two specific types of noise are of particular interest for machine learning practices. The first is a multidimensional white noise: $\eta \sim N(0, \sigma^2 I_D)$, where $\sigma$ is a positive scalar. The second type of noise is of Gaussian type with a covariance proportional to the Hessian, which is approximately equal to the noise caused by a minibatch gradient descent algorithm (Zhu et al., 2019; Xie et al., 2021): $C = a K$, where $a$ is a constant scalar, because

$$C(w) \approx \frac{1}{S} \frac{1}{N} \sum_{n=1}^N \nabla L(w, x_n) \nabla L(w, x_n)^T := \frac{1}{S} J(w) = \frac{1}{S} K,$$

where the first approximation is due to the fact that noise dominates the dynamics in the vicinity of a minimum, and the second approximation is according to the similarity between the Fisher information $J(w)$ and the Hessian $K$ near a minimum. This approximation is somewhat crude although it is often employed.
To be more general, one might wish to mix an isotropic Gaussian noise with the minibatch noise, namely $C = \sigma^2 I_D + aK$. The following theorem gives the distribution.

**Theorem 4.** Let $C = \sigma^2 I_D + a(\lambda) K$ and $\mu = 0$. Then the stationary distribution of $w$ is

$$
\mathcal{N} \left( 0, \lambda (\sigma^2 I_D + aK) [K (2 I_D - \lambda K)]^{-1} \right).
$$

Notice that, in this case, $C$ commutes with $K$. From this result, we can derive the following two special cases by setting $\sigma^2 = 0$ or $a = 0$.

**Corollary 2.** Let $\sigma^2 = 0$. Then $\Sigma = a \lambda (2 I_D - \lambda K)^{-1}$.

**Corollary 3.** Let $a = 0$. Then $\Sigma = \sigma^2 \lambda [K (2 I_D - \lambda K)]^{-1}$.

Interestingly, when $\lambda \ll 1$, the minibatch noise results in isotropic fluctuations, independent of the underlying geometry; the discrete time steps, however, causes fluctuations in the direction of large eigenvalues of the Hessian.

### 4.4. Approximation Error of SGD

We note that one important quantity for measuring the approximation error of SGD is $\text{Tr}[K\Sigma]$, because the expectation of a quadratic loss is

$$
L_{\text{train}} := \mathbb{E} \left( \frac{1}{2} w^T K w \right) = \frac{1}{2} \text{Tr}[K\Sigma],
$$

where the expectation is taken over the stationary distribution of $w$.

**Theorem 5.** (**Approximation error for discrete SGD with momentum**) Let the noise covariance $C$ and Hessian $K$ be any positive definite matrix. Then the training error for SGD with momentum is

$$
L_{\text{train}} = \frac{\lambda}{4(1 - \mu)} \text{Tr} \left[ \left( I_D - \frac{\lambda}{2(1 + \mu)} K \right)^{-1} C \right].
$$

This result suggests that a larger eigenvalue in the Hessian causes larger training error. Also, compared with the continuous result: $L_{\text{train}} = \frac{\lambda}{4(1 - \mu)} \text{Tr}[C]$, the discrete theory results in a larger approximation error.

### 4.5. Necessary Stability Condition

The main result in Theorem 2 also suggests a condition for the convergence of SGD. In order for a stationary distribution to exist at convergence, the covariance $\Sigma$ needs to exist and be positive definite, and this is only possible when $\left( 2 I_D - \frac{\lambda}{1 - \mu} K \right)$ is positive definite. This condition reflects the fact that discrete-time SGD becomes ill-conditioned as $\lambda$ increases, and so the continuous-time approximation becomes less valid. Also, an important implication is that using momentum may mitigate the ill-conditioning of the large learning rate, but only up to a factor of $1 + \mu < 2$, before the momentum causes another divergence problem due to the term $\frac{1}{1 - \mu}$. Therefore, when momentum is used, the necessary condition for convergence becomes $\lambda k^* \leq 2(1 + \mu) < 4$. We also comment that this is only a necessary condition for stability; the sufficient condition of stability is highly complicated and we leave this to a future work.

### 4.6. Regularization Effect of a Finite Learning Rate

For the continuous dynamics, the stationary distribution is known to obey the Boltzmann distribution, $P(w) \sim \exp[-L(w)/T]$ for some scalar $T$ determined by the noise strength (Landau and Lifshitz, 1980; Zhang et al., 2018). This implies that, close to a local minimum $w_*$, the stationary distribution is approximated by

$$
P(w|w^*) \sim \exp[-(w - w^*)^T K (w - w^*)/T].
$$

Comparing with the standard continuous-time solution (1), we see that this corresponds to an isotropic noise, namely $C = 2TI_D$, and $L_{\text{eff}} := -T \log P(w|w^*)$ may be defined as an effective loss function in analogy with an effective free energy in theoretical physics.

For discrete-time SGD, however, the stationary distribution has an additional term to leading order:

$$
P_D(w|w^*) = \exp \left[ -\frac{1}{T} L(w|w^*) \right] (\lambda + \frac{\lambda}{2} K^2) (w - w^*) + O(\lambda^2).
$$

This implies a different form of the effective loss function:

$$
L_{\text{eff}} \sim (w - w^*)^T K (w - w^*) + \frac{\lambda}{2} (w - w^*)^T K^2 (w - w^*),
$$

where the first term is the same as the continuous-time loss function, while the second term emerges as a discrete-time contribution due to a large learning rate. In particular, it encourages $w$ to have a smaller norm around the minimum $w_*$ in the kernel $K^2$. Therefore, to first order, the effect of the large learning rate can be understood as an effective weight decay term that encourages a smaller norm.

### 4.7. State-Dependent Noise

In reality, the minibatch noise depends on the current value of $w$, and the noise is, therefore, not constant (Simsekli et al., 2019; 2020; Hodgkinson and Mahoney, 2020; Meng et al., 2020; Ziyin et al., 2021; Mori et al., 2021). This means that the noise covariance is a function of the parameters covariance, namely $C = C(\Sigma)$. The general Theorem 2 is still applicable to this setting, and $\Sigma$ can be solved by setting $C = C(\Sigma)$, provided that $\Sigma$ exists. One example of state-dependent noise is given in section 6.4 where the noise covariance is set to $C \approx K\Sigma K$. After the publication of this work, we notice that Ziyin et al. (2021) has applied our formalism to study the exact shape and strength of SGD noise when minibatch sampling is used, and we refer the readers to this work for a more detailed study of the state-dependent noise in SGD.

### 5. Experiments

#### Gaussian Noise

We first consider the case when $w \in \mathbb{R}$ is one-dimensional. The loss function is $L(w) = \frac{1}{2} k w^2$ with $k = 1$. In Figure 2(a), we plot the variance of $w$ after 1000 training steps from $10^4$ independent runs. We compare the
We set the eigenvalues of the Hessian matrix to be \( \lambda k \) for \( 1 \leq k \leq 2 \). The predicted theory agrees excellently with the experiment, whereas the standard continuous-time approximation fails as \( \lambda \) increases. Moreover, the continuous-time approximation fails to predict the divergence of the variance at \( \lambda k \approx 2 \), whereas the discrete theory captures this very well. Now, we consider a multidimensional case. We set the loss function to be \( L(w) = \frac{1}{2} w^T K w \). For visualization, we choose \( D = 2 \), and we set the eigenvalues of the Hessian matrix to be 1 and 0.1, and plot the fluctuation along the eigenvectors of this Hessian matrix. See Figures 3(a)-(c). As before, we compare the discrete-time results with the theoretical predictions of Mandt et al. (2017). After diagonalization, the continuous-time dynamics predict \( \Sigma = \text{diag}(\lambda/2, \lambda/0.2) \), whereas the discrete theory predicts \( \Sigma = \text{diag}(\lambda/(2-\lambda), \lambda/(2-0.1\lambda)) \). The proposed theory exhibits no noticeable deviation from the experiment and successfully predicts a distortion along the direction with a large eigenvalue in the Hessian. In comparison, the continuous-time approximation always underestimates the fluctuation in the learning, and the discrepancy is larger as the learning rate gets larger; the prediction of the continuous-time theory can deviate arbitrarily far from the experiment as \( \lambda \) gets close to the divergence value.

**Minibatch Noise.** For minibatch noise, we solve a linear regression task with the loss function \( L(w) = \frac{1}{N} \sum_i (w^T x_i - y_i)^2 \), where \( N = 1000 \) is the number of data points; for the 1d case, the data points \( x_i \) are sampled independently from a normal distribution \( \mathcal{N}(0, 1) \); \( y_i = w^* x_i + \epsilon_i \) with a constant but fixed \( w^* \), \( \epsilon_i \) are noises, also sampled from a normal distribution. For sampling of minibatches, we set the batch size \( S = 100 \). The theoretical noise covariance matrix is approximated by \( C \approx K/S \). See Figure 2(b) for the 1d comparison. We see that the proposed theory agrees much better with the experiment than the continuous theory, both in trend and in magnitude. We also compare the predicted distribution for \( D = 2 \). Here, the data points \( x_i \) are sampled from \( \mathcal{N}(0, \text{diag}(1,0.1)) \), which makes the expected Hessian equal to \( \text{diag}(1,0.1) \). See Figures 3(d)-(f) for the illustration. Again, an overall agreement with the experiment is much better for the proposed theory. We notice that the prediction by Mandt et al. (2017) and the discrete theory agree well with each other when \( \lambda \) is small. Interestingly, the proposed theory slightly overestimates the variance of the parameters. This suggests the limitation of the commonly used approximation, \( C \sim K \) of the minibatch noise. It is possible to treat the minibatch noise in discrete-time regime rigorously in the proposed framework (Ziyin et al., 2021; Mori et al., 2021).

**SGD with Momentum.** For white noise, we set \( L(w) = k w^2/2 \) as before. In Figure 4(a), we plot the case with \( \lambda k > 2 \), where the dynamics will diverge if no momentum is present. The experiment does show this divergence at the value of \( \mu \rightarrow \lambda k/2 - 1 \) implied by the necessary stability condition, in contrast to the continuous-time theory that predicts no divergence. In Figure 4(b), we show the experiments with minibatch noise with the same \( \lambda \). The loss is the same as that of the minibatch noise. The predicted theory agrees better than the continuous-time approximation. On the other hand, the agreement becomes worse as the fluctuation in \( w \) becomes large. This suggests the limitation of the commonly used approximation of minibatch noise, i.e., \( C \sim H(w) = K \). More experimental results with a smaller learning rate are included in Appendix D.
We follow the setting of Mandt et al. (2017) for analyzing the escape rate, called the escaping efficiency, is proposed by Zhu et al. (2019) as

\[ E(t) := \mathbb{E}[L(w_t) - L(w_0)], \tag{17} \]

where \( w_0 \) is the exact minimum and \( t \) is a fixed time. This indicator qualitatively characterizes the ability of escape from the minimum \( w_0 \). It is related to the escape probability via the Markov inequality \( P(L(w_t) - L(w_0) \geq \delta) \leq \frac{\delta}{E(t)} \), for \( \delta > 0 \), if \( \delta = \Delta L \) is the height of the potential barrier.

**Theorem 7.** (Escaping efficiency from a sharp minimum)

Let the algorithm be updated according to Eq. (5) with \( \mu = 0 \). Then the escaping efficiency is

\[ E_d = \frac{\lambda}{4} \text{Tr} \left[ \left( I_D - \frac{\lambda K}{2} \right)^{-1} \left[ I_D - (I_D - \lambda K)^2t \right] C \right]. \tag{18} \]

The subscript \( d \) indicates discrete-time. In comparison, the escaping efficiency calculated from the continuous-time approximation is given by Zhu et al. (2019)

\[ E_c = \frac{\lambda}{4} \text{Tr} \left[ (I_D - e^{-2\lambda K t}) C \right], \tag{19} \]

where the subscript \( c \) indicates continuous-time. The two results can be easily compared in two limiting cases. First, in the short-time limit, the continuous-time theory predicts \( E_c = \frac{\lambda}{4} \text{Tr}[KC] \), which coincides with the single-step \( t = 1 \) result given by the discrete theory. Second, when \( t \gg 1 \), the continuous indicator becomes Hessian-independent: \( E_c = \frac{\lambda}{4} \text{Tr}[C] \), whereas the discrete result still depends on the curvature: \( E_d = \frac{\lambda}{4} \text{Tr}[(2I_D - \lambda K)^{-1}C] \). The conclusion that a flatter minimum relates to a smaller escaping efficiency still holds. If we take the small-\( \lambda \) limit, it recovers the trivial continuous result. In Figure 5, we compare the prediction of Eq. (18) with the continuous theory. We set \( C = I_D, K = I_D \) and compare at two different levels of learning rate. The result is averaged over 50000 runs. We see that our solution agrees with the experiment perfectly, while the continuous theory significantly underestimates the escape rate and fails at a large learning rate.

The following corollary shows that the discrete theory predicts a higher escape probability than the continuous one.

**Corollary 4.** \( \forall 0 < \lambda < 2/k^* \) and \( t \geq 0 \), \( E_d \geq E_c \).
We then investigate the effect of anisotropic noise on the escape efficiency as in Zhu et al. (2019). We consider different structure of noise with the same magnitude $\text{Tr}[C]$. We define an ill-conditioned Hessian $K$ as its descending ordered eigenvalues $k_1 \geq k_2 \cdots \geq k_D > 0$ satisfy $k_{l+1}, k_{l+2}, \ldots, k_D < k_l D^{-d}$ for some constant $l \ll D$ and $d > 1/2$. We assume that $C$ is aligned with $K$. Let $u_i$ be the corresponding unit eigenvector of eigenvalue $k_i$. For some coefficient $a > 0$, we have $u_i^T C u_1 \geq a k_1 \text{Tr}[C]/\text{Tr}[K]$. This is true if the maximal eigenvalues of $C$ and $K$ are aligned in proportion, namely $c_1/\text{Tr}[C] \geq a_1 k_1/\text{Tr}[K]$. 

**Theorem 8.** Under the conditions of an ill-conditioned Hessian and a noise covariance that aligns with the Hessian, the ratio between the escaping efficiencies of an anisotropic noise and its isotropic version $\bar{C} := \frac{\text{Tr}[C]}{\text{Tr}[K]} I_D$ is 

$$\frac{\text{Tr}[KC]}{\text{Tr}[K\bar{C}]} = O\left(a D^{2d-1}\right). \tag{20}$$

**Remark.** This result shows that the previous understanding that the anisotropy in noise may help escape from a sharp minimum still holds in a discrete-time regime. Therefore, the qualitative result in Zhu et al. (2019) still holds when a large learning rate is used.

### 6.3. Escape from Sharp Minima (Kramers Problem)

In physics, the Kramers escape problem (Kramers, 1940) concerns the approximate mean time for a particle confined in a local minimum of a potential $L(w)$ to escape across the potential barrier. For continuous-time dynamics, the standard approach to calculating this Kramers rate (or time) (Hanggi et al., 1990; Van Kampen, 1992) is to employ the Fokker-Planck equation for the distribution $P_c(w, t)$ (c for continuous-time) 

$$\frac{\partial P_c(w, t)}{\partial t} = -\nabla \cdot J(w, t), \tag{21}$$

where the probability current is defined as $J(w, t) := -\lambda P_c(w, t) \nabla L(w) - D \nabla^2 P_c(w, t)$. Here $D := \frac{1}{2} C$ is the diffusion matrix and $T$ is the effective “temperature”. The mean escape rate is defined as 

$$\gamma := \frac{P(w \in V_a)}{\int_{\partial a} J \cdot dS}, \tag{22}$$

where $P(w \in V_a)$ is the probability of a particle being inside the well $a$, and $\int_{\partial a} J \cdot dS$ is the probability flux through the boundary of the well $a$. For illustration, see Figure 6(a).

In continuous theory, the probability inside the well $a$ can be approximated by 1 because the distribution lies almost within the well. However, the discrete theory predicts larger fluctuations such that the distribution spreads out with large $\lambda$. By making proper approximations, we improve the result on the Kramers rate for the discrete SGD.

**Theorem 9.** Let $k_a$ and $k_b$ be the local Hessian at the local minimum $a$ and the barrier top $b$, respectively. Suppose $l$ is a midpoint on the most probable escape path between $a$ and $b$ such that $k(w) = k_a$ in the path $a \rightarrow l$ and $k(w) = k_b$ in $l \rightarrow b$. The approximate Kramers escape rate from a local minimum $a$ for the discrete-time SGD is

$$\gamma \approx \frac{1}{2\pi |k_b|} \sqrt{\frac{2}{2 - \lambda k_a}} \text{erf}\left(\frac{\sqrt{S(2 - \lambda k_a)\Delta L}}{\lambda k_a}\right) \times \exp\left[-\frac{2 S\Delta L}{\lambda} \frac{(1 - \lambda k_a/2) + 1 - l}{|k_b|}\right]. \tag{23}$$

where $\text{erf}(z)$ is the error function.

To compare, the mean escape rate obtained from the continuous-time theory (Xie et al., 2021) is

$$\gamma_c \approx \frac{1}{2\pi |k_b|} \exp\left[-\frac{2 S\Delta L}{\lambda} \left(\frac{l}{k_a} + 1 - l\right)\right]. \tag{24}$$

In Figure 6(b) we plot the quantity $\frac{\gamma}{|k_b|}$ while the experimental loss function by a factor $r: L \rightarrow rL$. The continuous theory predicts a constant escape rate by varying $r$, whereas the discrete theory expects a monotonic increase as $r$ becomes larger. Such monotonic increase is indeed observed in experiments. The theoretical curve is rescaled by a constant to make comparison easier. Our prediction is qualitatively consistent with the experiment, whereas the continuous theory is only valid in a rather limited range of small $r$. One surprising result here is that, the escape rate of continuous-time dynamics is invariant to the multiplication of the loss function by a constant $r$, while this is not true for discrete-time dynamics, where larger $r$ leads to a larger escape rate from a sharp minimum.
6.4. Second-Order Methods

In this subsection, we show that the proposed theory can also be extended to analyze the stochastic versions of second-order methods. Here, we consider the stochastic version of the damped Newton’s method (DNM) (Nesterov et al., 2018) where $\Lambda = \lambda K^{-1}$ and the natural gradient descent (NGD) algorithm in which the learning rate matrix is defined as $\Lambda := \lambda J(w)^{-1}$ with $J(w) := \mathbb{E}[\nabla L(\nabla L)^T]$ being the Fisher information. The NGD algorithm is known to be Fisher-efficient close to a local minimum (Amari, 1998; Amari and Nagaoka, 2007), namely, it is the fastest possible method to estimate a given statistical quantity, because the Fisher information is the “natural” metric in the probability space. The following corollaries give the model fluctuations of DNM and NGD under minibatch noise. A detailed discussion is given in Appendix E.2.7.

Corollary 5. Let the DNM algorithm be updated with noise covariance being $C = \frac{N-S}{NS} K$. The model fluctuation is

$$\Sigma = \frac{1 + \mu}{1 - \mu} \frac{\lambda}{2(1 + \mu) - \lambda} \frac{N - S}{NS} K^{-1}. \quad (25)$$

Corollary 6. Let the NGD algorithm be updated with noise covariance being $C = \frac{N-S}{NS} K \Sigma K$. Then,

$$\Sigma = \lambda \frac{(1 + \mu)}{2(1 - \mu^2)} \frac{N - S}{NS} + 1 - \mu \frac{\lambda}{2(1 - \mu^2)} K^{-1}. \quad (26)$$

Remark. For NGD, if $S \rightarrow N$, then $C \rightarrow 0$. In this situation, we have $\Sigma = \frac{\lambda}{2(1 + \mu)} K^{-1}$. This means that the algorithm involves nonzero fluctuations even if the noise is absent! Moreover, the divergence caused by $\lambda k^2 \rightarrow 2$ disappears for NGD. This suggests that, when the noise is due to minibatch sampling, NGD naturally corrects the ill-conditioned problem of discrete-time SGD.

Remark. Even more interestingly, both the DNM and the NGD algorithms induce fluctuations that are the same as those in the continuous-time SGD algorithm with Gaussian noise up to the coefficients, in the sense that the variance is proportional to $K^{-1}$ which is the local geometry of the minimum. Intuitively, this means that DNM and NGD need no correction even in the discrete-time case; moreover, they are likely to generalize better because they better align with the underlying loss function.

It would also be interesting to consider the stationary fluctuation of the adaptive gradient methods such as Adam (Kingma and Ba, 2014). Adam can be seen as an approximate second-order method with a diagonal preconditioning matrix $\hat{\Lambda} = \lambda / \sqrt{\mathbb{E} [\text{diag}(g^2) ]}$, where $g^2$ denotes element-wise square. This can be approximated by a non-diagonal preconditioning matrix $\hat{\Lambda} = \lambda / \sqrt{\mathbb{E} [gg^T] }$. Using $\mathbb{E} [gg^T]$, such approximation is in fact not bad; as shown in Zhiyi and Ziyin (2021), the diagonal assumption of $\mathbb{E} [gg^T]$ leads to surprisingly accurate predictions of the statistical properties of Adam for various modern neural architectures. Under this non-diagonal setting, we can solve the stationary covariance of Adam (without momentum).

Theorem 10. (Stationary fluctuation of Adam) Let the preconditioning matrix of Adam be $\Lambda = \lambda / \sqrt{\mathbb{E} [gg^T] }$ and the noise be $C = c K \Sigma K$ with a positive $c$, then

$$\Sigma_{\text{Adam}} = \frac{\lambda^2 (1 + c)}{4} I_D. \quad (27)$$

Remark. Similar to SGD, $\Sigma_{\text{Adam}}$ is isotropic. In general, we have $C \sim 1/S$. Two key differences exist: (1) that $\Sigma_{\text{SGD}}$ vanishes when $S \gg 1$, while that of Adam converge to a finite constant, like NGD; (2) $\Sigma_{\text{Adam}}$ scales as $\lambda^2$, while that of SGD scales as $\lambda$, and so Adam has much smaller fluctuation than SGD.

7. Concluding Remark

In this work, we have analyzed the SGD algorithm in a quadratic potential, which is a good approximation close to any local minimum and a common setting in the recent literature. Compared to the related works, our solution is exact, and relies on fewer assumptions than previous works, and, with the exact solutions, corrections to the previous results that were based on continuous-time approximation are obtained. In fact, we showed that even in the simplest settings, the prediction of the continuous-time solution may deviate significantly and eventually fails for a large learning rate. This suggests the fundamental limitation of making the continuous-time approximation in analyzing machine learning problems. Previous works have shown that, SGD leads to a flatter minimum due to the existence of anisotropic noise; this anisotropy is enhanced when the dynamics is discrete-time; this gives stronger mobility to model parameters along the sharper directions in the Hessian, and therefore, makes a flatter minimum more favorable. On the other hand, the distortion that the discrete-time SGD causes, in general, does not match the underlying landscape, indicating that using a large learning rate may cause a larger approximation error and worsened generalization. This tradeoff has been discussed only in a restricted setting in this work, and we hope the discussions here may stimulate further research in this direction.

Acknowledgment

We acknowledge Prof. Chikara Furusawa, Zeke Xie and Takashi Mori for valuable discussions. This work was supported by KAKENHI Grant No. JP18H01145 and a Grant-in-Aid for Scientific Research on Innovative Areas “Topological Materials Science (KAKENHI Grant No. JP15H05855) from the Japan Society for the Promotion of Science.
References


