Why does SGD prefer flat minima?: Through the lens of dynamical systems

Hikaru Ibayashi, [1] **Masaaki Imaizumi,** [2]

[1] University of Southern California
[2] The University of Tokyo / RIKEN Center for Advanced Intelligence Project
ibayashi@usc.edu, imaizumi@g.ecc.u-tokyo.ac.jp

Abstract

We show that stochastic gradient descent (SGD) escapes from sharp minima exponentially fast even before SGD reaches stationary distribution. SGD has been a de facto standard training algorithm for various machine learning tasks. However, there still exists an open question as to why SGDs find highly generalizable parameters from non-convex target functions, such as the loss function of neural networks. An "escaping" analysis has been an appealing framework to tackle this question. Escaping analysis measures how quickly SGD escapes from sharp minima, which is likely to have low generalization abilities. Despite its importance, the framework has the limitation that it works only when SGD reaches a stationary distribution after sufficient updates. In this paper, we prove that the SGD escapes from sharp minima exponentially fast even in a non-stationary setting. A key tool for the result is the Large Deviation Theory, a fundamental theory in dynamical systems. In particular, we found that a quantity called "quasi-potential" is a suitable tool to describe the SGD's stochastic behavior throughout its training process.

1 Introduction

Stochastic gradient descent (SGD) has become a de facto standard optimizer in modern machine learning, especially in deep learning. However, despite its prevalence, SGD leaves a theoretical question to us: why can SGD find generalizing solutions of complex models? This problem is particularly puzzling in modern machine learning with neural networks because their loss landscapes are known to be highly non-convex (Li et al. 2018), difficult to minimize (Blum and Rivest 1992), and full of non-generalizable minima (Zhang et al. 2017). Unraveling the generalization of SGD is an important open question for a modern machine learning community.

To answer the question, a narrative has emerged as one convincing hypothesis: "SGD can find generalizing solutions because it escapes from sharp minima" (Jastrzębski et al. 2017; Zhu et al. 2018; Xie, Sato, and Sugiyama 2020). "Sharp minima," in this context, mean local minima of loss functions that are sensitive to model parameters' perturbations, and they are known to deteriorate generalization ability (Keskar et al. 2016; Dziugaite and Roy 2017; Jiang et al. 2019). The "escaping" here means the behavior of SGD moving out of the neighborhood of minima. In a few words, the narrative explains that SGD quickly escapes from sharp minima and thus SGD tends to settle on a flat and generalizing solution (Zhu et al. 2018; Xie, Sato, and Sugiyama 2020). This explanation is aligned with actual phenomena. The left panel of Fig. 1 shows how SGD updates affect the sharpness of parameters throughout the SGD's training in a neural network. The figure shows that the sharpness oscillates drastically in the early phase of the training, and then becomes smaller toward the end. This plot suggests that SGD repeatedly jumps out of sharp minima and eventually reaches flat minima (Fig. 1, right).

There are active attempts to theoretically validate this hypothesis. Zhu et al. (2018) approximated SGD as a stochastic differential equation (SDE) with Gaussian noise and showed that the fast escaping is realized by the so-called "anisotropic noise" of SGD (the noise with the various magnitudes among directions.) Jastrzębski et al. (2017) formulated the effect of anisotropic noise in the **stationary regime**, where the SDE-model has reached a stationary distribution after many iterations. Xie, Sato, and Sugiyama (2020) further elaborated on this approach and found that escaping can be formulated as "Kramers escape rate," which is a well-used formula in physics (Kramers 1940). Their results revealed that the SGD escapes from minima **exponentially faster** as the sharpness of minima increases.

With these progressive refinements on escaping analysis, a remaining challenge is how to go beyond the stationary regime. Although, in physics, it is commonly assumed that systems have reached some stationary distributions (Eyring 1935; Hanggi 1986), such a stationary regime is not well applicable to SGD's analysis due to the following two reasons. First, it is shown that SGD forms a stationary distribution only over the very limited objective functions (Dieuleveut, Durmus, and Bach 2017; Chen, Mou, and Maguluri 2021). Secondly, even when such a stationary distribution exists, SGD takes $O(d)$ steps to reach it, where d is a number of parameters of a model (Raginsky, Rakhlin, and Telgarsky 2017). Since neural networks commonly have numerous parameters, the stationary regime may not be directly applicable to the actual SGD dynamics.

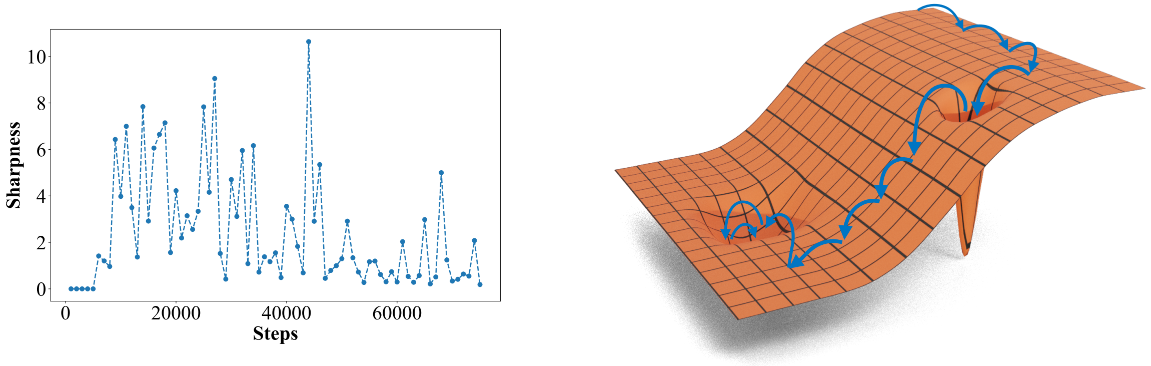


Figure 1: Sharpness dynamics throughout a training via SGD (left), where we used the sharpness defined in (Keskar et al. 2016, Metric 2.1). As is shown, sharpness fluctuates during the training and becomes small toward the end of training. This suggests SGD jumps out from sharp minima to find flat and generalizable minima (right). We used VGG (Simonyan and Zisserman 2014) fed with CIFAR-10 (Krizhevsky, Hinton et al. 2009) with cross-entropy loss.

In this paper, we propose a novel formulation of **exponentially fast** escape of the SGD in the **non-stationary regime**, by introducing the Large Deviation Theory, from dynamical systems (Dembo and Zeitouni 2010; Freidlin and Wentzell 2012). Based on the Large Deviation Theory, we show the following main result on SGD's time to escape:

Theorem 2 (informal):

$$\text{SGD's time to escape} \sim \exp\left[\frac{B}{\eta}\Delta L \lambda_{\max}^{-1}\right],$$

where B is a batch size, η is a learning rate, ΔL is the depth of the minimum, and $\lambda_{\max}$ is the maximum eigenvalue of the Hessian matrix at the minimum, i.e. sharpness of the minimum (Definition 3). We can see that as the sharpness of the minimum increases, i.e. $\lambda_{\max}$ increases, the time to escape decreases exponentially with the sharpness of minima. This is the first result showing that SGD escapes from sharp minima exponentially fast, in the non-stationary regime. As a further benefit, our formulation can be easily extended to the discrete update rule of SGD (Theorem 3).

2 Problem Formulation

Notations: For a $k \times k$ matrix M, $\lambda_j(M)$ is the j-th largest eigenvalue of M. We especially write $\lambda_{\max}(M) = \lambda_1(M)$ and $\lambda_{\min}(M) = \lambda_k(M)$. $\mathcal{O}(\cdot)$ denotes Landau's Big-O notation. $\|\cdot\|$ denotes the Euclidean norm. Given a time-dependent function θ_t, $\dot{\theta}_t$ denotes the differentiation of θ_t with respect to t. $N(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with the mean μ, and the covariance Σ.

2.1 Stochastic Gradient Descents

We consider a learning model parameterized by $\theta \in \mathbb{R}^d$, where d is a number of parameters. Given training examples $\{x_i\}_{i=1}^N$ and a loss function $\ell(\theta, x_i)$, we consider a training loss $L(\theta) := \frac{1}{N}\sum_{i=1}^N \ell(\theta, x_i)$ and a mini-batch loss $L^B(\theta)$ is a mini-batch training loss with size B.

We consider two types of stochastic gradient descent (SGD) methods; a discrete SGD and a continuous SGD.

Discrete SGD Given an initial parameter $\theta_0 \in \mathbb{R}^d$ and a learning rate $\eta > 0$, SGD generates a sequence of parameters $\{\theta_k\}_{k\in\mathbb{N}}$ by the following update rule:

$$\theta_{k+1} = \theta_k - \eta \nabla L^B(\theta_k),\ k \in \mathbb{N}. \tag{1}$$

We model SGD as a gradient descent with a Gaussian noise perturbation. We decompose $-\nabla L^B(\theta_k)$ in (1) into a gradient term $-\nabla L(\theta_k)$ and a noise term $\nabla L(\theta_k) - \nabla L^B(\theta_k)$, and model the noise as a Gaussian noise. With this setting, the update rule in (1) is rewritten as

$$\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k) + \sqrt{\frac{\eta}{B}} W_k, \tag{2}$$

where $W_k \sim N(0, \eta C(\theta_k))$ is a parameter-dependent Gaussian noise with its covariance, $C(\theta) := \mathbb{E}_{i\sim \mathrm{Uni}(\{1,\dots,N\})}[(\nabla L(\theta) - \nabla \ell(\theta, x_i))(\nabla L(\theta) - \nabla \ell(\theta, x_i))^\top]$.

Continuous SGD We define an SDE-model so that its discretization corresponds to the discrete SGD (2) using the Euler scheme (e.g. Definition 5.1.1 of (Gobet 2016)). We call this SDE model "continuous SGD." Given a time index $t \geq 0$ and an initial parameter $\theta_0 \in \mathbb{R}^d$, the continuous dynamic of continuous SGD is written as follows:

$$\dot{\theta}_t = -\nabla L(\theta_t) + \sqrt{\frac{\eta}{B}} C(\theta_t)^{1/2} \dot{w}_t \tag{3}$$

where w_t is a d-dimensional Wiener process, i.e. an $\mathbb{R}^d$-valued stochastic process with t such that $w_0 = 0$ and $w_{t+u} - w_t \sim N(0, uI)$ for any $t, u > 0$. We note that this system can be seen as a Gaussian perturbed dynamical system with a noise magnitude $\sqrt{\eta/B}$ because η and B do not evolve by time.

Remark 1 (Gaussianity of W_k)**.** *In this work, the Gaussianity of the noise on gradients (W_k) is justified by the following reasons: (i) if the batch size B is sufficiently large, the central limit theorem ensures the noise term becomes Gaussian (Zhu et al. 2018), and (ii) an empirical study shows that the noise term follows Gaussian distribution (Xie, Sato, and Sugiyama 2020).*

We also remark that the properties of SGD's noise are still actively studied and some studies showed W_k is heavy-tailed (Simsekli, Sagun, and Gurbuzbalaban 2019; Nguyen et al. 2019). We refer readers to Appendix A for in-depth discussions.

2.2 Mean Exit Time

We consider the problem of how discrete and continuous SGD's escape from the minima of loss surfaces. This is formally quantified by a notion of *mean exit time*. We define $\theta^* \in \mathbb{R}^d$ as a local minimum of loss surfaces, and also define its neighborhood $D \subset \mathbb{R}^d$ as an open set which contains θ^*. We define the mean exit time as follows:

Definition 1 (Mean exit time from D)**.** *Consider a continuous SGD (3) starting from $\theta_0 \in D$. Then, a mean exit time of the continuous SGD from D is defined as*

$$\mathbb{E}[\tau] := \mathbb{E}[\min\{t : \theta_t \notin D\}].$$

Intuitively, a continuous SGD with small $\mathbb{E}[\tau]$ easily escapes from the neighborhood D. The mean exit time of continuous SGD (3) formally corresponds to the SGD's "time to escape" in this paper.

Similarly, we define the discrete mean exit time as follows. Here, $k\eta$ (time step × learning rate) corresponds to t, since the η is regarded as a width of the discretization.

Definition 2 (Discrete mean exit time from D)**.** *Consider a discrete SGD (2) starting from $\theta_0 \in D$. Then, a discrete mean exit time of the discrete SGD from D is defined as*

$$\mathbb{E}[\nu] := \mathbb{E}[\min\{k\eta : \theta_k \notin D\}].$$

Remark 2 (Other measures for escaping)**.** *In previous work, there exist several terms and definitions to quantify the escaping behaviors. (Zhu et al. 2018) defined "escaping efficiency" as $\mathbb{E}_{\theta_t}\left[L\left(\theta_t\right) - L\left(\theta_0\right)\right]$. (Xie, Sato, and Sugiyama 2020) defined an "escape rate" as a ratio between the probability of coming out from θ^*'s neighborhood and the probability mass around θ^*. They also defined an "escape time" by the inverse of the escape rate.*

2.3 Basic Assumptions for SGD's Escape Problem

We provide basic assumptions that are commonly used in the literature on the escape problem (Mandt, Hoffman, and Blei 2016; Zhu et al. 2018; Jastrzębski et al. 2017; Xie, Sato, and Sugiyama 2020).

Assumption 1 ($L(\theta)$ is locally quadratic in D)**.** *There exists a matrix $H^* \in \mathbb{R}^{d\times d}$ such that for any $\theta \in D$, the following equality holds:*

$$\begin{aligned}\forall\theta \in D, L(\theta) =& L\left(\theta^*\right) + \nabla L\left(\theta^*\right)\left(\theta - \theta^*\right) \\ &+ \frac{1}{2}\left(\theta - \theta^*\right)^\top H^*\left(\theta - \theta^*\right)\end{aligned}$$

Assumption 2 (Covariance matrix at θ^*)**.** $C(\theta^*) = H^*$.

Assumption 3. *For all θ in D, all the entries of $C(\theta)$ are Lipschitz continuous in θ.*

Assumption 4. *For all θ in D, there exists a constant $k > 0$ that bounds the eigenvalues of $C\left(\theta\right)$: $k \geq \lambda_1 \geq \ldots \geq \lambda_d \geq \frac{1}{k}$.*

It is known that Assumption 2 hold at global minima (Jastrzębski et al. 2017; Zhu et al. 2018) and it is empirically shown that it is a reasonable approximation at local minima as well in (Xie, Sato, and Sugiyama 2020, Section 2). Although Assumption 4 is admittedly strong, it is shown that even neural networks have parameters satisfying it when their activation functions are bounded (Zhong et al. 2017, Appendix D.2).

Finally, we adopt the following metric as sharpness in our analysis.

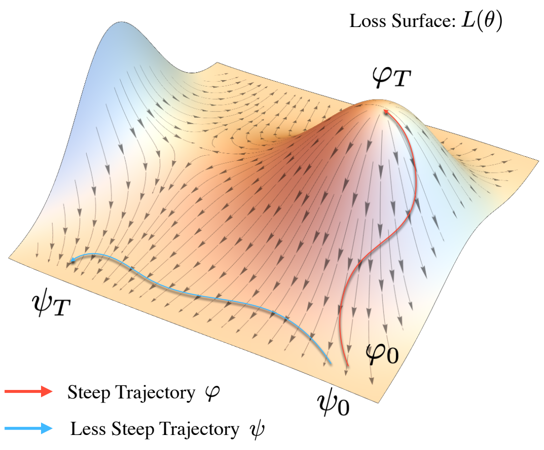

Loss Surface: $L(\theta)$
φ_T
ψ_T
φ_0
ψ_0
Steep Trajectory φ
Less Steep Trajectory ψ

Figure 2: Visual illustration of steepness (Definition 4). The steepness of φ, $S_T(\varphi)$, is greater than $S_T(\psi)$ because φ moves against the vector field of gradient $-\nabla L(\theta)$.

Definition 3 (Sharpness of a minimum θ^*)**.** *Sharpness of θ^* is the maximum eigenvalue of H^* in Assumption 1, that is,*

$$\lambda_{\max} = \lambda_{\max}(H^*).$$

This is one of the most common definitions of sharpness (Goldblum et al. 2020; Jastrzebski et al. 2020). We note, however, that there is ongoing controversy over its definition (Dinh et al. 2017).

3 Large Deviation Theory for SGD

We introduce the basic notions from the Large Deviation Theory (Dembo and Zeitouni 2010; Freidlin and Wentzell 2012). The Large Deviation Theory is a theoretical framework for stochastic processes and one of its results, *mean exit time analysis*, can formally quantity escaping in our setup.

First, we define *steepness* of a trajectory on a loss surface $L(\theta)$, followed by the continuous SGD (3). Let $\varphi = \{\varphi_t\}_{t\in[0,T]} \subset \mathbb{R}^d$ be a trajectory in the parameter space over a time interval $[0, T]$ with a terminal time T, where $\varphi_t \in \mathbb{R}^d$ is a parameter which continuously changes in t (see Figure 2). Also, φ is regarded as a continuous map from $[0, T]$ to $\mathbb{R}^d$, i.e. is an element of $\mathbf{C}_T(\mathbb{R}^d)$ (a set of continuous trajectories in $\mathbb{R}^d$) which is a support of continuous SGD during $[0, T]$. Here we define the following quantity.

Definition 4 (Steepness of φ for SGD)**.** *Steepness of a trajectory φ for (3) is defined as*

$$S_T(\varphi) := \frac{1}{2}\int_0^T \left(\dot{\varphi}_t + \nabla L\left(\varphi_t\right)\right)^\top C\left(\varphi_t\right)^{-1}\left(\dot{\varphi}_t + \nabla L\left(\varphi_t\right)\right)dt$$

Intuitively, steepness $S_T(\varphi)$ means the hardness for the system (3) to follow this trajectory φ up the hill on $L(\theta)$, as illustrated in Figure 2. The steepness is introduced to describe a distribution of trajectories of continuous SGD. If a trajectory φ has a large steepness $S_T(\varphi)$, the probability that the system takes the trajectory decreases. We formally summarize the properties of steepness in Appendix C. Although steepness is a tailored definition for our setup, this quantity is a special case of a more general notion in the Large Deviation Theory. The corresponding notions are called "rate function" in (Dembo and Zeitouni 2010, Section 1.2) and "normalized action functional" in (Freidlin and Wentzell 2012, Section 3.2).

We secondly define *quasi-potential*, which is the smallest steepness from a minimum θ^* to a boundary ∂D. It plays an essential role in the mean exit time.

Definition 5 (Quasi-potential)**.** *Given the system (3) whose initial point is a local minima θ^*, quasi-potential of a parameter $\theta \in D$ is defined as*

$$V(\theta) := \inf_{T>0} \inf_{\varphi:(\varphi_0,\varphi_T)=(\theta^*,\theta)} S_T(\varphi).$$

Same as steepness, quasi-potential can be seen as the minimum effort that the system (3) needs to climb from θ^* up to θ on $L(\theta)$. Given those definitions, Large Deviation Theory provides the mean exit time of a continuous SGD (3) based on the quasi-potential.

Theorem 1 (Fundamental Theorem of Exit Time)**.** *Consider the continuous SGD (3) whose initial point is the local minima $\theta_0 = \theta^*$. Suppose Assumption 1, 3, and 4 hold. Then, the mean exit time (Definition 1) has the following limit:*

$$\lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}\left[\tau\right] = V_0$$

holds, where $V_0 := \min_{\theta \in \partial D} V(\theta)$.

We obtain Theorem 1 by adapting a more general theorem (Dembo and Zeitouni 2010, Theorem 5.7.11 (a)) to our setting with Assumption 1. Rigorously, we verify that several requirements of the general theorem, such as asymptotic stability and attractiveness, are satisfied with our setup. A precise description of the assumptions can be found in Appendix B, and the proof of Theorem 1 under our setup can be found in Appendix G.

4 Mean Exit Time Analysis for SGD

In this section, we give an asymptotic analysis of the mean exit time as our main result. As preparation, we provide an approximate computation of the quasi-potential in our setting, then we give the main theorem.

4.1 Approximate Computation of Quasi-potential

We develop an approximation of the quasi-potential $V(\theta)$, which is necessary to study the mean exit time by the fundamental theorem (Theorem 1). However, the direct calculation with a general $C(\theta)$ is a difficult problem, and at best we get a necessary condition for the exact formula (Hu et al. 2017). Instead, we consider a *proximal system* which is a simplified version of the continuous SGD (3) with a state-independent noise covariance.

Proximal System with $C(\theta) = I$ We define the following proximal system which generates a sequence $\{\widehat{\theta}_t\}$:

$$\dot{\widehat{\theta}}_t = -\nabla L\left(\widehat{\theta}_t\right) + \sqrt{\frac{\eta}{B}}\dot{w}_t \qquad (4)$$

This system is obtained by replacing the covariance $C(\theta)$ of the continuous SGD (3) with an identity I. That is, this proximal system is regarded as a Gaussian gradient descent with isotropic noise.

We further define steepness and quasi-potential of the proximal system as follow:

Steepness

For each $\varphi \in \mathbf{C}_T(\mathbb{R}^d)$, $\widehat{S}_T(\varphi) := \dfrac{1}{2}\displaystyle\int_0^T \|\dot{\varphi}_t + \nabla L\left(\varphi_t\right)\|^2 dt$

Quasi-potential

For each $\theta \in D$,$\widehat{V}(\theta) := \displaystyle\inf_{T>0} \inf_{\varphi:(\varphi_0,\phi_r)=(\theta^*,\theta)} \widehat{S}_T(\varphi)$

Owing to the noise structure of the proximal system, we achieve a simple form of the quasi-potential. For the quasi-potential $\widehat{V}(\theta)$, the following lemma holds:

Lemma 1. *Under Assumption 1, $\widehat{V}(\theta) = 2\left(L(\theta) - L(\theta^*)\right)$.*

Proof. If the function φ_t for $t \in [0, T]$ does not exit from $D \cup \partial D$,

$$\begin{aligned}\widehat{S}_T(\varphi) &= \frac{1}{2}\int_0^T \left\|\dot{\varphi}_t - \nabla L\left(\varphi_t\right)\right\|^2 dt + 2\int_0^T \dot{\varphi}_t^\top \nabla L\left(\varphi_t\right) dt \\ &= \frac{1}{2}\int_0^T \left\|\dot{\varphi}_t - \nabla L\left(\varphi_t\right)\right\|^2 dt + 2\left(L(\varphi_T) - L(\varphi_0)\right) \\ &\geq 2\left(L(\varphi_t) - L(\varphi_0)\right)\end{aligned}$$

The equality holds when $\dot{\varphi}_t = \nabla L\left(\varphi_t\right)$. Since quasi-potential at θ is the infimum of the steepness from θ^* to θ, $\widehat{V}(\theta) = 2(L(\theta) - L(\theta^*))$ is obtained. □

Lemma 1 shows that the quasi-potential with the proximal system is simply represented as the height of θ from a minimum θ^*. By the quasi-potential, we simply obtain the following result by combining Theorem 1:

Proposition 1 (Mean Exit Time of Proximal System)**.** *Consider the proxy system (4) whose initial point is the local minima $\theta_0 = \theta^*$. Suppose that Assumption 1, 2, and 5 hold. Then, the mean exit time of (4) from the neighborhood D, $\mathbb{E}[\widehat{\tau}]$, has the following limit*

$$\lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}[\widehat{\tau}] = 2\Delta L.$$

Approximation of Quasi-potential $V_0 := \min_{\theta \in \partial D} V(\theta)$ We approximate the target quasi-potential $V(\theta)$ using $\lambda_{\max}^{-1}\widehat{V}(\theta)$ from the proximal system. For this sake, we impose the following assumption:

Assumption 5. *There exists $K > 0$ such that for any $\theta \in D$, if $S_T(\varphi) = V(\theta)$, then $\forall t \in [0, T] : \dot{\varphi}_t \leq K$ holds.*

This claims that the velocities of trajectories do not become infinitely large. With this mild assumption, we obtain the estimation of V_0 as follows.

Lemma 2. *Under Assumption 1, 2, and 5, there exists a constant A such that*

$$\left|V_0 - \lambda_{\max}^{-1}\widehat{V}_0\right| \leq A\left(\left(1 + C_0 rk\right)\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)$$

where C_0 is a Lipschitz constant of $C(\theta)$ and r is the radius of D.

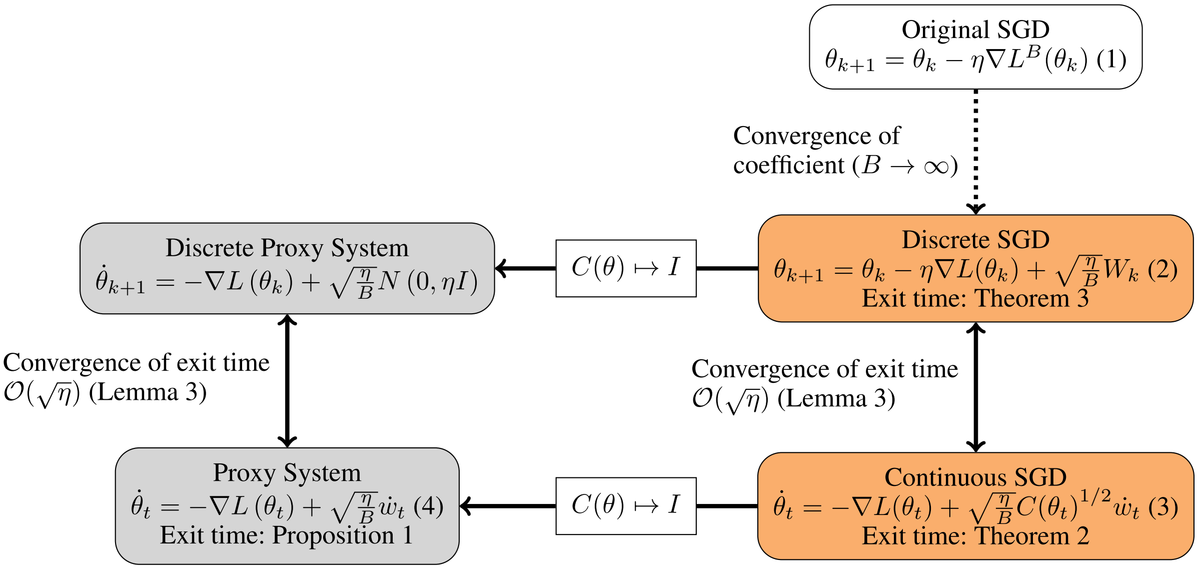

Original SGD
$\theta_{k+1} = \theta_k - \eta \nabla L^B(\theta_k)$ (1)
Convergence of coefficient ($B \to \infty$)
Discrete Proxy System
$\dot{\theta}_{k+1} = -\nabla L(\theta_k) + \sqrt{\frac{\eta}{B}} N(0, \eta I)$
$C(\theta) \mapsto I$
Discrete SGD
$\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k) + \sqrt{\frac{\eta}{B}} W_k$ (2)
Exit time: Theorem 3
Convergence of exit time
$\mathcal{O}(\sqrt{\eta})$ (Lemma 3)
Convergence of exit time
$\mathcal{O}(\sqrt{\eta})$ (Lemma 3)
Proxy System
$\dot{\theta}_t = -\nabla L(\theta_t) + \sqrt{\frac{\eta}{B}} \dot{w}_t$ (4)
Exit time: Proposition 1
$C(\theta) \mapsto I$
Continuous SGD
$\dot{\theta}_t = -\nabla L(\theta_t) + \sqrt{\frac{\eta}{B}} C(\theta_t)^{1/2} \dot{w}_t$ (3)
Exit time: Theorem 2

Figure 3: The whole structure of our results.

In the statement of Lemma 2, the right-hand side is small when (i) H^* is well-conditioned (i.e. $\lambda_{\min} \approx \lambda_{\max}$) and (ii) $C(\theta)$ does not change much around θ^* (i.e. C_0 is small.) Under such conditions, the quasi-potential of continuous SGD becomes close to $\lambda_{\max}^{-1} \widehat{V}_0$. But we note that this approximation can be loose because of the hidden factors in A (Appendix D).

4.2 Main Results: Mean Exit Time Analysis

As our main results, we give inequalities that characterize a limit of the mean escape time. We recall the definition of the depth of a minimum θ^* as $\Delta L := \min_{\theta \in \partial D} L(\theta) - L(\theta^*)$.

Continuous SGD First, we study the case of continuous SGD (3). This result is obtained immediately by combining the fundamental theorem (Theorem 1) with the approximated quasi-potential (Lemma 2):

Theorem 2 (Mean Exit Time of Continuous SGD)**.** *Consider the continuous SGD (3) whose initial point is the local minima $\theta_0 = \theta^*$. Suppose that Assumption 1, 2, and 5 hold. Then, the mean exit time (Definition 1) from the neighbourhood D has the following limit:*

$$2\frac{\Delta L}{\lambda_{\max}} - A\left(\left(1 + C_0 rk\right) \lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right) \leq \lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}[\tau]$$
$$\leq 2\frac{\Delta L}{\lambda_{\max}} + A\left(\left(1 + C_0 rk\right) \lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)$$

Excluding the effect of the approximation $A(\lambda_{\min}^{-1} - \lambda_{\max}^{-1} + C_0 rk^2)$, this result indicates that continuous SGD needs $\exp(2\frac{B}{\eta}\lambda_{\max}^{-1}\Delta L)$ time of update to escape from the neighborhood D of the local minima θ^*. Compared to the quasi-potential of the proxy system $2\Delta L$ (Proposition 1), the covariance matrix reduces quasi-potential in the factor of $\lambda_{\max}$. This result endorses the fact that SGD's noise structure, $C(\theta)$, exponentially accelerates the escaping (Xie, Sato, and Sugiyama 2020), because quasi-potential exponentially affects mean exit time (Theorem 1). A more rigorous comparison is given in Section 5.

Discrete SGD Next, we give the mean escape time analysis for discrete SGD (2). Our approach is to combine the following discretization error analysis to the continuous SGD results (Theorem 2):

Lemma 3 (Discretization Error)**.** *For a stochastic system with Gaussian perturbation and its discrete correspondence, the discretization error of exit time has the following convergence rate $\mathbb{E}[\nu] - \mathbb{E}[\tau] = \mathcal{O}(\sqrt{\eta})$.*

The following lemma can be simply derived as a special case of (Gobet and Menozzi 2010, Theorem 17) by substituting $g(\cdot) = 0$, $f(\cdot) = 1$, and $k(\cdot) = 0$ in their definition.

Based on the analysis, we obtain the following result:

Theorem 3 (Mean Exit Time of Discrete SGD)**.** *Consider the discrete Gaussian SGD (2) whose initial point is the local minima $\theta_0 = \theta^*$. Suppose that Assumption 1, 2, and 5, hold. Then, the mean exit time (Definition 1) from the neighbourhood D has the following limit:*

$$2\frac{\Delta L}{\lambda_{\max}} - A\left(\left(1 + C_0 rk\right) \lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right) \leq \lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}[\nu]$$
$$\leq 2\frac{\Delta L}{\lambda_{\max}} + A\left(\left(1 + C_0 rk\right) \lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)$$

This result indicates that continuous and discrete SGD have identical asymptotic mean exit times. In other words, the discretization error is asymptotically negligible in this analysis of escape time. Fig 3 summarizes the whole structure of our results.

5 Comparison with Existing Escape Analyses

We compare our analysis with the closely related existing analyses and discuss the technical differences in detail. As summarized in Table 1, we picked as closely related analysis, (Hu et al. 2017; Jastrzębski et al. 2017; Zhu et al. 2018; Nguyen et al. 2019; Xie, Sato, and Sugiyama 2020), which analyzes how the SGD's noise affects escape efficiency.

Comparison on exit time From Table 1, we obtain three implications. (i) In all the results, either or both the learning rate η and H^* play an important role. (ii) There are four

Studies	Exponential escape	Sharpness analysis	No escape paths	Non-stationary	Discreteness	Exit Time (Order)
(Hu et al. 2017)			$\surd$	$\surd$	$\surd$	$\exp(1/\eta)$
(Jastrzębski et al. 2017)		$\surd$				$\exp\left(\frac{B}{\eta}\Delta L + d\right)$
(Zhu et al. 2018)		$\surd$	$\surd$	$\surd$		$1/\mathrm{Tr}\left(C(\theta^*)^{-1}H^*\right)$
(Nguyen et al. 2019)			$\surd$	$\surd$	$\surd$	$1/\eta^{(\alpha-\delta)/2}$
(Xie, Sato, and Sugiyama 2020)	$\surd$	$\surd$				$\exp\left(\frac{B}{\eta}\Delta L\bar{\lambda}^{-1}\right)$
Ours	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\exp\left(\frac{B}{\eta}\left(\Delta L\lambda_{\max}^{-1} \pm \Xi\right)\right)$

Table 1: Technical difference among analyses. The specific meanings of each column are described in the main passages of Section 5. For the results of each work, we only show their order by ignoring constants. $\Xi = A\left(\left(1 + C_0 rk\right)\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)$ is the approximation error, $\bar{\lambda}$ is some value in $[\lambda_{\min}, \lambda_{\max}]$ defined in (Xie, Sato, and Sugiyama 2020), and α, δ are parameters related to the tail probability (Nguyen et al. 2019).

results where the exit time is expressed as an exponential form, and the sharpness-related values $\lambda_{\max}$ and $\bar{\lambda}$ appear in the results of (Xie, Sato, and Sugiyama 2020) and our study. (iii) Our study and (Xie, Sato, and Sugiyama 2020) have different orders for the parameters for sharpness. This fact will be discussed in the latter half of this section.

Escaping path assumption We remark that the assumptions of our theorem have an essential difference from (Jastrzębski et al. 2017) and (Xie, Sato, and Sugiyama 2020). Their analyses assume that SGD escapes along a linear path, named "escape path," where the gradient perpendicular to the path direction is zero. Escaping path is a convenient assumption to reduce the escape analysis to one-dimensional problems. However, the existence of such paths is supported only weakly by (Draxler et al. 2018), and it is unlikely that the stochastic process continuously moves linearly. The fact that we eliminated the escaping path assumptions is a substantial technical improvement.

Effect of sharpness The technical significance of our theory is that it can analyze the sharpness effect. Because of its non-linearity, sharpness analyses tend to become nontrivial, thus a limited number of existing works have tackled it. Among the selected results, the sharpness effect appears in (Jastrzębski et al. 2017) and (Zhu et al. 2018) as H^*, and in (Xie, Sato, and Sugiyama 2020) as λ. We note that the results of (Jastrzębski et al. 2017) and (Xie, Sato, and Sugiyama 2020) include auxiliary sharpness values, such as $\bar{\lambda} \in [\lambda_{\min}, \lambda_{\max}]$ respectively. Those terms appear because of the escaping path assumption and our results show that those terms are not fundamental.

Heavy tailed noise Among the selected works, only (Nguyen et al. 2019) use a heavy-tailed noise model, i.e. the noise whose distribution has a heavier tail than exponential distribution. Although it is known that the heavy-tailed noise models the empirical behavior of SGD well (Simsekli, Sagun, and Gurbuzbalaban 2019), it is quite difficult to mathematically formulate it. (Nguyen et al. 2019) use the Lévy process for their analysis, where α represents the degree of the heavy tail, and $\delta \in (0, 1)$ includes miscellaneous constants. Analyzing the sharpness under the heavy-tailed setup is still an open problem.

6 Conclusion

In this paper, we showed that SGD escapes from sharp minima exponentially fast even in the non-stationary regime. To obtain the result, we used the Large Deviation Theory from dynamical system and identified that quasi-potential plays the key role in the exponential escape in the non-stationary regime. Our results are the novel theoretical clue to explain the mechanics as to why SGD can find generalizing minima.

Acknowledgments

Hikaru Ibayashi and Masaaki Imaizumi were supported from JST Presto (JPMJPR1852) and JST Crest (JPMJCR21D2).

References

Absil, P.-A.; and Kurdyka, K. 2006. On the stable equilibrium points of gradient systems. *Syst. Control Lett.*, 55(7): 573–577.

Blum, A. L.; and Rivest, R. L. 1992. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1): 117–127.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177–186. Springer.

Chen, Z.; Mou, S.; and Maguluri, S. T. 2021. Stationary Behavior of Constant Stepsize SGD Type Algorithms: An Asymptotic Characterization. *arXiv preprint arXiv:2111.06328*.

Cheng, X.; Yin, D.; Bartlett, P.; and Jordan, M. 2020. Stochastic gradient and langevin processes. In *International Conference on Machine Learning*, 1810–1819. proceedings.mlr.press.

Daneshmand, H.; Kohler, J.; Lucchi, A.; and Hofmann, T. 2018. Escaping Saddles with Stochastic Gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 1155–1164. PMLR.

De Stefano, C.; Fontanella, F.; Maniaci, M.; and di Freca, A. S. 2011. A method for scribe distinction in medieval manuscripts using page layout features. In *International Conference on Image Analysis and Processing*, 393–402. Springer.

Dembo, A.; and Zeitouni, O. 2010. *Large Deviations Techniques and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2nd edition.

Dieuleveut, A.; Durmus, A.; and Bach, F. 2017. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*.

Dinh, L.; Pascanu, R.; Bengio, S.; and Bengio, Y. 2017. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 1019–1028.

Draxler, F.; Veschgini, K.; Salmhofer, M.; and Hamprecht, F. 2018. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, 1309–1318. PMLR.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Dziugaite, G. K.; and Roy, D. M. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Eyring, H. 1935. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2): 107–115.

Freidlin, M. I.; and Wentzell, A. D. 2012. *Random Perturbations of Dynamical Systems 3rd Ed.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Garipov, T.; Izmailov, P.; Podoprikhin, D.; Vetrov, D.; and Wilson, A. G. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8803–8812.

Geiping, J.; Goldblum, M.; Pope, P.; Moeller, M.; and Goldstein, T. 2022. Stochastic Training is Not Necessary for Generalization. In *International Conference on Learning Representations*.

Gobet, E. 2016. *Monte-Carlo Methods and Stochastic Processes: From Linear to Non-Linear*. CRC Press.

Gobet, E.; and Menozzi, S. 2010. Stopped diffusion processes: Boundary corrections and overshoot. *Stochastic Process. Appl.*, 120(2): 130–162.

Goldblum, M.; Geiping, J.; Schwarzschild, A.; Moeller, M.; and Goldstein, T. 2020. Truth or backpropaganda? An empirical investigation of deep learning theory. In *International Conference on Learning Representations*.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Hanggi, P. 1986. Escape from a metastable state. *Journal of Statistical Physics*, 42(1): 105–148.

He, F.; Liu, T.; and Tao, D. 2019. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32: 1143–1152.

He, H.; Huang, G.; and Yuan, Y. 2019. Asymmetric Valleys: Beyond Sharp and Flat Local Minima. *Advances in Neural Information Processing Systems*, 32: 2553–2564.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1995. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, 529–536.

Hochreiter, S.; and Schmidhuber, J. 1997. Flat minima. *Neural computation*, 9(1): 1–42.

Hoffer, E.; Hubara, I.; and Soudry, D. 2017. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1729–1739.

Hu, W.; Li, C. J.; Li, L.; and Liu, J.-G. 2017. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705. 07562*.

Huang, W. R.; Emam, Z.; Goldblum, M.; Fowl, L.; Terry, J. K.; Huang, F.; and Goldstein, T. 2020. Understanding Generalization Through Visualizations. In Zosa Forde, J.; Ruiz, F.; Pradier, M. F.; and Schein, A., eds., *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, 87–97. PMLR.

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

Jastrzebski, S.; Szymczak, M.; Fort, S.; Arpit, D.; Tabor, J.; Cho, K.; and Geras, K. 2020. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*.

Jiang, Y.; Neyshabur, B.; Mobahi, H.; Krishnan, D.; and Bengio, S. 2019. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, B.; Li, Y.; and Yuan, Y. 2018. An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, 2698–2707. PMLR.

Kramers, H. A. 1940. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4): 284–304.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. *Advances in Neural Information Processing Systems*, 31.

Li, Z.; and Arora, S. 2020. An Exponential Learning Rate Schedule for Deep Learning. In *International Conference on Learning Representations*.

Li, Z.; Lyu, K.; and Arora, S. 2020. Reconciling Modern Deep Learning with Traditional Optimization Analyses: The Intrinsic Learning Rate. In *NeurIPS*.

Li, Z.; Malladi, S.; and Arora, S. 2021. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 34.

Mandt, S.; Hoffman, M.; and Blei, D. 2016. A Variational Analysis of Stochastic Gradient Algorithms. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 354–363. New York, New York, USA: PMLR.

Masters, D.; and Luschi, C. 2018. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Nguyen, T. H.; Şimşekli, U.; Gürbüzbalaban, M.; and Richard, G. 2019. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *arXiv preprint arXiv:1906.09069*.

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

Raginsky, M.; Rakhlin, A.; and Telgarsky, M. 2017. Nonconvex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In Kale, S.; and Shamir, O., eds., *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, 1674–1703. PMLR.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Schmidt, R. M.; Schneider, F.; and Hennig, P. 2021. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*, 9367–9376. PMLR.

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

Şimşekli, U.; Gürbüzbalaban, M.; Nguyen, T. H.; Richard, G.; and Sagun, L. 2019. On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks. *arXiv preprint arXiv:1912.00018*.

Simsekli, U.; Sagun, L.; and Gurbuzbalaban, M. 2019. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5827–5837. PMLR.

Smith, S. L.; Dherin, B.; Barrett, D. G.; and De, S. 2021. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*.

Smith, S. L.; Kindermans, P.-J.; Ying, C.; and Le, Q. V. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*.

Teschl, G. 2000. Ordinary differential equations and dynamical systems. *Grad. Stud. Math.*, 140: 08854–08019.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, L.; Ma, C.; et al. 2018. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31: 8279–8288.

Wu, L.; Zhu, Z.; et al. 2017. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*.

Xie, Z.; Sato, I.; and Sugiyama, M. 2020. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. In *International Conference on Learning Representations*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization (2016). *arXiv preprint arXiv:1611.03530*.

Zhong, K.; Song, Z.; Jain, P.; Bartlett, P. L.; and Dhillon, I. S. 2017. Recovery Guarantees for One-hidden-layer Neural Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 4140–4149. PMLR.

Zhu, Z.; Wu, J.; Yu, B.; Wu, L.; and Ma, J. 2018. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*.

A Related Work

Sharpness and Generalization of Neural Networks The shape of loss surfaces has long been a topic of interest. The argument that the flatness of loss surfaces around local minima improves generalization was first studied by (Hochreiter and Schmidhuber 1995, 1997), and the observation has recently been reconfirmed in deep neural networks by (Keskar et al. 2016). Although it is difficult to formally define sharpness (Dinh et al. 2017), many empirical studies have shown its connections to generalization where various definitions of sharpness are studied, such as the volume around a minimum (Huang et al. 2020), the maximum eigenvalue of Hessian matrix (Goldblum et al. 2020, Remark 2.2), and the highest loss value around a minimum (Jiang et al. 2019). Our analysis adopts the maximum eigenvalue of the Hessian matrix (Definition 3) as (Goldblum et al. 2020) did.

As the other investigations on the loss surface geometry, (He, Huang, and Yuan 2019) discussed the asymmetry of loss surfaces, (Draxler et al. 2018; Garipov et al. 2018) studied how multiple local minima are internally connected, and (Li et al. 2018) developed a random dimensional reduction method to visualize loss surfaces in low dimensions.

Learning Rate and Generalization Although our analysis is based on the asymptotic limit of $\eta \to 0$, the practical SGD has finite learning rates (Goyal et al. 2017). There are lines of research investigating particularly the finite learning rates. (Smith et al. 2021) have shown that the finite learning rate of SGD works as an implicit regularizer. (Li, Malladi, and Arora 2021) have investigated the validity of SDE approximation of SGD with finite learning rates. The differences in convergence properties among different learning rates are also an active research area (Wu, Ma et al. 2018; Li, Lyu, and Arora 2020). A shortcoming of our framework is that it cannot deal with the effect of finite learning rates.

Although the practical SGD commonly uses dynamical learning rates (or schedulers) (Li and Arora 2020), it is fair to assume that the learning rate is constant in our analysis because we consider the behavior of SGD during a short period of time, (i.e. after SGD falls to a minimum until it escapes.)

SGD and Machine Learning The detailed nature of SGD itself is also an object of interest. SGD was first proposed in (Robbins and Monro 1951), as a lazy version of gradient descent using random subsets of training data. Thus, SGD has been intended to be a convenient heuristic rather than a refined algorithm. However besides its computational convenience, SGD works as effectively as gradient descent does in many optimization problems, and its convergence properties have been solidified on the convex objective functions (Bottou 2010). The recent success in the field of neural networks is particularly remarkable because it is shown that SGD performs greatly on various non-convex functions as well. SGD-based training algorithms have been achieving state-of-the-art one after another, e.g., Adagrad (Duchi, Hazan, and Singer 2011), Adam (Kingma and Ba 2014) and many others (Schmidt, Schneider, and Hennig 2021). Although there is a study showing that SGD is not a unique approach for generalization (Geiping et al. 2022), SGD remains one of the standard methods in the long run because of its computational convenience.

SGD's Noise Analyzing SGD's noise has been an appealing topic in the research community. It is known that the magnitude of the gradient noise in SGD has versatile effects on its dynamics (Kleinberg, Li, and Yuan 2018) thus it has been closely investigated especially in relation to a learning rate and batch size. An effect of large batch sizes on the reduction of gradient noise is investigated in (Hoffer, Hubara, and Soudry 2017; Smith et al. 2018; Masters and Luschi 2018). Another area of interest is the shape of a gradient noise distribution. (Zhu et al. 2018; Hu et al. 2017; Daneshmand et al. 2018) investigated the anisotropic nature of gradient noise and its advantage. (Simsekli, Sagun, and Gurbuzbalaban 2019) discussed the fact that a gradient noise distribution has a heavier tail than Gaussian distributions. (Nguyen et al. 2019; Şimşekli et al. 2019) showed the benefits of these heavy tails for SGD. (Panigrahi et al. 2019) rigorously examined gradient noise in deep learning and how close it is to a Gaussian. (Xie, Sato, and Sugiyama 2020) studied a situation where the distribution is Gaussian, and then analyzes the behavior of SGD in a theoretical way.

Discretization of SGD We summarize the approximation we used in Table 3. We used the continuous SGD (3) as an approximation of the discrete SGD (2) because (3) is exactly discretized to (2). This approximation is commonly used because it is well known that the trajectories of those two system show, so-called, "strong convergence" in the order of $O(\sqrt{\eta})$, i.e. $\mathbb{E}(\sup_{0\leq t\leq T}|\theta_k^{\text{discrete}} - \theta_{k\eta}|) = O(\eta^{\frac{1}{2}})$ (see e.g. (Gobet 2016; Cheng et al. 2020)). We note that strong convergence validates the similarity of trajectories, but it does not necessarily guarantee the similarity of escaping behavior. Our work is the first completed argument with Lemma 3 introduced.

As a final remark, (2) is also an approximated model of the original SGD (the dotted arrow in Fig. 3). Although this approximation is justified via the central limit theorem (Jastrzębski et al. 2017; He, Liu, and Tao 2019), it is admittedly heuristic and the quantitative validation for the approximation is assumed to be a (highly non-trivial) open problem.

B Assumptions of Fundamental Theorem

Although there are miscellaneous assumptions (Assumption 6, 7, and 8) in (Dembo and Zeitouni 2010, Theorem 5.7.11 (a)), they can be simply derived from the assumptions in SGD's escape problem. We provide the following brief justification.

Stability of θ^* is an essential assumption for (Dembo and Zeitouni 2010, Theorem 5.7.11 (a)) and especially commonly assumed in dynamical system (Hu et al. 2017; Wu, Zhu et al. 2017).

Assumption 6 (θ^* is asymptotically stable)**.** *For any neighborhood U that contains θ^*, there exists a small neighborhood V of θ^* such that gradient flow with any initial value $\theta_0 \in V$ does not leave U for $t \geq 0$ and $\lim_{t\to\infty}\theta_t = \theta^*$.*

Assumption 7 (D is attracted to θ^*)**.** *$\forall \theta_0 \in D$, a system $\dot{\theta}_t = -\nabla L(\theta_t)$ with initial value θ_0 converges to θ^* without leaving D as $t \to \infty$.*

But it does not explicitly appear in SGD's escaping analysis (Zhu et al. 2018; Jastrzębski et al. 2017; Xie, Sato, and Sugiyama 2020), because they usually adopt stronger assumptions. Assumption 6 is known to be equivalent to the local minimality of θ^* under the condition that $L(\theta)$ is real analytic around θ^* (Absil and Kurdyka 2006). Also, by definition of asymptotic stability in Assumption 6, we can always find a region D that satisfies Assumption 7. The more detailed properties of stability can be found, such as in (Teschl 2000, Section 6.5).

For the rigorous asymptotic analysis, the quasi-potential needs to be finite.

Assumption 8 (Finite V_0)**.** $V_0 \triangleq \inf_{\theta' \in \partial D} V(\theta') < \infty$.

But this is naturally satisfied in our setup. Because of Assumption 1 and 3, we can choose $\varphi_t = \nabla L(\varphi_t)$ to obtain finite steepness. This implies the quasi-potential is always finite for any θ.

Finally, to order to ensure the continuity, we need to impose the following assumption.

Assumption 9. *There exists an $N < \infty$ such that, for all $\rho > 0$ small enough and all x, y with $|x - z| + |y - z| \leq \rho$ for some $z \in \partial D \cup \{\theta^*\}$, there is a function u satisfying that $\|u\| < N$ and $\varphi_{T(\rho)} = y$, where*

$$\phi_t = x - \int_0^t \nabla L\left(\varphi_s\right) ds + \int_0^t C^{1/2}\left(\varphi_s\right) u_s ds$$

and $T(\rho) \to 0$ as $\rho \to 0$.

This can be derived from Assumption 2.

C Formal properties of Steepness and Quasi-potential

Fundamental Lemmas of Steepness Formally, steepness (Definition 4) is a useful measure because it satisfies the following Lemmas 4 and 5. To state the lemmas, readers shall view θ as a probability measure on $\mathbb{R}^d$, i.e. $\theta := (\mathbb{R}^d, \mathcal{F})$, where $\mathcal{F}$ is the Borel σ-field on $\mathbb{R}^d$.

Lemma 4. *If all the entries of $\nabla L(\cdot)$ and $C(\cdot)$ are bounded, uniformly Lipschitz continuous functions, then given $\{\theta_t\}$, the solution of (3), for all $\Gamma \in \mathcal{F}$*

$$\liminf_{\varepsilon \to 0} \varepsilon \ln \theta(\Gamma) \geq -\inf_{\varphi \in \Gamma^o} S_T(\varphi)$$

where Γ^o denotes the interior of Γ.

Lemma 5. *If all the entries of $\nabla L(\cdot)$ and $C(\cdot)$ are bounded, uniformly Lipschitz continuous functions, then given $\{\theta_t\}$, the solution of (3), for all $\Gamma \in \mathcal{F}$*

$$\limsup_{\varepsilon \to 0} \varepsilon \ln \theta(\Gamma) \leq -\inf_{\varphi \in \bar{\Gamma}} S_T(\varphi)$$

where $\bar{\Gamma}$ denotes the closure of Γ.

Proof. Definition 4 corresponds to the rate function defined in (Dembo and Zeitouni 2010, (5.6.6)) in our setup (see the (Dembo and Zeitouni 2010, Remark on p214) as well). By (Dembo and Zeitouni 2010, Theorem 5.6.7), our steepness satisfies Large Deviation Principle, which is stated in (Dembo and Zeitouni 2010, (1.2.4)). This statement immediately proves Lemma 4 and 5. □

Preliminary Lemmas of Quasi-potential Derived from Lemma 4 and 5, the following lemmas play essential roles in bounding the mean exit time. We introduce the statements in our notations and briefly describe their implications.

In the lemmas below, we use Θ_t to denote gradient flow, i.e., $\dot{\Theta}_t = -\nabla L(\Theta_t)$ and for $\mu > 0$, $\mathcal{B}_\mu$ denotes an μ-neighbourhood of θ^*, that is, $\mathcal{B}_\mu := \{\theta' \in \mathbb{R}^d \mid \|\theta' - \theta^*\| \leq \mu\}$. We also define

$$\pi_\mu \triangleq \inf\left\{t : t \geq 0, \theta_t \in \mathcal{B}_\mu \cup \partial D\right\}.$$

Lemma 6 (Lemma 5.7.18 in (Dembo and Zeitouni 2010))**.** *Under Assumption 9, for any $\xi > 0$ and any $\mu > 0$ small enough, there exists $T_0 < \infty$ such that*

$$\liminf_{\varepsilon \to 0} \varepsilon \ln \inf_{\theta_0 \in \mathcal{B}_\mu} \mathrm{P}\left(\tau \leq T_0\right) > -(V_0 + \xi)$$

Lemma 6 means if ε is small enough, there exists a trajectory that starts from $\mathcal{B}_\mu$, exits with finite time duration T_0, and has its steepness less than $V_0 + \xi$.

Lemma 7 (Lemma 5.7.19 in (Dembo and Zeitouni 2010))**.** *Under Assumption 1, 2, and 4, for any $\mu > 0$ small enough such that $\mathcal{B}_\mu \subset D$, we have*

$$\lim_{t \to \infty} \limsup_{\varepsilon \to 0} \varepsilon \ln \sup_{\theta_0 \in D} \mathrm{P}\left(\pi_\mu > t\right) = -\infty.$$

Lemma 7 means if ε is small enough and t is large enough, θ will either fall into $\mathcal{B}_\mu$ or each ∂D at some point.

Lemma 8 (Lemma 5.7.21 in (Dembo and Zeitouni 2010))**.** *Under Assumption 1 and 3, for any $\mu > 0$ small enough such that $\mathcal{B}_\mu \subset D$, for any closed set $N \subset \partial D$*

$$\lim_{\mu \to 0} \limsup_{\varepsilon \to 0} \varepsilon \ln \sup_{\theta_0 \in \mathcal{B}_\mu} \mathrm{P}\left(\theta_{\pi_\mu} \in N\right) \leq -\inf_{\theta' \in N} V(\theta'),$$

Lemma 8 means if ε is small enough and $\mathcal{B}_\mu$ is a small enough neighborhood of θ^*, there exists a trajectory from $\mathcal{B}_\mu$ to N such that its steepness at least $\inf_{\theta' \in N} V\left(\theta'\right)$.

Lemma 9 (Lemma 5.7.22 in (Dembo and Zeitouni 2010))**.** *For any $\mu > 0$ small enough such that $\mathcal{B}_\mu \subset D$ and all $\theta_0 \in D$, we have*

$$\lim_{\varepsilon \to 0} \mathrm{P}\left(\theta_{\pi_\mu} \in \mathcal{B}_\mu\right) = 1$$

Lemma 9 means if ε is small enough, θ eventually falls into $\mathcal{B}_\mu$.

Lemma 10 (Lemma 5.7.23 in (Dembo and Zeitouni 2010))**.** *For every $\delta > 0$ and every $\xi > 0$, there exists a constant $T(\xi, \delta) < \infty$ such that*

$$\limsup_{\varepsilon \to 0} \varepsilon \ln \sup_{\theta_0 \in D} \mathrm{P}\left(\sup_{t \in [0, T(\xi,\delta)]} |\theta_t - \theta_0| \geq \delta\right) < -\xi$$

Lemma 10 means over short time intervals, θ_t can get far from its starting point with an exponentially small probability.

D Deferred Proof of Lemma 2

Proof. First, we use $\lambda_{\max}^{-1}\widehat{S}_T(\varphi)$ as a "proxy steepness" to estimate $S(\varphi)$ and $V(\theta)$. For any trajectory φ that defines quasi-potential, the following bound holds.

$$\begin{aligned}
&\left|S_T(\varphi) - \lambda_{\max}^{-1}\widehat{S}_T(\varphi)\right| \\
&= \left|\frac{1}{2}\int_0^T \Phi^\top \left(C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I\right)\Phi dt\right| \\
&\text{(where } \Phi := \left(\dot{\varphi}_t + \nabla L\left(\varphi_t\right)\right)) \\
&\le \frac{1}{2}\int_0^T \left|\Phi^\top \left(C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I\right)\Phi\right| dt
\end{aligned} \tag{5}$$

Since $C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I$ is positive semi-definite,

$$\begin{aligned}
(5) &\le \frac{1}{2}\int_0^T \left\|\Phi\right\|^2 \lambda_{\max}\left(C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I\right) dt \\
&= \frac{1}{2}\int_0^T \left\|\dot{\varphi}_t + \nabla L\left(\varphi_t\right)\right\|^2 \lambda_{\max}\left(C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I\right) dt
\end{aligned} \tag{6}$$

Since D is a finite set and $L(\theta)$ is a locally quadratic function (Assumption 1), there exists a constant $M > 0$ that satisfies $\forall\theta \in D : \|\nabla L(\theta)\| \le M$. we can further obtain the following bound.

$$\begin{aligned}
(6) &\le \frac{T}{2}(K+M)^2 \sup_{0\le t\le T}\left\{\lambda_{\max}\left(C\left(\varphi_t\right)^{-1} - \lambda_{\max}^{-1} I\right)\right\} \\
&\quad (\because \|\nabla L(\theta)\| \le M \text{ and Assumption 5}) \\
&= \frac{T}{2}(K+M)^2 \sup_{0\le t\le T}\left\{\lambda_{\max}\left(C\left(\varphi_t\right)^{-1}\right) - \lambda_{\max}^{-1}\right\} \\
&= \frac{T}{2}(K+M)^2 \left\{\sup_{0\le t\le T}\lambda_{\max}\left(C\left(\varphi_t\right)^{-1}\right) - \lambda_{\max}^{-1}\right\}
\end{aligned} \tag{7}$$

The following inequalities show that $\sup_{0\le t\le T}\lambda_{\max}(C(\varphi_t)^{-1})$ is close to $\lambda_{\min}^{-1}(= \lambda_{\max}(C(\theta^*)^{-1})$ by Assumption 2).

For any $\theta \in \{\varphi_t\}_{0\le t\le T}$

$$\begin{aligned}
&\lambda_{\max}\left(C\left(\theta\right)^{-1}\right) - \lambda_{\max}\left(C\left(\theta^*\right)^{-1}\right) \\
&= \left\|C\left(\theta\right)^{-1}\right\|_{\text{op}} - \left\|C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&\le \left\|C\left(\theta\right)^{-1} - C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&= \left\|C\left(\theta\right)^{-1}\left(C\left(\theta^*\right) - C\left(\theta\right)\right)C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&\le \left\|C\left(\theta\right)^{-1}\right\|_{\text{op}} \left\|C\left(\theta^*\right) - C\left(\theta\right)\right\|_{\text{op}} \left\|C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&\le \left\|C\left(\theta\right)^{-1}\right\|_{\text{op}} \left\|C\left(\theta^*\right) - C\left(\theta\right)\right\|_{\text{F}} \left\|C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&\le \left\|C\left(\theta\right)^{-1}\right\|_{\text{op}} C_0\left|\theta^* - \theta\right|_\infty \left\|C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&(\because \text{There exists } C_0 > 0 \text{ by Assumption 3}) \\
&\le \left\|C\left(\theta\right)^{-1}\right\|_{\text{op}} C_0 r \left\|C\left(\theta^*\right)^{-1}\right\|_{\text{op}} \\
&\text{(There exists a constant radius, } r\text{, because } D \text{ is finite.)} \\
&\le C_0 r k \lambda_{\min}^{-1} \quad (\because \text{Assumption } 4)
\end{aligned}$$

Thus,

$$(7) \le \frac{T}{2}(K+M)^2\left(\left(1 + C_0 r k\right)\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right).$$

With this upper bound, V_0 can also be bounded in the followings. By definition,

$$\begin{aligned}
V_0 &= \inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi:(\varphi_0,\varphi_T)=(\theta^*,\theta)} S_T(\varphi) \\
\widehat{V}_0 &= \inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi:(\varphi_0,\varphi_T)=(\theta^*,\theta)} \widehat{S}_T(\varphi)
\end{aligned}$$

From here below, we denote $\inf_{\varphi:(\varphi_0,\varphi_T)=(\theta^*,\theta)}$ by $\inf_{\varphi(\theta,T)}$ for brevity.

Since ∂D is a continuous finite boundary, we have the following $\theta^\dagger$ and θ^*.

$$\begin{aligned}
\theta^\dagger &:= \operatorname*{arginf}_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} S_T(\varphi) \\
\theta^* &:= \operatorname*{arginf}_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} \widehat{S}_T(\varphi).
\end{aligned}$$

The followings hold.

$$\begin{aligned}
&\inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} S_T(\varphi) - \inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} \widehat{S}_T(\varphi) \\
&\qquad \le \inf_{T>0}\inf_{\varphi(\theta^*,T)} S_T(\varphi) - \inf_{T>0}\inf_{\varphi(\theta^*,T)} \widehat{S}_T(\varphi) \\
&\inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} \widehat{S}_T(\varphi) - \inf_{\theta\in\partial D}\inf_{T>0}\inf_{\varphi(\theta,T)} S_T(\varphi) \\
&\qquad \le \inf_{T>0}\inf_{\varphi(\theta^\dagger,T)} \widehat{S}_T(\varphi) - \inf_{T>0}\inf_{\varphi(\theta^\dagger,T)} S_T(\varphi)
\end{aligned}$$

Similarly, we can the following finite $T^\dagger$ and T^* for each θ.

$$\begin{aligned}
T^\dagger(\theta) &:= \operatorname*{arginf}_{T>0}\inf_{\varphi(\theta,T)} S_T(\varphi) \\
T^*(\theta) &:= \operatorname*{arginf}_{T>0}\inf_{\varphi(\theta,T)} \widehat{S}_T(\varphi),
\end{aligned}$$

and the followings hold

$$\begin{aligned}
&\inf_{T>0}\inf_{\varphi(\theta^*,T)} S_T(\varphi) - \inf_{T>0}\inf_{\varphi(\theta^*,T)} \widehat{S}_T(\varphi)\\
&\le \inf_{\varphi(\theta^*,T^*(\theta^*))} S_{T^*(\theta^*)}(\varphi) - \inf_{\varphi(\theta^*,T^*(\theta^*))} \widehat{S}_{T^*(\theta^*)}(\varphi)\\
&\inf_{T>0}\inf_{\varphi(\theta^\dagger,T)} \widehat{S}_T(\varphi) - \inf_{T>0}\inf_{\varphi(\theta^\dagger,T)} S_T(\varphi)\\
&\le \inf_{\varphi(\theta^\dagger,T^\dagger(\theta^\dagger))} \widehat{S}_{T^\dagger(\theta^\dagger)}(\varphi) - \inf_{\varphi(\theta^\dagger,T^\dagger(\theta^\dagger))} S_{T^\dagger(\theta^\dagger)}(\varphi).
\end{aligned}$$

Similarly, since $L(\theta)$ and $C(\theta)$ are continuous, for each θ and T, we have

$$\begin{aligned}
\varphi^\dagger(\theta,T) &:= \operatorname*{arginf}_{\varphi(\theta,T)} S_T(\varphi)\\
\varphi^*(\theta,T) &:= \operatorname*{arginf}_{\varphi(\theta,T)} \widehat{S}_T(\varphi).
\end{aligned}$$

and we get

$$\begin{aligned}
&\inf_{\varphi(\theta^*,T^*(\theta^*))} S_{T^*(\theta^*)}(\varphi) - \inf_{\varphi(\theta^*,T^*(\theta^*))} \widehat{S}_{T^*}(\varphi)\\
&\le S_{T^*(\theta^*)}(\varphi^*(\theta^*,T^*(\theta^*))) - \widehat{S}_{T^*}(\varphi^*(\theta^*,T^*(\theta^*)))\\
&\le \frac{T^*(\theta^*)}{2}(K+M)^2\left((1+C_0rk)\,\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{\varphi(\theta^\dagger,T^\dagger(\theta^\dagger))} \widehat{S}_{T^\dagger(\theta^\dagger)}(\varphi) - \inf_{\varphi(\theta^\dagger,T^\dagger(\theta^\dagger))} S_{T^\dagger(\theta^\dagger)}(\varphi)\\
&\le \widehat{S}_{T^\dagger(\theta^\dagger)}(\varphi^\dagger(\theta^\dagger,T^\dagger(\theta^\dagger))) - S_{T^\dagger(\theta^\dagger)}(\varphi^\dagger(\theta^\dagger,T^\dagger(\theta^\dagger)))\\
&\le \frac{T^\dagger(\theta^\dagger)}{2}(K+M)^2\left((1+C_0rk)\,\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right)
\end{aligned}$$

Thus we get

$$\left|V_0 - \lambda_{\max}^{-1}\widehat{V}_0\right| \le A\left((1+C_0rk)\,\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\right),$$

where $A := \frac{\max\{T^\dagger(\theta^\dagger),T^*(\theta^*)\}}{2}(K+M)^2$. □

E Deferred Proof of Theorem 3

Proof. For the continuous SGD, by Theorem 1, we have $\lim_{\eta\to 0}\frac{\eta}{B}\ln\mathbb{E}[\tau] = V_0$, where $V_0 = \min_{\theta'\in\partial D} V(\theta')$. With this result, it remains to evaluate the discretization error of exit time.

Here, without loss of generality, we assume $\mathbb{E}[\nu] > 1$. Also, we consider a case with $\mathbb{E}[\nu] - \mathbb{E}[\tau] \ge 0$. For the opposite case $\mathbb{E}[\nu] - \mathbb{E}[\tau] < 0$, we can obtain the same result by repeating the following proof. By Lemma 3, for sufficiently small η, there exists a constant c such that $0 \le \mathbb{E}[\nu] - \mathbb{E}[\tau] \le c\sqrt{\eta}$ holds. Therefore, the discrete exit time can be lower-bounded as

$$\frac{\eta}{B}\ln\mathbb{E}[\nu] \ge \frac{\eta}{B}\ln\mathbb{E}[\tau],$$

and also upper-bounded as

$$\begin{aligned}
\frac{\eta}{B}\ln\mathbb{E}[\nu] &\le \frac{\eta}{B}\ln\left(\mathbb{E}[\tau] + c\sqrt{\eta}\right) = \frac{\eta}{B}\ln\left(1 + \mathbb{E}[\tau] - 1 + c\sqrt{\eta}\right)\\
&\le \frac{\eta}{B}\ln\left(\mathbb{E}[\tau]\right) + \frac{\eta}{B}\ln\left(1 + c\sqrt{\eta}\right).
\end{aligned}$$

The last inequality follows that $\log(1+a+b) \le \log(1+a) + \log(1+b)$ for any $a, b > 0$. Using the lower and upper bound, we obtain

$$\lim_{\eta\to 0}\frac{\eta}{B}\ln\mathbb{E}[\nu] = \lim_{\eta\to 0}\left\{\frac{\eta}{B}\ln\left(\mathbb{E}[\tau]\right) + \frac{\eta}{B}\ln\left(1 + c\sqrt{\eta}\right)\right\} = V_0.$$

Combined with Lemma 2 and Theorem 1, we obtain the statement of Theorem 3. □

F Numerical Validation

We provide numerical experiments to validate our result under practical scenarios. We use a multi-layer perceptron with one hidden layer with 5000 units, mean square loss function, fed with the AVILA dataset (De Stefano et al. 2011). To obtain the local minimum θ^*, we run the gradient descent network for a sufficiently long time (1000 epochs) to obtain asymptotically stable θ^*. The region D is defined as a neighborhood of θ^*. With θ^* as an initial value, we measure the exit times with SGD 100 times independently. We measure the average number of steps at which SGD exits from D as the discrete mean exit time. To observe the dependency on the essential hyper-parameters ($\lambda_{\max}$, η, B, and ΔL,), we compute the Pearson correlation coefficient, i.e. the linear correlation. The sharpness of θ^* is controlled by mapping $L(\theta)$ to $L(\sqrt{\alpha}\theta)$ with a parameter $\alpha > 0$. Since this mapping changes $\lambda_{\max}$ to $\alpha\lambda_{\max}$ with other properties remaining the same, we use α as a surrogate of the sharpness $\lambda_{\max}$. In a similar manner, ΔL is controlled by mapping $L(\theta)$ to $\beta L(\theta)$, where, β is a surrogate of the depth of a minimum ΔL.

Fig. 4 shows the discrete mean exit time has an exponential dependency on $\lambda_{\max}^{-1}$, η^{-1}, B, and ΔL, which is aligned with Theorem 3. As a reference, we provide the same experiment with $C(\theta) \mapsto I$ (i.e. (4)). We also conducted a reference experiment to verify Proposition 1. In contrast to Fig. 4, Fig. 5 shows the discrete mean exit time is independent of sharpness while η and ΔL show the same trend. All the codes are available. [1]

Our experiment is limited to the one-hidden layer model because Assumption 4 is proven to hold in the simple model Zhong et al. (2017). Although this experiment shows that our theory is valid in the practical scale of time steps ν and learning rate η, we admit that it would be challenging to obtain a compelling result with large-scale architecture such as transformer (Vaswani et al. 2017) and ResNets (He et al. 2016).

G Proof of Theorem 1

Main proof For simplicity, we use ε to denote η/B. To prove this result, we provide the proof for an upper bound (Lemma 11) and a lower bound (Lemma 12). Preliminary Lemmas that support the proofs are summarized in Appendix C (Lemmas 6, 7, 8, 9, and 10).

Throughout the proofs, we sometimes use P_{θ_0} or $\mathrm{P}_{\theta_0'}$ to clearly indicate which trajectory we are referring to.

First, we develop the upper bound on the mean exit time.

[1] Source code: https://github.com/ibayashi-hikaru/SGD_exit_time.

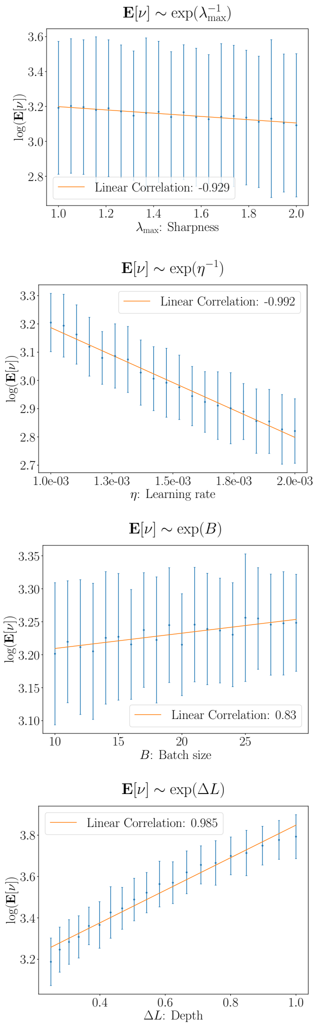


Figure 4: Numerical validation of Theorem 3, where the mean exit time shows exponential dependency on $\lambda_{\max}^{-1}$, η^{-1}, B, and ΔL. The error bars indicate the standard deviation.

Lemma 11. *For any $c > 0$, there exists an ε_0 such that for $\varepsilon < \varepsilon_0$, $\varepsilon \ln \mathbb{E}[\tau] < V_0 + c$ holds, where $V_0 := \min_{\theta' \in \partial D} V(\theta')$.*

Proof. Given Lemma 6 with $\xi := \frac{c}{2}$ and $\mu := \mu_0$, there exists a T_0 such that

$$\liminf_{\varepsilon \to 0} \varepsilon \ln \inf_{\theta_0 \in \mathcal{B}_{\mu_0}} \mathrm{P}\left(\tau \le T_0\right) > -(V_0 + \frac{c}{2})$$

Also, given Lemma 7 with $\mu = \mu_0$, there exists a T_1 such that

$$\limsup_{\varepsilon \to 0} \varepsilon \ln \sup_{\theta_0 \in D} \mathrm{P}\left(\tau_{\mu_0} > T_1\right) < 0.$$

Let $T = T_0 + T_1$. Then there exists some $\varepsilon_0 > 0$ such that for all $\varepsilon \le \varepsilon_0$,

$$\begin{aligned} q \triangleq \inf_{\theta_0 \in D} \mathrm{P}\left(\tau \le T\right) &\ge \inf_{\theta_0 \in D} \mathrm{P}\left(\tau_{\mu_0} \le T_1\right) \inf_{\theta_0 \in \mathcal{B}_{\mu_0}} \mathrm{P}\left(\tau \le T_0\right) \\ &\ge e^{-(V_0 + \frac{c}{2})/\varepsilon} \end{aligned}$$

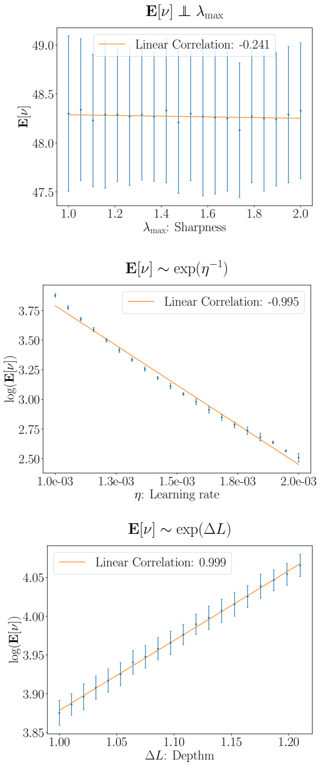


Figure 5: Reference experiment using the proxy system (4). Different from the result of SGD, the exit time has no dependency on $\lambda_{\max}$ while it has exponential dependency on η and ΔL similarly to SGD. This result is aligned with Proposition 1.

Considering the events $\{\tau > kT\}$ for $k = 1, 2, \ldots$ yields

$$\begin{aligned} \mathrm{P}\left(\tau > (k+1)T\right) &= \left[1 - \mathrm{P}\left(\tau \le (k+1)T \mid \tau > kT\right)\right] \mathrm{P}\left(\tau > kT\right) \\ &\le (1-q)\mathrm{P}\left(\tau > kT\right) \end{aligned}$$

Iterating over $k = 1, 2, \ldots$ gives

$$\sup_{\theta_0 \in D} \mathrm{P}\left(\tau > kT\right) \le (1-q)^k$$

Therefore,

$$\begin{aligned} \sup_{\theta_0 \in D} \mathbb{E}\left[\tau\right] &\le T\left[1 + \sum_{k=1}^{\infty} \sup_{\theta_0 \in D} \mathrm{P}\left(\tau > kT\right)\right] \\ &\le T \sum_{k=0}^{\infty} (1-q)^k = \frac{T}{q} \end{aligned}$$

and since $q \ge e^{-(V_0+c)/\varepsilon}$,

$$\sup_{\theta_0 \in D} \mathbb{E}\left[\tau\right] \le T e^{(V_0 + \frac{c}{2})/\varepsilon}$$

If we take ε_0 small enough,

$$\varepsilon \ln \mathbb{E}\left[\tau\right] \le V_0 + \frac{c}{2} < V_0 + c$$

holds for all $\varepsilon \le \varepsilon_0$ and $\theta_0 \in D$.

□

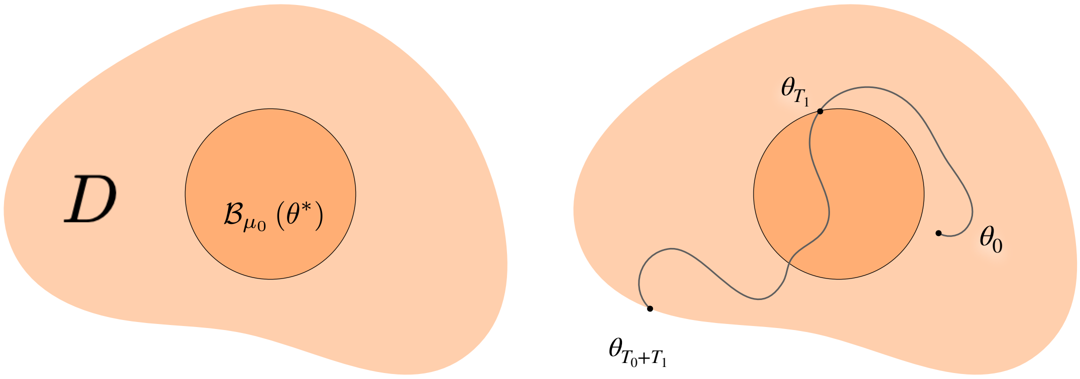
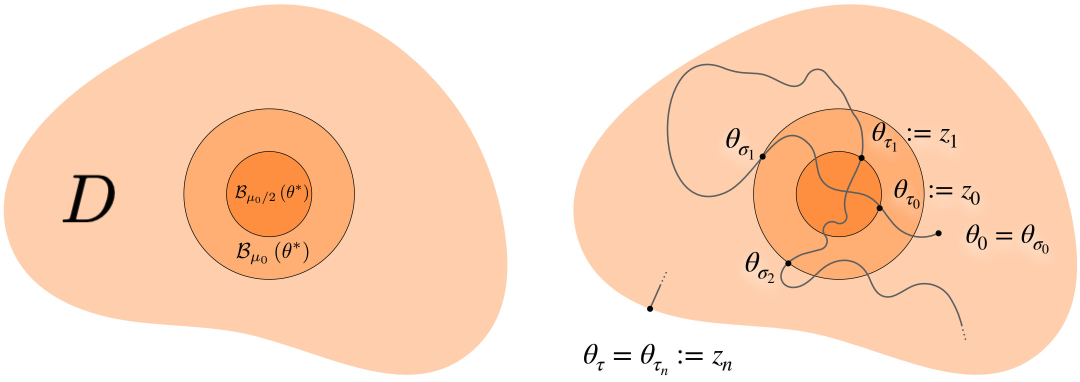


Figure 6: Domains and trajectory right appearing in the proof of upper bound.



D
$\mathcal{B}_{\mu_0/2}(\theta^*)$
$\mathcal{B}_{\mu_0}(\theta^*)$
θ_{σ_1}
$\theta_{\tau_1} := z_1$
$\theta_{\tau_0} := z_0$
$\theta_0 = \theta_{\sigma_0}$
θ_{σ_2}
$\theta_\tau = \theta_{\tau_n} := z_n$

Figure 7: Domains and trajectory right appearing in the proof of lower bound.

Next, we develop the lower bound on the exit time.

Lemma 12. *For any $c > 0$, there exists an ε_0 such that for $\varepsilon < \varepsilon_0$, $\varepsilon \ln \mathbb{E}[\tau] > V_0 - c$ holds, where $V_0 := \min_{\theta' \in \partial D} V(\theta')$.*

Proof. To prove the lower bound on τ, let $\mu_0 > 0$ be small enough that $\mathcal{B}_{\mu_0} \subset D$. Let $\sigma_0 = 0$ and for $n = 0, 1, \ldots$ define the following series of time stamps

$$\tau_n = \inf\left\{t : t \geq \sigma_n, \theta_t \in \mathcal{B}_{\mu_0/2} \cup \partial D\right\},$$
$$\sigma_{n+1} = \inf\left\{t : t > \tau_n, \theta_t \in \mathcal{B}_{\mu_0}\right\}$$

with the convention that $\sigma_{n+1} = \infty$ if $\theta_{\tau_n} \in \partial D$. Note that necessarily $\tau = \tau_n$ for some integer n. Moreover, since τ_n are exit times and θ_t is a strong Markov process, the process $z_n \triangleq \theta_{\tau_n}$ is a Markov chain. Note that ∂D is a closed set and choose $\mu_0 > 0$ small enough as needed by Lemma 8 for

$$\limsup_{\varepsilon \to 0} \varepsilon \ln \sup_{\theta_0' \in \mathcal{B}_{\mu_0}} \mathrm{P}_{\theta_0'}\left(\theta_{\pi_{\mu_0}} \in \partial D\right) < -V_0 + \frac{c}{2}.$$

Now, let $\xi := V_0$ and $\delta := \mu_0$, and let $T_0 = T(V_0, \mu_0)$ be as determined by Lemma 10. Then there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \leq \varepsilon_0$ and all $n \geq 1$,

$$\sup_{\theta_0 \in D} \mathrm{P}\left(\tau = \tau_n\right) \leq \sup_{\theta_0' \in \mathcal{B}_{\mu_0}} \mathrm{P}_{\theta_0'}\left(\theta_{\pi_{\mu_0}} \in \partial D\right) \leq e^{-(V_0 - c/2)/\varepsilon}$$

and

$$\sup_{\theta_0 \in D} \mathrm{P}\left(\sigma_n - \tau_{n-1} \leq T_0\right) \leq \sup_{\theta_0 \in D} \mathrm{P}\left(\sup_{t \in [0, T_0]} |\theta_t - \Theta_t| \geq \mu_0\right)$$
$$\leq e^{-(V_0 - c/2)/\varepsilon}$$

The event $\{\tau \leq kT_0\}$ implies that either one of the first $k+1$ among the mutually exclusive events $\{\tau = \tau_n\}$ occurs, or else that at least one of the first k excursions $[\tau_n, \tau_{n+1}]$ off $\mathcal{B}_{\mu_0/2}$ is of length at most T_0. Thus, by the union of events bound, utilizing the preceding worst-case estimates, for all $\theta_0 \in D$ and any integer k,

$$\mathrm{P}\left(\tau \leq kT_0\right) \leq \sum_{n=0}^{k} \mathrm{P}\left(\tau = \tau_n\right) + \mathrm{P}\left(\min_{1 \leq n \leq k} \{\sigma_n - \tau_{n-1}\} \leq T_0\right)$$
$$\leq \mathrm{P}\left(\tau = \tau_0\right) + 2ke^{-(V_0 - \frac{c}{2})/\varepsilon}$$

Recall the identity $\{\tau = \tau_0\} \equiv \left\{\theta_{\pi_{\mu_0}} \notin B_\rho\right\}$ and apply the preceding inequality with $k = \left[T_0^{-1} e^{(V_0 - \frac{c}{2})/\varepsilon}\right] + 1$ to obtain (for small enough ε)

$$\mathrm{P}\left(\tau \leq e^{(V_0 - \frac{c}{2})/\varepsilon}\right) \leq \mathrm{P}\left(\tau \leq kT_0\right) \leq \mathrm{P}\left(\theta_{\pi_{\mu_0}} \notin \mathcal{B}_{\mu_0}\right) + 4T_0^{-1} e^{-\frac{c}{2\varepsilon}}$$

By Lemma 9, the left side of this inequality approaches zero as $\varepsilon \to 0$; hence, the following holds.

$$\lim_{\varepsilon \to 0} \mathrm{P}\left(e^{(V_0 - \frac{c}{2})/\varepsilon} \leq \tau \right) = 1$$

By Chebycheff's bound, if we take ε_0 small enough,

$$\varepsilon \ln \mathbb{E}\left[\tau\right] > V_0 - c$$

holds for all $\varepsilon \leq \varepsilon_0$. □