QUASI-POTENTIAL THEORY FOR ESCAPE PROBLEM: QUANTITATIVE SHARPNESS EFFECT ON SGD'S ESCAPE FROM LOCAL MINIMA

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a quantitative theory on the escape problem of stochastic gradient descent (SGD) and investigate the effect of the sharpness of loss surfaces on escape. Deep learning has achieved tremendous success in various domains, however, it has opened up theoretical problems. For instance, it is still an ongoing question as to why an SGD can find solutions that generalize well over non-convex loss surfaces. An approach to explain this phenomenon is the escape problem, which investigates how efficiently the SGD escapes from local minima. In this paper, we develop a novel theoretical framework for the escape problem using "quasi-potential," the notion defined in a fundamental theory of stochastic dynamical systems. We show that quasi-potential theory can handle the geometric property of loss surfaces and a covariance structure of gradient noise in a unified manner through an eigenvalue argument, while previous research studied them separately. Our theoretical results imply that sharpness contributes to slowing down escape, but the SGDs noise structure cancels the effect, which ends up exponentially accelerating its escape. We also conduct experiments to empirically validate our theory using neural networks trained with real data.

1 INTRODUCTION

In recent years, the successes of deep learning have been a major driving force of machine learning development (LeCun, 2019). Owing to its strong generalization capability, deep learning has diverged into a wide range of domains, such as computer vision (Krizhevsky et al., 2012), speech recognition (Mikolov et al., 2011), and natural language processing (Collobert et al., 2011). This high performance of deep learning is underpinned by gradient-based learning algorithms, including stochastic gradient descent (SGD) and its variations (Kingma & Ba, 2014; Schmidt et al., 2021). However, at the same time, those unprecedented successes raise a question:

"Why does SGD learn parameters of neural networks with high generalization performance?"

Although non-convex optimization problems in neural networks have been thought to be difficult to solve (Blum & Rivest, 1992), SGD is known for finding nearly optimal solutions, and further, the obtained solutions generalize well Keskar et al. (2016); Brutzkus et al. (2017). Analyzing the role of SGD on the performance of deep learning is an area of research that is currently attracting strong interest (Masters & Luschi, 2018; Jastrzebski et al., 2021).

One of the promising directions to answer this question is to study the geometric properties of a training loss landscape. An increasing number of empirical studies have found that the minima obtained by SGD have distinctive geometric properties. Keskar et al. (2016) have shown that the shape of the minima obtained by SGD tends to be flat. He et al. (2019b) have deepened the investigation by picturing that SGD settles on the flatter side of asymmetric loss surface, which they named "asymmetric valley." Draxler et al. (2018) and Garipov et al. (2018) have shown that separate minima obtained by independent training processes are internally connected through pathways. Li et al. (2017) have proposed a dimension reduction technique to visualize the geometry of the loss surface, visually confirming flat minima. Most significantly, Jiang et al. (2019)'s large-scale experiment verified that minima in flat and wide regions have the strongest correlations with generalization capabilities. To

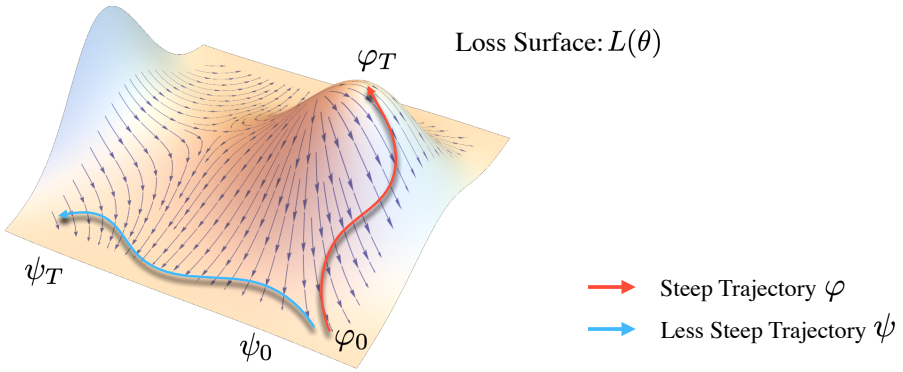


Figure 1: Visual illustration of steepness (Definition 3). The steepness of φ, $S_{0T}(\varphi)$, is greater than $S_{0T}(\psi)$ because φ climbs up to a higher point of loss surface.

attain a theoretical understanding of SGD, it is key to quantitatively analyze the connection between SGD and the geometric properties of the loss surface.

An *escape problem* is a scheme of analyzing the dynamic of SGD escaping from local minima of loss surfaces (Zhu et al., 2019; Jastrzębski et al., 2017; Hu et al., 2019; Nguyen et al., 2019; Xie et al., 2020). This scheme allows us to investigate why SGD avoids (potentially) bad local minima and settles on good minima. Zhu et al. (2019) first investigated the SGD's escape phenomenon and showed that SGD's escape is enhanced by its unique noise structure, called the "anisotropic noise structure." Invoked by their analysis, many studies have been attempting to theoretically quantify this phenomenon. Hu et al. (2019) rigorously identified the role of learning rate in escaping. Nguyen et al. (2019) used the Levy process to provide the precise description of SGD as well as its escaping phenomena. Jastrzębski et al. (2017) developed a theory of stochastic differential equation and quantified how the anisotropic noise affects its fast escape from sharp minima. Xie et al. (2020) refined the mathematical aspect and showed that the SGD's noise structure exponentially enhances escaping under a setup of diffusion theory.

In this paper, we apply the *quasi-potential theory* to the escape problem, and investigate a *mean exit time*, which formally quantifies escaping. The notion of quasi-potential is defined in a fundamental theory of stochastic dynamical systems, named a large deviation theory (Freidlin & Wentzell, 2012; Dembo & Zeitouni, 2010), and it is used to quantify a distribution of trajectories that a stochastic dynamical system takes. To illustrate quasi-potential for our problem setup, we introduce an intuitive notion, *steepness* of a trajectory (Fig. 1 and Definition 3), and show that it is an effective tool to analyze the escape of SGD. To the best of our knowledge, this is the first work that applies the quasi-potential to formalize the relationship between SGD's escape and the geometric properties of loss surface.

Our main findings and contributions are as follows:

- We develop the novel quasi-potential theory that rigorously describes the escape of SGD by a batch-size, a learning rate, and geometric parameters of loss surfaces. In particular, our theory can analyze gradient noise of SGD and sharpness of loss surfaces in a unified manner. For this, we use eigenvalues of Hessian matrices of the loss surfaces.

- We incorporate several practical settings that were not always covered in the previous theories: an effect of discrete update of SGD, state-dependent noise on gradients in SGD, and no assumption of the stationary distribution.

- We obtain a theoretical finding that a loss surface with sharp minima, i.e. with its Hessian matrices having large eigenvalues, is an obstacle to the escape of SGD. This is obtained by the unified analysis of gradient noise and loss surfaces by our theory, and has not been found in the existing studies.

Study	Time to escape	Non-stationary	Parameter dependent noise	Discrete setup
Hu et al. (2017)	$\propto \exp\left[\eta^{-1}\right]$	$\surd$	$\surd$	
Jastrzębski et al. (2017)	$\exp\left[\frac{B}{2\eta}\Delta L\right]\sqrt{\frac{\det H'}{\det H}}$		$\surd$	
Zhu et al. (2019)	N/A	$\surd$	$\surd$	
Nguyen et al. (2019)	$\frac{\alpha}{2}\frac{r^{\alpha}}{\sqrt{\eta}^{\alpha}}\left(1+\mathcal{O}\left(\eta^{\delta/2}\right)\right)$	$\surd$		$\surd$
Xie et al. (2020)	$\frac{2\pi}{\lvert\lambda'\rvert}\exp\left[2\frac{B}{\eta}\Delta L\left(\frac{s}{\lambda}+\frac{1-s}{\lvert\lambda'\rvert}\right)\right]$		$\surd$	
Ours (continuous SGD)	$\exp\left[2\frac{B}{\eta}r^2\lambda^{\frac{1}{2}}\right]$	$\surd$	$\surd$	
Ours (discrete SGD)	$\exp\left[2\frac{B}{\eta}r^2\lambda^{\frac{1}{2}}\right]+\mathcal{O}\left(\sqrt{\eta}\right)$	$\surd$	$\surd$	$\surd$

Table 1: Comparison of the studies on the escape problem. B is batch size, η is a learning rate, r is a radius of the region around a minimum, H is a Hesse matrix of loss functions at a minima, and $\lambda = \lambda_{\min}(H)$. Further, H' is a Hesse matrix on one of the neighboring points of the minimum and λ' is one of its eigenvalues of H'. ΔL is a difference of training loss values within a neighborhood of minimum, $\alpha \in (0, 2]$ is an index of heavy-tailedness of gradient noise in SGD, and $\delta \in (0, 1)$ and $s \in (0, 1)$ are values that implicitly include various factors of the escaping problem. "Non-stationary" denotes whether the result holds without assuming that SGD reaches a stationary distribution before escaping. "Parameter dependent noise" denotes whether noise in SGD depends on current parameters. "Discrete setup" means whether the analysis is valid with a discrete update by SGD. Our theory has two main advantages: (i) it explicitly quantifies various elements of SGD and loss surfaces without relying on auxiliary variables to incorporate them, such as s and δ, and (ii) it is applicable to a wide range of the settings.

1.1 Comparison with Existing Studies on Escape Problem

In Table 1, we compare the exit time derived with the results of the other studies that analyze the escape problem. There are two main points of focus. The first is the time to escape that we derive. Our theory realizes a unified analysis of exit time incorporating all the related parameters, batch size, learning rate, radius of the region around a minimum, and sharpness of a minimum. As a consequence, we show that the eigenvalues of the Hesse matrix increase the time to escape, which has not been found in the other studies. The second is the flexibility of the setting. Different from Jastrzębski et al. (2017) and Xie et al. (2020), our theory does not require that SGD reaches the stationary distribution before escaping, which is known to take exponentially many time steps (Xu et al., 2017; Raginsky et al., 2017). Additionally, our theory can evaluate the correspondence with the practical SGD, which has a discrete update rule and state-dependent noise.

Notations: For a $k \times k$ matrix A, $\lambda_j(A)$ is the j-th largest eigenvalue of A, and $\lambda_{\max}(A) = \lambda_1(A)$ and $\lambda_{\min}(A) = \lambda_k(A)$ denote the largest and the smallest eigenvalue of a square matrix. $\mathcal{O}(\cdot)$ 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 Setting and Problem

2.1 Stochastic Gradient Descent and Dynamical System

Consider a learning model parameterized by $\theta \in \mathbb{R}^d$. 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

$L^B(\theta) := \frac{1}{B}\sum_{x_i \in \mathcal{B}} \ell(\theta, x_i)$, where $\mathcal{B}$ is a randomly sampled subset of the training data such that $|\mathcal{B}| = B$. We assume that $L(\theta)$ is differentiable and its derivative $\nabla L(\theta)$ is Lipschitz continuous.

We mainly consider two types of SGD: a discrete SGD and a continuous SGD. Although a discrete SGD is used in practice, we study continuous SGD as a starting point of our analysis because of its mathematical convenience. This is a widely used approach in general SGD analyses (Ali et al., 2019; Advani et al., 2020) as well as in the escaping analyses (Jastrzębski et al., 2017; Xie et al., 2020).

Discrete SGD: First, we give the usual discrete formulation of SGD. Given an initial parameter $\theta_0 \in \mathbb{R}^d$, SGD generates a sequence of parameters $\{\theta_k\}_{k\in\mathbb{N}}$ by the following update rule:

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

for $k \in \mathbb{N}$, where $\eta > 0$ is a learning rate.

In particular, we focus on SGD whose noise on gradients has a Gaussian distribution. 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 consider a case that the noise is Gaussian. 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 $W_k \sim N\left(0, \eta C(\theta_k)\right)$ is a parameter-dependent Gaussian noise with its covariance $C(\theta) := \mathbb{E}_{i\sim \mathrm{Uni}(\{1,\ldots,N\})}\left[(\nabla L(\theta) - \nabla \ell\left(\theta, x_i\right))^\top (\nabla L(\theta) - \nabla \ell\left(\theta, x_i\right))\right]$.We assume that $C(\theta)$ is Lipschitz continuous.

The Gaussianity of the noise on gradients 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 noise, and (ii) several empirical studies show that the noise term becomes Gaussian noise (Mandt et al., 2016; Jastrzębski et al., 2017; He et al., 2019a), although different findings have been obtained in other settings (Simsekli et al., 2019).

Continuous SGD: We also give a continuous SGD, which is exactly discretized to (2) by a classic Euler scheme (Definition 5.1.1 of Gobet (2016)). With a time index $t \geq 0$ and the given initial parameter $\theta_0 \in \mathbb{R}^d$, the continuous dynamic of SGD is written as follows:

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

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 this system can be seen as a Gaussian perturbed dynamical system with a noise magnitude $\sqrt{\frac{\eta}{B}}$ because η and B do not evolve by time.

2.2 Escape Problem and Mean Exit Time

We consider the problem on how SGD escapes from minima of loss surfaces. In this paper, our target of interest is quantified by a notion of *mean exit time* for continuous SGD and *discrete mean exit time* of discrete SGD. Let $\theta^* \in \mathbb{R}^d$ be a local minimum of loss surfaces, and $D \subset \mathbb{R}^d$ be a r-neighborhood of θ^* with $r > 0$. We define the mean exit time as follows:

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

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

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

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

These definitions are common in quasi-potential theory (Freidlin & Wentzell, 2012; Gobet, 2016). Intuitively, the smaller $\mathbb{E}[\tau]$ or $\mathbb{E}[\nu]$ becomes, the faster the system escapes from a region D. In other words, the system has a stronger tendency to escape from θ^*.

We remark that there are other formulations to analyze the escape problem. Zhu et al. (2019) define escaping efficiency as $\mathbb{E}_{\theta_t}\left[L\left(\theta_t\right) - L\left(\theta_0\right)\right]$. Jastrzębski et al. (2017) and Xie et al. (2020) study a ratio between the probability of coming out from θ^*'s neighborhood and the probability mass around θ^*.

2.3 Basic Assumptions for the Escape Problem

We provide basic assumptions for the escape problem, commonly used in the literature (Mandt et al., 2016; Zhu et al., 2019; Jastrzębski et al., 2017; Xie et al., 2020).

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

$$\forall\theta \in D, L(\theta) = L(\theta^*) + \nabla L(\theta^*)(\theta - \theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*)$$

Assumption 2 (Hesse covariance matrix)**.** *For any $\theta \in D$, $C(\theta)$ is approximately equal to H^*.*

Although Assumption 2 has been commonly used, its validity is still being discussed (Section 2 in Xie et al. (2020) and Appendix A in Jastrzębski et al. (2017)). Thus, as further investigation, we consider a setup under the following relaxed assumption.

Assumption 3 (Relaxation of Assumption 2)**.** *There exist constants $0 < c_1 \leq c_2 < \infty$ such that for $\theta \in D$, $C(\theta) = H^*G$ holds with a symmetric matrix G as $0 < c_1 \leq \lambda_{\min}(G) \leq \lambda_{\max}(G) \leq c_2$.*

Assumption 3 allows a Hesse matrix in a neighborhood of minima to be a wide range of positive definite matrices.

3 Quasi-potential Theory

We introduce the basic notions of the quasi-potential theory. We start with defining a notion of *steepness* of a trajectory followed by the systems (3) on a loss surface $L(\theta)$. Let $\varphi = \{\varphi_t\}_{t\in[0,T]}$ be a trajectory over a finite time interval $[0, T]$. Or more formally, φ is a continuous map from $[0, T]$ to $\mathbb{R}^d$, and is an element of $\mathbf{C}_{0T}(\mathbb{R}^d)$ which is a support of a dynamical process in $[0, T]$. Given a trajectory φ and the system (3), we define the following quantity:

Definition 3 (Steepness)**.** *Steepness of a trajectory φ followed by (3) is defined as*

$$S_{0T}(\varphi) := \frac{1}{2}\int_0^T (\dot{\varphi}_t + \nabla L(\varphi_t))^\top C(\varphi_t)^{-1/2}(\dot{\varphi}_t + \nabla L(\varphi_t))\, dt.$$

Steepness $S_{0T}(\varphi)$ can be intuitively interpreted as the hardness of climbing that the system (3) is exposed to while following the trajectory φ on $L(\theta)$ (Fig. 1). This notion is generally utilized in the field of dynamical systems, for example, and is called “normalized action functional" in Section 3.2 of Freidlin & Wentzell (2012) and “rate function" in Section 1.2 of Dembo & Zeitouni (2010).

Steepness is useful to describe a distribution of trajectories generated by dynamical systems. If a trajectory φ has a large steepness $S_{0T}(\varphi)$, the probability that the dynamic system takes the trajectory decreases exponentially. Formally, the distribution is analyzed as follows.

Lemma 1 (Theorem 3.1 in Section 3.3 Freidlin & Wentzell (2012))**.** *For any $\delta, \zeta > 0, \varphi \in \mathbf{C}_{0T}\left(\mathbb{R}^d\right)$, and sufficiently small $\varepsilon > 0$, the following holds:*

$$\mathrm{P}_{\varphi'}\left(\varphi' \in \mathbf{C}_{0T}(\mathbb{R}^d) \mid \rho(\varphi', \varphi) < \delta\right) \geq \exp\left\{-\varepsilon^{-2}\left[S_{0T}(\varphi) + \zeta\right]\right\},$$

where $\rho(\varphi', \varphi) = \sup_{t\in[0,T]} \|\varphi'_t - \varphi_t\|$.

Lemma 2 (Theorem 3.1 in Section 3.3 Freidlin & Wentzell (2012))**.** *Let $\Phi(s) = \left\{\varphi \in \mathbf{C}_{0T}(\mathbb{R}^d) \mid S_{0T}(\varphi) \leq s\right\}$. For all $\delta, \zeta, s > 0$, and sufficiently small $\varepsilon > 0$, the following holds:*

$$\mathrm{P}_{\varphi'}\left\{\varphi' \in \mathbf{C}_{0T}(\mathbb{R}^d) \mid \rho(\varphi', \Phi(s)) \geq \delta\right\} \leq \exp\{-\varepsilon^{-2}(s - \zeta)\},$$

where $\rho(\varphi', \Phi(s)) = \inf_{\varphi\in\Phi(s)} \rho(\varphi', \varphi)$.

Although we restrict our attention to the system (3), the same discussion is applicable to a general class of diffusion processes and dynamical systems with Markov perturbations (For details, see section 5.7 in Dembo & Zeitouni (2010) or Section 6.5 in Freidlin & Wentzell (2012)).

Although there are several trajectories from θ^* to $\theta \in D$ with different steepness, a dominating factor for mean exit time is the smallest steepness among them, which is called *quasi-potential*:

Definition 4 (Quasi-potential)**.** *Quasi-potential of $\theta \in D$ is defined as*

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

Similar to steepness, quasi-potential can be seen as the minimum effort the system (3) needs to climb from θ^* up to θ on $L(\theta)$. (For more details, see Section 5.3 of Freidlin & Wentzell (2012)).

4 Mean Exit Time Analysis for SGD

4.1 Assumptions

To analyze the mean exit time, the quasi-potential theory requires several assumptions regarding the *stability* of the system (3) at θ^*.

Assumption 4 (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 5 (D is attracted to θ^*)**.** *$\forall \theta_0 \in D$, gradient flow with initial value θ_0 converges to θ^* without leaving D as $t \to \infty$.*

where "gradient flow" means a continuous gradient descent defined as $\dot{\theta}_t = -\nabla L(\theta_t)$.

Stability is a commonly used notion in dynamical systems (Hu et al., 2017; Wu et al., 2017), although it does not always appear in SGD's escaping analysis (Zhu et al., 2019; Jastrzębski et al., 2017; Xie et al., 2020). Assumption 4 is known to be equivalent to the local minimality of θ^* under the condition that $L(\theta)$ is real analytic around θ^* (Absil & Kurdyka, 2006). Also, by definition of asymptotic stability in Assumption 4, we can always find a region D that satisfies Assumption 5. The more detailed properties of stability can be found, such as in Section 6.5 of Teschl (2000). Assumption 4 and 5 are necessary to obtain the result (5) in the following section.

Also, we require the following assumption as a boundary condition of Theorem 4.

Assumption 6. $L(\theta^*) = 0$

Assumption 6 is only for simplifying our proofs without changing the essence of our problem.

4.2 Main Results

We analyze the mean escape time of SGD under the above assumptions. In preparation, we state two facts. First, $V(\theta)$ is calculated as a solution of the following equation:

$$\frac{1}{2}\nabla V(\theta)^\top C(\theta)^{1/2}\nabla V(\theta) - \nabla L(\theta)^\top \nabla V(\theta) = 0. \tag{4}$$

Second, if $\frac{B}{\eta}$ is sufficiently small, the mean exit time can be expressed using $V(\theta)$ as

$$\mathbb{E}[\tau] = \exp\left[\frac{B}{\eta}V_0\right], \tag{5}$$

where $V_0 := \min_{\theta' \in \partial D} V(\theta')$. Although these facts have been investigated in the literature (for example, see Hu et al. (2019) and Section 4.4 in Freidlin & Wentzell (2012)), we give our own theorems and proofs in Appendix A and B for completeness.

The followings are our main results. We start with the mean exit time of Continuous SGD. Let $\mathbb{E}[\tau_{\mathrm{SGD}}]$ be the mean exit time of the continuous SGD, and let $\mathbb{E}[\tau_{\mathrm{isoSGD}}]$ be the mean exit time of an isotropic continuous SGD whose $C(\theta)$ is set to I.

Theorem 1 (Continuous SGD)**.** *Suppose that Assumption 1, 2, 4, 5, and 6 hold. Then, for sufficiently small $\frac{\eta}{B}$,*

$$\mathbb{E}[\tau_{\mathrm{isoSGD}}] = \exp\left[2\frac{B}{\eta}r^2\lambda\right], \quad \mathbb{E}[\tau_{\mathrm{SGD}}] = \exp\left[2\frac{B}{\eta}r^2\lambda^{\frac{1}{2}}\right].$$

This result gives an exact expression for the expected escape time with the explicit values of SGD. The results also have two implications. First, these result both of those results show that the mean escape time exponentially increases in the smallest eigen value of H^*, i.e. λ. This implies that sharper minima generally slow down the escaping, which is seemingly opposite to the implication of the existing literature (Jastrzębski et al., 2017; Xie et al., 2020). But in fact this is consistent with the existing literature because some of the sharpness factor is implicitly included in other variables such as ΔL. Second, our result endorses the fact that SGD's anisotropic noise exponentially accelerates the escaping (Xie et al., 2020), because the result shows that the mean exit time of SGD is smaller than that of isotropic SGD by $\exp\left[\lambda^{\frac{1}{2}}\right]$.

We also develop additional result under the weaker assumption (Assumption 3).

Theorem 2 (Continuous SGD under weaker assumption)**.** *Suppose that Assumption 1, 3, 4, 5, and 6 hold. Then, for sufficiently small $\frac{\eta}{B}$,*

$$\begin{aligned} & \mathbb{E}[\tau_{\mathrm{isoSGD}}] &= \exp\left[2\frac{B}{\eta}r^2\lambda\right] \\ \exp\left[2\frac{B}{\eta}r^2\frac{1}{\sqrt{c_2}}\lambda^{\frac{1}{2}}\right] \leq & \mathbb{E}[\tau_{\mathrm{SGD}}] &\leq \exp\left[2\frac{B}{\eta}r^2\frac{1}{\sqrt{c_1}}\lambda^{\frac{1}{2}}\right] \end{aligned}$$

It claims that replacing Assumption 2 by Assumption 3 has no effect in the isotropic case, but has a constant effect in the anisotropic case.

Our theory can be extended to the discrete case. By $\mathbb{E}[\nu_{\mathrm{SGD}}]$, we denote the discrete mean exit time of the discrete SGD, and by $\mathbb{E}[\nu_{\mathrm{isoSGD}}]$ we denote the one of an isotropic version of the discrete SGD, i.e. with $C(\theta)$ being I. The escaping problem of a discrete SGD, the discrete mean exit time, is formulated as a special case of Gobet & Menozzi (2010). By substituting $g(\cdot) = 0, f(\cdot) = 1$, and $k(\cdot) = 0$ in Theorem 17 in Gobet & Menozzi (2010), we can obtain the following simplified statement.

$$\max\left\{\mathbb{E}[\nu_{\mathrm{isoSGD}}] - \mathbb{E}[\tau_{\mathrm{isoSGD}}], \mathbb{E}[\nu_{\mathrm{SGD}}] - \mathbb{E}[\tau_{\mathrm{SGD}}]\right\} = \mathcal{O}(\sqrt{\eta}).$$

which immediately prove the following theorem:

Theorem 3 (Discrete SGD)**.** *Given, Assumption 1, 2, 4, 5, and 6, for sufficiently small $\frac{\eta}{B}$,*

$$\mathbb{E}[\nu_{\mathrm{isoSGD}}] = \exp\left[2\frac{B}{\eta}r^2\lambda\right] + \mathcal{O}\left(\sqrt{\eta}\right), \quad \mathbb{E}[\nu_{\mathrm{SGD}}] = \exp\left[2\frac{B}{\eta}r^2\lambda^{\frac{1}{2}}\right] + \mathcal{O}\left(\sqrt{\eta}\right)$$

This result suggests that the discrete error does not majorly affect the escape. We note this is the first study that confirms the validity of using a continuous SGD model (3) for escape analysis.

We describe the proof for Theorem 2. We omit the proof of Theorem 1 because it is a special case of Theorem 2 ($c_1 = c_2 = 1$).

4.3 Proof Outline

We describe a proof of Theorem 2. We begin with the isotropic case and then investigate the non-isotropic case. The following lemmas are useful, whose proofs are in Appendix C.

Lemma 3. *Let A, B be $n \times n$ symmetric matrices whose eigen values are non-negatives. Then,*

$$\lambda_i(AB) \leq \lambda_i(A)\lambda_{j-i+1}(B) \quad \forall i \leq j$$

Lemma 4. *For positive definite symmetric matrices A and B, the following inequality holds*

$$\lambda_{\min}(AB) \geq \lambda_{\min}(A)\lambda_{\min}(B)$$

Isotropic case, $\mathrm{E}[\tau_{\mathrm{isoSGD}}]$: We substitute I to $C(\theta)$. By the Jacobi equation (4) which is formally given by Theorem 4 in Appendix A, we have the following form for $\theta \in D$:

$$\frac{1}{2}\nabla V(\theta)^\top \nabla V(\theta) - \nabla L\left(\theta\right)^\top \nabla V(\theta) = 0. \tag{6}$$

We have $\nabla V(\theta) = 2\nabla L(\theta)$ as a solution of (6). Given that $V(\theta^*) = 0$ by the definition of steepness and $L(\theta^*) = 0$ by Assumption 6, we obtain $V(\theta) = 2L(\theta)$ for $\theta \in D$. Therefore, we have

$$V_0 = \min_{x \in \partial D} 2L(\theta) = \min_{x \in \partial D} 2\theta^\top H^*\theta = 2r^2\lambda.$$

The second equality follows Assumption 1 and 6. Combined with the fact (5), which is formally shown in Theorem 5 in Appendix B, we obtain the statement of Theorem 2 in the anisotropic case.

Anisotropic case, $\mathrm{E}[\tau_{\mathrm{SGD}}]$**:** Similar to the isotropic case, the equation (4) (or Theorem 4) gives

$$\frac{1}{2}\nabla V(\theta)^\top C(\theta)^{\frac{1}{2}}\nabla V(\theta) - \nabla L(\theta)^\top \nabla V(\theta) = 0.$$

$\nabla V(\theta) = 2C(\theta)^{-\frac{1}{2}}\nabla L(\theta)$ is a solution of (6). Then, $\nabla V(\theta)$ is simply written as

$$\begin{aligned}\nabla V(\theta) &= 2C(\theta)^{-\frac{1}{2}}\nabla L(\theta) = 2C(\theta)^{-\frac{1}{2}}\nabla\left(\theta^\top H^*\theta\right) = 2C(\theta)^{-\frac{1}{2}}2H^*\theta \\ &= 2(H^*G)^{-\frac{1}{2}}2H^* = 4G^{-\frac{1}{2}}H^{*-\frac{1}{2}}H^*\theta = 4G^{-\frac{1}{2}}H^{*\frac{1}{2}}\theta.\end{aligned}$$

The second equation follows Assumption 1 and 6, and the fourth equation follows Assumption 3. Given that $V(\theta^*) = 0$, we obtain $V(\theta) = \theta^\top G^{-\frac{1}{2}}H^{*\frac{1}{2}}\theta$ for $\theta \in D$. Then, we rewrite V_0 is as

$$V_0 = 2r^2\lambda_{\min}(G^{-\frac{1}{2}}H^{*\frac{1}{2}}), \tag{7}$$

By Lemma 3, we develop an upper bound of $\lambda_{\min}(G^{-1/2}H^{*1/2})$ as $\lambda_{\min}(G^{-1/2}H^{*1/2}) = \lambda_{\min}(H^{*1/2}G^{-1/2}) \leq \lambda_{\min}(H^{*1/2})\lambda_{\max}(G^{-1/2}) \leq c_1^{-1/2}\lambda^{1/2}$. Similarly, Lemma 4 gives us the following lower bound $\lambda_{\min}(G^{-1/2}H^{*1/2}) \geq \lambda_{\min}(G^{-1/2})\lambda_{\min}(H^{*1/2}) \geq c_2^{-1/2}\lambda^{1/2}$. We substitute the two inequalities into the solution (7), then obtain the following form of V_0:

$$2r^2\frac{1}{\sqrt{c_2}}\lambda^{\frac{1}{2}} \leq V_0 \leq 2r^2\frac{1}{\sqrt{c_1}}\lambda^{\frac{1}{2}}$$

Combined with (5), or Theorem 5 in Appendix B, we finish the proof of Theorem 2. □

5 Experiment

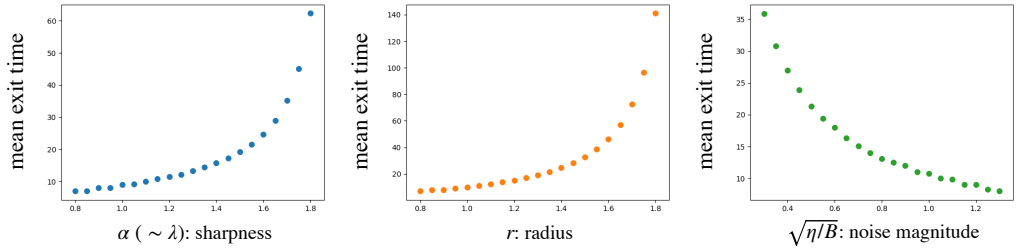


Figure 2: Empirical validation of Theorem 1, where the empirical mean exit time has exponential dependency on sharpness α ($\sim \lambda$), radius r, and noise magnitude $\sqrt{\eta/B}$.

We conduct an experiment to validate our result of discrete setup (Theorem 3), using a neural network and real-world datasets. We use a multi-layer perception and the AVILA dataset (De Stefano et al., 2011) to observe that the discrete mean exit time of SGD has exponential dependence on eigenvalue λ, radius r and a ratio of the learning rate and the batch size η/B.

In order for our essential assumptions to hold, we use the mean square loss with ℓ_2 regularizer for $L(\theta)$ (Assumption 1) and train the model with the gradient descent for a sufficiently long time to obtain θ_0 near θ^*. We set the r-neighborhood of θ_0 as D (Assumption 4 and 5). To measure the discrete mean exit time, we repeatedly execute a vanilla SGD from θ_0 for 1000 times and take an average number of steps at which SGD exit from D (i.e. when the distance from θ_0 becomes farther than r).

To control λ, we follow the approach of Xie et al. (2020). We obtain sharper minima by mapping the loss function L to L_α such that $L_\alpha(\theta) := L(\sqrt{\alpha}\theta)$ $(\alpha > 0)$ and setting $\theta_0 := \theta_0/\alpha$. Since this mapping coverts λ to $\alpha\lambda$ with other properties remaining the same, we use α as a surrogate of λ.

We show the results in Figure 2. As Theorem 3 suggests, the noise magnitude $\sqrt{\eta/B}$ exponentially accelerates the escaping under our experiment setup, and eigen value and radius have the effect

of exponentially slowing down the escaping The experiment can be reproduced by using the code below.[1]

6 Related Works

We summarize relevant studies related to the topics on loss surfaces and the stochastic gradient descent algorithm. We mainly consider the following three factors.

Loss surface shape: Shape of loss surfaces have long been a topic of interest. The argument that the flatness of loss surfaces around local minima improves generalization was first studied by Hochreiter & Schmidhuber (1995; 1997), and the observation has recently reconfirmed in deep neural networks by Keskar et al. (2016). Sagun et al. (2017) empirically examined the flatness of loss surfaces. The theoretical advantage of the flatness was criticized by Dinh et al. (2017) in terms of scale-sensitivity of flatness, but Tsuzuku et al. (2020) and Rangamani et al. (2019) tackled the criticism by developing scale invariant flatness measures. An effect of the shape of loss surfaces on SGD was investigated in Wu et al. (2017); Ge et al. (2018), and Chaudhari et al. (2019); Foret et al. (2020) developed a variant of SGD which made use of this fact. In addition to the flatness, He et al. (2019b) proposed a new notion of asymmetry of loss surfaces, and Draxler et al. (2018); Garipov et al. (2018) studied how several local minima in a loss surface are connected. Li et al. (2018) developed a random dimensional reduction method to visualize loss surfaces on high dimensional spaces.

Exit/Stability of SGD: How SGD behaves in neighborhoods of local minima in loss surfaces is investigated from two aspects: stability and escape efficiency. For stability, a way in which SGD finds local minima and stabilizes was analyzed by Wu et al. (2018); Kleinberg et al. (2018); Achille et al. (2019); Li et al. (2017). Smith & Le (2017) used Bayesian ideas to analyze the stability. For exiting aspects, Jastrzębski et al. (2017) investigated an effect of a Hesse matrix of local minima on the ease of escaping ineffective local minima, and Xie et al. (2020) elaborated this effect via quantitative analysis. Zhu et al. (2019) showed that anisotropic structure of gradient noise by SGD is useful in escaping inefficient local minima, and Nguyen et al. (2019) studied an effect of non-Gaussianity of the gradient noise.

SGD property: Detailed nature of SGD itself is also an object of interest. The magnitude of the gradient noise by SGD is an important factor, including its relation to a learning rate and a batch size. An effect of large batch sizes on the reduction of gradient noise is investigated in Hoffer et al. (2017); Smith et al. (2018); Masters & Luschi (2018). Another area of interest is shape of a gradient noise distribution. Zhu et al. (2019); Hu et al. (2017); Daneshmand et al. (2018) investigated the anisotropic nature of gradient noise and its advantage. Simsekli et al. (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 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 et al. (2020) studied a situation where the distribution is Gaussian, and then analyzes the behavior of SGD in a theoretical way.

7 Conclusion

In this paper, we develop a novel quasi-potential theory for the escape problem of SGD. Our theory gives an intuitive picture of SGD's escaping dynamic, and also but is an effective tool for formal analysis. In our main result, our theory explicitly describes how the escape of SGD is affected by a batch-size, a learning rate, and radius of regions, and sharpness (Theorem 1). Furthermore, due to its flexibility, our theory allows the extended analyses, such as SGD's escape under even weaker assumption on covariance matrix (Theorem 2) and the escape problem of a discrete SGD (Theorem 3). We believe our theory provides a solid insight for SGD dynamics and also flexible theory for further studies.

[1] experiment code place holder. Hidden for the anonymous review. See supplemental material

REFERENCES

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

Alessandro Achille, Giovanni Paolini, and Stefano Soatto. Where is the information in a deep neural network? *arXiv preprint arXiv:1905.12213*, 2019.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Netw.*, 132:428–446, December 2020.

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1370–1378. PMLR, 2019.

Avrim L Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

P Chaudhari, A Choromanska, S Soatto, and others. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12(ARTICLE): 2493–2537, 2011.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1155–1164. PMLR, 2018.

Claudio De Stefano, Francesco Fontanella, Marilena Maniaci, and Alessandra Scotto di Freca. A method for scribe distinction in medieval manuscripts using page layout features. In *International Conference on Image Analysis and Processing*, pp. 393–402. Springer, 2011.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2nd edition, 2010.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Mark I Freidlin and Alexander D Wentzell. *Random Perturbations of Dynamical Systems 3rd Ed*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8803–8812, 2018.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

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

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

Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32:1143–1152, 2019a.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in Neural Information Processing Systems*, 32:2553–2564, 2019b.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, pp. 529–536, 1995.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Elad Hoffer, Itay Hubara, and Daniel Soudry. 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*, pp. 1729–1739, 2017.

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705. 07562*, 2017.

Wenqing Hu, Zhanxing Zhu, Haoyi Xiong, and Jun Huan. Quasi-potential as an implicit regularizer for the loss function in the stochastic gradient descent. *arXiv preprint arXiv:1901.06054*, 2019.

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pp. 4772–4784. PMLR, 2021.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 25:1097–1105, 2012.

Yann LeCun. 1.1 deep learning hardware: Past, present, and future. In *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 12–19. ieeexplore.ieee.org, February 2019.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110. PMLR, 2017.

Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In Maria Florina Balcan and Kilian Q Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 354–363, New York, New York, USA, 2016. PMLR.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černocký. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 196–201. ieeexplore.ieee.org, December 2011.

Thanh Huy Nguyen, Umut Şimşekli, Mert Gürbüzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *arXiv preprint arXiv:1906.09069*, 2019.

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

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703. PMLR, 2017.

Akshay Rangamani, Nam H Nguyen, Abhishek Kumar, Dzung Phan, Sang H Chin, and Trac D Tran. A scale invariant flatness measure for deep network minima. *arXiv preprint arXiv:1902. 02434*, 2019.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Robin M Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*, pp. 9367–9376. PMLR, 2021.

Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A Tail-Index analysis of stochastic gradient noise in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837. PMLR, 2019.

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

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

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020.

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

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

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pp. 7654–7663. PMLR, 2019.

A JACOBI EQUATION OF QUASI-POTENTIAL

Theorem 4. *For all $\theta \in D$, $V(\theta)$ satisfies the following Jacobi equation,*

$$\frac{1}{2}\nabla V(\theta)^\top C(\theta)^{1/2}\nabla V(\theta) - \nabla L(\theta)^\top \nabla V(\theta) = 0$$

Proof. For simplicity, we introduce the inner product and the norm regarding a point $\theta \in D$, denoted as follows,

$$\langle u, v\rangle_\theta := u^\top C(\theta)^{-1/2} v \tag{8}$$

$$\|u\|_\theta := \sqrt{\langle u, u\rangle_\theta} \tag{9}$$

Given these definitions, the $S_{0T}(\varphi)$ can be expressed as follows,

$$S_{0T}(\varphi) = \frac{1}{2}\int_0^T \|\dot{\varphi}_t + \nabla L(\varphi_t)\|_{\varphi_t}^2 \, dt. \tag{10}$$

For the integrand of (10), the following inequality holds.

$$\|\dot{\varphi}_t + \nabla L(\varphi_t)\|_{\varphi_t}^2 \tag{11}$$

$$= \|\dot{\varphi}_t\|_{\varphi_t}^2 + \|\nabla L(\varphi_t)\|_{\varphi_t}^2 + 2\langle \dot{\varphi}_t, \nabla L(\varphi_t)\rangle_{\varphi_t} \tag{12}$$

$$= \left(\|\dot{\varphi}_t\|_{\varphi_t} - \|\nabla L(\varphi_t)\|_{\varphi_t}\right)^2 + 2\|\dot{\varphi}_t\|_{\varphi_t}\|\nabla L(\varphi_t)\|_{\varphi_t} + 2\langle \dot{\varphi}_t, \nabla L(\varphi_t)\rangle_{\varphi_t} \tag{13}$$

$$\geq 2\|\dot{\varphi}_t\|_{\varphi_t}\|\nabla L(\varphi_t)\|_{\varphi_t} + 2\langle \dot{\varphi}_t, \nabla L(\varphi_t)\rangle_{\varphi_t} \tag{14}$$

For each trajectory φ, the equality in (14) can be realized by appropriately parameterizing φ. Specifically, we use a parameter $t^* \in \mathbb{R}$ such that $t = f(t^*)$ and

$$\|\dot{\varphi}^*_{t^*}\|_{\varphi^*_{t^*}} = \|\nabla L(\dot{\varphi}^*_{t^*})\|_{\varphi^*_{t^*}}, \tag{15}$$

where $\varphi^*_{t^*} := \varphi(f(t^*))$.

This parameterizing reduces the quasi-potential to the minimization of the following quantity,

$$(S_{0T}(\varphi) \geq S_{0f(T)}(\varphi^*) =) \int_0^{f(T)} \|\dot{\varphi}^*_{t^*}\|_{\varphi^*_{t^*}} \|\nabla L(\varphi^*_{t^*})\|_{\varphi^*_{t^*}} + \langle \dot{\varphi}^*_{t^*}, \nabla L(\varphi^*_{t^*})\rangle_{\varphi^*_{t^*}} \, dt^* \tag{16}$$

subject to

$$\|\dot{\varphi}^*_{t^*}\|_{\varphi^*_{t^*}} = \|\nabla L(\dot{\varphi}^*_{t^*})\|_{\varphi^*_{t^*}}. \tag{17}$$

Since the integrand of (16) includes the first order derivative regarding t^*, (16) is invariant over different parameterizing. For convenience, we use a new parameter r such that $t^* = g(r)$ and

$$\left\|\dot{\hat{\varphi}}_r\right\|_{\hat{\varphi}_r} = 1, \tag{18}$$

where $\hat{\varphi}_r := \varphi_g^*(r)$ and the quasi-potential is reduced to the following formula, [2]

$$V(\theta) = \inf_{\forall r \in [0,R]: \|\dot{\hat{\varphi}}_r\|_{\hat{\varphi}_r}=1} \int_0^R \left(\left\| \dot{\hat{\varphi}}_r \right\|_{\hat{\varphi}_r} \|\nabla L(\hat{\varphi}_r)\|_{\hat{\varphi}_r} + \left\langle \dot{\hat{\varphi}}_r, \nabla L(\hat{\varphi}_r) \right\rangle_{\hat{\varphi}_r} \right) dr, \tag{19}$$

where $\hat{\varphi}_R = x$. The right hand side can be expanded further by Bellman optimality,

$$V(\theta) = \inf_{\forall r \in [0,R]: \|\dot{\hat{\varphi}}_r\|_{\hat{\varphi}_r}=1} \left(\int_{R-\delta}^R \left(\left\| \dot{\hat{\varphi}}_r \right\|_{\hat{\varphi}_r} \|\nabla L(\hat{\varphi}_r)\|_{\hat{\varphi}_r} + \left\langle \dot{\hat{\varphi}}_r, \nabla L(\hat{\varphi}_r) \right\rangle_{\hat{\varphi}_r} \right) dr + V(\hat{\varphi}_{R-\delta}) \right) \tag{20}$$

Taylor expansion around $r = R$ gives

$$\int_{R-\delta}^R \left(\left\| \dot{\hat{\varphi}}_r \right\|_{\hat{\varphi}_r} \|\nabla L(\hat{\varphi}_r)\|_{\hat{\varphi}_r} + \left\langle \dot{\hat{\varphi}}_r, \nabla L(\hat{\varphi}_r) \right\rangle_{\hat{\varphi}_r} \right) dr + V(\hat{\varphi}_{R-\delta}) \tag{21}$$

$$= \delta \left(\|\nabla L(\hat{\varphi}_R)\|_{\hat{\varphi}_R} + \left\langle \nabla L(\hat{\varphi}_R), \dot{\hat{\varphi}}_R \right\rangle_{\hat{\varphi}_R} - \dot{\hat{\varphi}}_R^\top \nabla V(\hat{\varphi}_R) \right) + V(\hat{\varphi}_R) + o\left(\delta^2\right) \tag{22}$$

Taking $\delta \to 0$ and noticing $\hat{\varphi}_R = \theta$, (20) can be simplified to the following equation.

$$0 = \inf_{\forall r \in [0,R]: \|\dot{\hat{\varphi}}_r\|_{\hat{\varphi}_r}=1} \left(\|\nabla L(\theta)\|_\theta + \left\langle \nabla L(x)^\top, \dot{\hat{\varphi}}_R \right\rangle_\theta - \dot{\hat{\varphi}}_R \nabla V(\theta) \right) \tag{23}$$

It remains to choose $\hat{\varphi}$ such that $\dot{\hat{\varphi}}_R$ minimizes the following quantity,

$$\|\nabla L(\theta)\|_\theta + \left\langle \nabla L(x)^\top, \dot{\hat{\varphi}}_R \right\rangle_\theta - \dot{\hat{\varphi}}_R^\top \nabla V(\theta). \tag{24}$$

Since the following equality holds,

$$\left\langle \nabla L(\theta)^\top, \dot{\hat{\varphi}}_R \right\rangle_\theta - \dot{\hat{\varphi}}_R \nabla V(\theta) = \left\langle \nabla L(\theta)^\top - \nabla V(\theta)^\top C(\theta)^{-1/2}, \dot{\hat{\varphi}}_R \right\rangle_\theta \tag{25}$$

it is easy to see that a trajectory $\hat{\varphi}^*$ such that

$$\dot{\hat{\varphi}}_R^* = -\frac{\nabla L(x)^\top - \nabla V(\theta) C(\theta)^{1/2}}{\|\nabla L(\theta)^\top - \nabla V(\theta) C(\theta)^{1/2}\|_\theta} \tag{26}$$

minimizes (25) and, thus minimizes (24).

With this $\hat{\varphi}^*$, (23) simplifies to

$$\|\nabla L(\theta)\|_\theta = \|\nabla L(\theta)^\top - \nabla V(\theta) C(\theta)^{1/2}\|_\theta. \tag{27}$$

Taking the square of both sides, we get

$$\frac{1}{2} \nabla V(\theta)^\top C(\theta)^{1/2} \nabla V(\theta) - \nabla L(\theta) \cdot \nabla V(\theta) = 0 \tag{28}$$

□

B Proof for Theorem 5

Theorem 5. *If ε is sufficiently small,*

$$\mathbb{E}[\tau] = \exp\left[\varepsilon^{-2} V_0\right]$$

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

[2] One might think that if we parametrize as above (18), the equality condition for (14) is violated. Indeed

$$S_{0T}(\hat{\varphi}) \neq \int_0^T \|\dot{\varphi}_t\|_{\hat{\varphi}_t} \|\nabla L(\hat{\varphi}_t)\|_{\hat{\varphi}_t} + \langle \dot{\varphi}_t, \nabla L(\hat{\varphi}_t) \rangle_{\hat{\varphi}_t} dt$$

for $\hat{\varphi}$. However, $\hat{\varphi}$ is introduced just for the simple calculation of $S_{0T}(\varphi^*)$. Although $\hat{\varphi}$ frequently appears in the proof, our attention is still on φ^* and $S_{0T}(\varphi^*)$, not on $S_{0T}(\hat{\varphi})$.

Here we provide the proof for upper-bound (Lemma 5) and lower-bound (Lemma 6). Throughout the proofs, we use ε to denote $\sqrt{\frac{\eta}{B}}$ for simplicity and use $\tilde{\theta}_t$ to denote gradient flow, i.e., $\frac{d\tilde{\theta}_t}{dt} = -\nabla L(\tilde{\theta}_t)$ to clearly distinguish from the system θ_t in (3). Also, we use $\mathbf{C}_{0T,\theta_0}$, P_{θ_0}, instead of $\mathbf{C}_{0T}$ or P, to clearly indicate the initial values trajectories.

Lemma 5. *If ε is sufficiently small,*

$$\mathbb{E}\left[\tau\right] = O\Big(\exp\left[\varepsilon^{-2}V_0\right]\Big)$$

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

Proof. We show for any positive constant $c > 0$, there exists a small ε_0 such that $\forall \varepsilon \leq \varepsilon_0$: $\mathbb{E}\left[\tau\right] \leq \exp\left[\varepsilon^{-2}(V_0 + c)\right]$.

We show the following two independent facts:

1. The time for (3) to start from $\theta_0 = \theta_0^1 \in D$ and reach inside of Γ, τ_1, is no more than T_1 with high probability
2. The time for (3) to start from $\theta_0 = \theta_0^2 \in \Gamma$ and reach ∂D, τ_2, is no more than T_2 with high probability.

Part 1:

Given the trajectory φ_1 provided by Lemma 1 and Lemma 9, we know if $\varepsilon < \varepsilon_{\mathrm{stp1}}(\varphi_1, \mu/2, 1)$, the following inequality holds

$$\mathrm{P}_{\theta_0^1}\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0^1}(\mathbb{R}^d) \mid \rho\left(\varphi', \varphi_1\right) < \mu/2\Big\} \geq \exp\left\{-\varepsilon^{-2}\right\}. \tag{29}$$

Therefore, if we take $\varepsilon < \min\{\sqrt{\frac{1}{\ln 2}}, \varepsilon_{\mathrm{stp1}}(\varphi_1, \mu/2, 1)\}$, we have

$$\mathrm{P}_{\theta_0^1}\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0^1}(\mathbb{R}^d) \mid \rho\left(\varphi', \varphi_1\right) < \mu/2\Big\} \geq \frac{1}{2} \tag{30}$$

Since the event of $\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0^1}(\mathbb{R}^d) \mid \rho\left(\varphi', \varphi_1\right) < \mu/2\Big\}$ means that x_t reaches Γ in no later than T_1, we obtain

$$\mathrm{P}_{\theta_0^1}\left\{\tau_1 < T_1\right\} \geq \frac{1}{2}. \tag{31}$$

This finishes **Part 1**.

Part 2:

For clarity, we introduce θ_t^2 to denote an independent copy of (3) whose initial value is θ_0^2, i.e.

$$\dot{\theta}_t^2 = -\nabla L\left(\theta_t^2\right) + \varepsilon D(\theta_t^2)\dot{w}_t, \tag{32}$$

Given the trajectory φ_2 provided by Lemma 9 tells us if $\varepsilon < \varepsilon_{\mathrm{stp1}}(\varphi_2, g, \frac{c}{2})$, the following inequality holds

$$\mathrm{P}_{\theta^2}\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0^2}(\mathbb{R}^d) \mid \rho\left(\varphi', \varphi_2\right) < g\Big\} \geq \exp\Big\{-\varepsilon^{-2}\Big(S_{0T}\left(\varphi_2\right) + \frac{c}{2}\Big)\Big\}. \tag{33}$$

Since $\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0^2}(\mathbb{R}^d) \mid \rho\left(\varphi', \varphi_2\right) < g\Big\}$ is the event that y_t goes out of B in no more than the time T_2 and we know $S_{0T} < (\varphi_2) < V_0 + \frac{\varepsilon}{2}$, we can conclude the following,

$$\mathrm{P}_{\theta^2}\left\{\tau_2 < T_2\right\} \geq \exp\left\{-\varepsilon^{-2}\left(V_0 + c\right)\right\}. \tag{34}$$

This finishes **Part 2**.

Combining (31) and (34), we can obtain

$$\mathrm{P}_{\theta}\left\{\tau < T_1 + T_2\right\} \geq \frac{1}{2}\exp\left\{-\varepsilon^{-2}\left(V_0 + c\right)\right\} \tag{35}$$

Since this is a simple exponential distribution, we can obtain the following expectation.

$$\mathbb{E}\left[\tau\right] \leq 2\left(T_1 + T_2\right) \exp\left\{\varepsilon^{-2}\left(V_0 + c\right)\right\} \tag{36}$$

By setting

$$\varepsilon < \frac{1}{\sqrt{\ln 2(T_1 + T_2)}} \min\left\{\sqrt{\frac{1}{\ln 2}}, \varepsilon_{\text{stp1}}(\varphi_1, \mu/2, 1), \varepsilon_{\text{stp1}}(\varphi_2, g, \frac{c}{2})\right\}, \tag{37}$$

we can get

$$\mathbb{E}\left[\tau\right] \leq \exp\left\{\varepsilon^{-2}\left(V_0 + c\right)\right\}. \tag{38}$$

□

Lemma 6. *If ε is sufficiently small,*

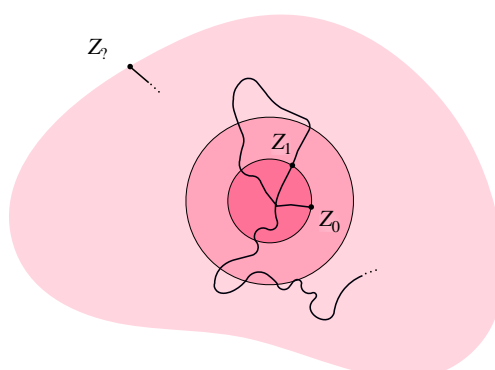
$$\mathbb{E}\left[\tau\right] = \Omega\Big(\exp\left[\varepsilon^{-2} V_0\right]\Big)$$

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

Proof. We show for any positive constant $c > 0$, there exists a small ε_0 such that $\forall \varepsilon \leq \varepsilon_0$: $\mathbb{E}\left[\tau\right] \geq \exp\left[\varepsilon^{-2}(V_0 - c)\right]$.

We consider the specific cases where the initial value of (3) is in $\partial\gamma$, which can be trivially extended to general cases. Consider a Markov chain Z_k defined as follows,

1. $\tau_0 = 0$
2. $\sigma_k = \inf\left\{t > \tau_k \ : \ \theta_t \in \partial\Gamma\right\}$
3. $\tau_k = \inf\left\{t > \sigma_{k-1} : \theta_t \in \partial\gamma \cup \partial D\right\}$
4. $Z_k = \theta_{\tau_k}$



Let $\kappa := \inf\{k \mid Z_k \in \partial D\}$, we have $\tau = \tau_n$ and

$$\mathbb{E}\left[\tau\right] = \sum_{k=0}^{\infty}\Big(\mathrm{P}_{\theta_0}\left\{\kappa \geq k\right\} - \mathrm{P}_{\theta_0}\left\{\kappa \geq k+1\right\}\Big)\tau_k \tag{39}$$

$$= \sum_{k=1}^{\infty}\mathrm{P}_{\theta_0}\left\{\kappa \geq n\right\}\left(\tau_k - \tau_{k-1}\right) \tag{40}$$

This can be further evaluated as follows,

$$\mathbb{E}\left[\tau\right] > \sum_{k=1}^{\infty}\mathrm{P}_{\theta_0}\left\{\kappa \geq k\right\}\left(\tau_k - \sigma_{k-1}\right) \quad (\because \tau_{k-1} < \sigma_{k-1}) \tag{41}$$

$$> \sum_{k=1}^{\infty}\mathrm{P}_{\theta_0}\left\{\kappa \geq n\right\}\Big(\inf_{\theta_0 \in \Gamma}\mathbb{E}[\tau_1]\Big) \tag{42}$$

Since γ is a strict subset of Γ and Γ is a strict subset of D, it takes a positive amount of time to transit from Γ to either γ or D and there exists a positive lower bound t_1 for $\inf_{\theta_0 \in \Gamma} \mathbb{E}[\tau_1]$ that is independent of ε. We get

$$\mathbb{E}[\tau] > t_1 \sum_{k=1}^{\infty} \mathrm{P}_{\theta_0}\{\kappa \geq k\}. \tag{43}$$

The immediate implication of Lemma 7 is that $\mathrm{P}_{\theta_0}\{\kappa > k\} \geq \left[1 - \exp\left\{-\varepsilon^{-2}(V_0 - d)\right\}\right]^{k-1}$. So we get

$$\mathbb{E}[\tau] > t_1 \sum_{k=1}^{\infty} \left[1 - \exp\left\{-\varepsilon^{-2}(V_0 - c)\right\}\right]^{k-1} \tag{44}$$

$$= t_1 \exp\left\{\varepsilon^{-2}(V_0 - c)\right\} \tag{45}$$

This implies $\mathbb{E}[\tau] \geq \exp\left\{\varepsilon^{-2}(V_0 - c)\right\}$ holds if ε is small enough. □

Lemma 7.

$$\mathrm{P}(Z_{k+1} \in \partial D \mid Z_k \in \gamma) \leq \exp\left\{-\varepsilon^{-2}(V_0 - c)\right\}$$

Proof. First of all, decompose $\mathrm{P}(Z_{k+1} \in \partial D \mid Z_k \in \gamma)$ into two parts in the following way.

$$\mathrm{P}_{\theta_0}(Z_{k+1} \in \partial D \mid Z_k \in \gamma) \leq \max_{\theta_0' \in \gamma} P_{\theta_0'}\{\tau_1 = \tau\} \tag{46}$$

$$= \max_{\theta_0' \in \gamma} \left[\mathrm{P}_{\theta_0'}\{\tau = \tau_1 < T\} + \mathrm{P}_{\theta_0'}\{\tau = \tau_1 \geq T\}\right] \tag{47}$$

This is true for arbitrary T so we pick $T = T'$ large enough so that this inequality holds for the first term.

$$\mathrm{P}_{\theta_0'}\{\tau = \tau_1 \geq T'\} \leq \frac{1}{2} \exp\left\{-\varepsilon^{-2}(V_0 - c)\right\} \tag{48}$$

The existence of such T' is guaranteed by the fact that V_0 is finite and the following lemma.

Lemma 8 (Freidlin & Wentzell (2012) Chapter 4 Lemma 2.2 (b))**.** *For any $\alpha > 0$, there exist positive constants c and T_0, such that for all sufficiently small $\varepsilon > 0$ and any $\theta_0 \in D \cup \partial D \backslash \mathcal{E}_\alpha(\theta^*)$ we have the inequality*

$$\mathrm{P}_{\theta_0}\{\zeta_\alpha > T\} \leq \exp\left\{-\varepsilon^{-2} c (T - T_0)\right\}$$

where $\zeta_\alpha = \inf\{t : X_t^\varepsilon \notin D \backslash \mathcal{E}_\alpha(\theta^*)\}$

Given a constant T', we consider bounding $\mathrm{P}_{\theta_0'}\{\tau = \tau_1 < T'\}\Big]$.

Consider the following set of trajectories,

$$\Phi(V_0 - c/2) = \{\varphi : \varphi_0 = \theta_0', S_{0T}(\varphi) \leq V_0 - c/2\}. \tag{49}$$

Since it takes at least V_0 to reach ∂D from θ^*, the following inequality holds,

$$\mathrm{P}_{\theta_0'}\{\tau = \tau_1 < T'\} \leq \mathrm{P}_{\theta_0'}\{\varphi_y \notin \Phi(V_0 - c/2)\}. \tag{50}$$

Also, Lemma 2 implies, for all $\varepsilon \leq \varepsilon_{\mathrm{stp2}}(V_0 - c/2, \delta, c/2)$

$$\mathrm{P}_{\varphi'}\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0'}(\mathbb{R}^d) \mid \rho\big((\varphi' - \Phi(V_0 - c/2)\big) \geq \delta\Big\} \leq \exp\{-\varepsilon^{-2}((V_0 - c/2) - c/2)\} \tag{51}$$

$$= \exp\{-\varepsilon^{-2}(V_0 - c)\} \tag{52}$$

Since δ can be arbitrarily small, the event of $\{\varphi_y \notin \Phi(V_0 - c/2)\}$ is equal to the event of $\Big\{\varphi' \in \mathbf{C}_{0T,\theta_0'}(\mathbb{R}^d) \mid \rho\big((\varphi' - \Phi(V_0 - c/2)\big) \geq \delta\Big\}$. So we can claim

$$\mathrm{P}_{\theta_0'}\{\tau = \tau_1 < T\} < \exp\{-\varepsilon^{-2}(V_0 - c)\}. \tag{53}$$

If we set $\varepsilon \leq \frac{1}{\sqrt{\ln 2}} \varepsilon_{\mathrm{stp2}}(V_0 - c/2, \delta, c/2)$, we can conclude

$$\mathrm{P}_{\theta_0'}\{\tau = \tau_1 < T\} < \frac{1}{2} \exp\{-\varepsilon^{-2}(V_0 - c)\}. \tag{54}$$

□

Lemma 9. *There exist positive constant numbers $\mu, g, T_1, T_2 > 0$ such that,*

1. *$\forall \theta \in D$, there exists a trajectory φ^1 such that $\varphi^1_0 = \theta_0$, $\varphi^1_T \in \gamma$, $0 < T \leq T_1$ and $S_{0T}(\varphi^1) = 0$*

2. *$\forall \theta \in \partial\Gamma$, there exists a trajectory φ^2 such that $\varphi^2_0 = \theta_0$, $\varphi^2_T \in \partial D_g$, $0 < T \leq T_2$ and $S_{0T}(\varphi^2) < V_0 + \frac{c}{2}$*

where γ, Γ, and D_g are $\mu/2$-neighborhood of θ^*, μ-neighborhood of θ^* and g-neighborhood of D respectively.

Proof. The first statement is immediately implied by the fact that D is attracted to a asymptotically stable equilibrium position θ^* (Assumption 4 and 5).

For the second statement, since $V_0 := \min_{x' \in \partial D} V(x')$, there exists a trajectory φ^a such that $\varphi^a_0 = \theta^*$, $\varphi^a_{T_a} \in \partial D$ and $S_{0T_a}(\varphi^a) = V_0$, where T_a is finite by Lemma 2.2 (a) in Freidlin & Wentzell (2012) Chapter 4. We cut off the first portion of φ^a up to the first intersection of Γ and define it as φ^b. $\varphi^b_0 \in \Gamma$, $\varphi^b_{T_b} \in \partial D$ and $S_{0T_b}\left(\varphi^b\right) < V_0$. By Lemma 2.3 in Freidlin & Wentzell (2012) Chapter 4, there exists a trajectory from $\varphi^b_{T_b}$ to y such that the rate function is less than $L|y - \varphi^b_{T_b}|$ with a constant L. Then, if we take a small enough g, we can obtain φ^c such $\varphi^c_0 = \varphi^b_{T_b}$, $\varphi^c_{T_c} \in \partial D_g$ and $S_{0T_c}(\varphi^c) < \frac{c}{2}$. Connecting φ^b and φ^c, we can obtain φ^2. □

C Proofs for Lemma 3 and Lemma 4

Proof for Lemma 3

Proof. Let a_i, b_i and c_i be the unitary eigenvectors of A, B and $A + B$, respectively. Namely, $Aa_i = \lambda_i(A)a_i$ and so on.

Let V_a, V_b and V_c are vector spaces spanned by the vectors $\{a_i, \ldots, a_n\}$ and $\{b_{j-i+1}, \ldots, b_n\}$ and $\{c_1, \ldots, c_j\}$, respectively. Then, $V_a \cap V_b \cap V_c \neq 0$. In fact, for arbitrary vector spaces X, Y, we can say $\dim(X \cap Y) = \dim(X) + \dim(X) - \dim(X + Y)$. Using this twice, it follows that

$$
\begin{aligned}
&\dim\left(V_a \cap V_b \cap V_c\right) \\
&= \dim\left(V_a \cap (V_b \cap V_c)\right) \\
&= \dim\left(V_a\right) + \dim\left(V_b \cap V_c\right) - \dim\left(V_a + (V_b \cap V_c)\right) \\
&= \dim\left(V_a\right) + \dim\left(V_b\right) + \dim\left(V_c\right) - \dim\left(V_b + V_c\right) - \dim\left(V_a + (V_b \cap V_c)\right) \\
&\geq \dim\left(V_a\right) + \dim\left(V_b\right) + \dim\left(V_c\right) - n - n \\
&= (n - i + 1) + (n - j + i - 1) + 1 + j - n - n \\
&= 1
\end{aligned}
$$

and thus there is a nonzero unit vector $x \in V_a \cap V_b \cap V_c$. Because $x \in V_c$, we can write $x = \sum_{v=1}^{j} x^v c_v$ and then

$$
\begin{aligned}
\langle x, (AB)x\rangle &= \langle \sum_{v=1}^{j} x^{\nu} c_v, (AB) \sum_{v=1}^{j} x^{\nu} c_v\rangle \\
&= \langle \sum_{v=1}^{j} x^{\nu} c_v, \sum_{v=1}^{j} x^{\nu} \lambda_v(AB) c_v\rangle \\
&= \sum_{v=1}^{j} |x^v|^2 \lambda_v(AB) \\
&\geq \sum_{v=1}^{j} |x^v|^2 \lambda_j(AB) \\
&= \lambda_j(AB)
\end{aligned}
$$

Similarly, because $x \in V_a$, we can write $x = \sum_{v=i}^{n} x^v a_v$ and then

$$
\begin{aligned}
\langle x, Ax\rangle &= \langle \sum_{v=i}^{n} x^{\nu} a_v, A \sum_{v=i}^{n} x^{\nu} a_v\rangle \\
&= \langle \sum_{v=i}^{n} x^{\nu} a_v, A \sum_{v=i}^{n} x^{\nu} a_v\rangle \\
&= \sum_{v=i}^{n} |x^v|^2 \lambda_v(A) \\
&\leq \sum_{v=i}^{j} |x^{\nu}|^2 \lambda_i(A) \\
&= \lambda_i(A)
\end{aligned}
$$

Similarly, we get

$$\langle x, Bx\rangle \leq \lambda_{j-i+1}(B)$$

Combining the inequalities above and Cauchy Schwartz inequality, we get

$$\lambda_j(AB) \leq \langle x, (AB)x\rangle \leq |\langle A^* x, Bx\rangle| \leq |A^* x|\,|Bx| \leq \lambda_i(A)\lambda_{j-i+1}(B)$$

□

PROOF FOR LEMMA 4

Proof. For positive definite matrices, $\lambda_{\max}(\cdot)$ is equivalent to the spectral norm of the matrix. By the sub-multiplicative property of spectral norm, we can claim

$$\lambda_{\max}(A)\lambda_{\max}(B) \geq \lambda_{\max}(AB), \tag{55}$$

Therefore, since $\lambda_{\max}(X^{-1}) = \frac{1}{\lambda_{\min}(X)}$,

$$
\begin{aligned}
& \lambda_{\max}(A^{-1})\lambda_{\max}(B^{-1}) \geq \lambda_{\max}(A^{-1}B^{-1}) && (56)\\
\iff\quad & \lambda_{\max}(A^{-1})\lambda_{\max}(B^{-1}) \geq \lambda_{\max}((AB)^{-1}) && (57)\\
\iff\quad & \frac{1}{\lambda_{\min}(A)}\frac{1}{\lambda_{\min}(B)} \geq \frac{1}{\lambda_{\min}(AB)} && (58)\\
\iff\quad & \lambda_{\min}(AB) \geq \lambda_{\min}(A)\lambda_{\min}(B) && (59)
\end{aligned}
$$

□