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

Hikaru Ibayashi
University of Southern California
ibayashi@usc.edu

Masaaki Imaizumi
The University of Tokyo
imaizumi@g.ecc.u-tokyo.ac.jp

ABSTRACT

We develop a quantitative theory on an escape problem of a stochastic gradient descent (SGD) algorithm and investigate the effect of sharpness of loss surfaces on the escape. Deep learning has achieved tremendous success in various domains, however, it has opened up various theoretical open questions. One of the typical questions is why an SGD can find parameters that generalize well over non-convex loss surfaces. An escape problem is an approach to tackle this question, which investigates how efficiently an SGD escapes from local minima. In this paper, we develop a quasi-potential theory for the escape problem, by applying a theory of stochastic dynamical systems. We show that the quasi-potential theory can handle both geometric properties of loss surfaces and a covariance structure of gradient noise in a unified manner, while they have been separately studied in previous works. Our theoretical results imply that (i) the sharpness of loss surfaces contributes to the slow escape of an SGD, and (ii) the SGD’s noise structure cancels the effect and exponentially accelerates the escape. We also conduct experiments to empirically validate our theory using neural networks trained with real data.

Keywords Deep learning · SGD · Flat minima

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 and 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 and Rivest, 1992], it is known that SGD finds nearly optimal solutions empirically, 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 and Luschi, 2018; Jastrzebski et al., 2021].

To study geometric properties of a training loss landscape is one of the promising directions to answer the question. Many empirical studies have found that 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., 2019a] 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...
Quasi-potential theory for escape problem

A reduction technique to visualize the geometry of loss surfaces, visually confirming flat minima. Most significantly, Jiang et al. [2019] conducted large-scale experiments and verified that minima in flat and wide regions have the strongest correlations with generalization capabilities. To attain a theoretical understanding of SGD, it is key to quantitatively analyze the connection between SGD and the geometric properties of loss surfaces.

An escape problem is a scheme of analyzing the dynamic of SGD escaping from local minima of loss surfaces [Zhu et al., 2019, Jastrz˛ebski 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. Jastrz˛ebski 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 a 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 and Wentzell, 2012, Dembo and 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 surfaces.

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 the gradient noise of SGD and the 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.
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 paper considers the escape problem in SGD, which is a key step in the optimization process. The escape problem deals with how quickly a learning algorithm can find a minimum of the loss function.

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, a radius of the region around a minimum, and sharpness of a minimum. As a consequence, we show that the eigenvalues of the Hesse matrix on one of the neighboring points of the minimum and \( \lambda \) is one of its eigenvalues of \( H' \). \( \Delta L \) is a difference of training loss values within a neighborhood of minimum, \( \alpha \in (0, 1) \) 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 \( \delta \), 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, a 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 Jastrzبهski 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.

| Study                      | Time to escape                   | Non-stationary | Parameter dependent noise | Discrete setup |
|----------------------------|----------------------------------|----------------|---------------------------|----------------|
| [Hu et al., 2017]          | \( \propto \exp \left[ \eta^{-1} \right] \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| Jastrzبهski et al. [2017]  | \( \exp \left[ \frac{B}{2 \eta} \Delta L \right] \sqrt{\frac{\det H'}{\det H}} \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| [Zhu et al., 2019]         | N/A                              | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| [Nguyen et al., 2019]      | \( \frac{\alpha}{2} \exp \left( \frac{s}{\eta} \right) \left( 1 + \mathcal{O}(\eta^{3/2}) \right) \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| Xie et al. [2020]          | \( \exp \left[ \frac{2B}{\eta} \Delta L \left( \frac{s}{\lambda} + \frac{1-s}{|\lambda|^2} \right) \right] \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| **Ours (continuous SGD)**  | \( \exp \left[ \frac{2B}{\eta} \Delta L \right] \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |
| **Ours (discrete SGD)**    | \( \exp \left[ \frac{2B}{\eta} \Delta L \right] + \mathcal{O}(\sqrt{\eta}) \) | \( \checkmark \) | \( \checkmark \) | \( \checkmark \) |

Table 1: Comparison of the studies on the escape problem. \( B \) is batch size, \( \eta \) 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 \( \lambda' \) is one of its eigenvalues of \( H' \). \( \Delta L \) 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 \( \delta \), and (ii) it is applicable to a wide range of the settings.

### 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 \( L_B(\theta) := \frac{1}{|B|} \sum_{x_i \in B} \ell(\theta, x_i) \), where \( B \) is a randomly sampled subset of the training data such that \( |B| = B \). 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
We provide basic assumptions for the escape problem, commonly used in the literature [Mandt et al., 2016, Zhu et al., 2019]. We remark that there are other formulations to analyze the escape problem. Zhu et al. [2019] define escaping efficiency. In particular, we focus on SGD whose noise on gradients has a Gaussian distribution. We decompose $-\nabla L(\theta_k)$ into a gradient term $-\nabla L(\theta_k)$ and a noise term $\nabla L(\theta_k) - \nabla L(\theta_k)$, and consider a case that the noise is Gaussian. With this setting, the update rule in (1) is rewritten as

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

(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 $-\nabla L^B(\theta_k)$ 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

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

(2)

where $W_k \sim N(0, \eta C(\theta_k))$ is a parameter-dependent Gaussian noise with its covariance $C(\theta) := \mathbb{E}_{i \sim \text{Uni}(1,\ldots,N)} [\nabla L(\theta) - \nabla L(\theta, x_i)]^\top (\nabla L(\theta) - \nabla L(\theta, x_i))$. 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, Jastrz˛ebski et al., 2017, He et al., 2019], 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:

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

(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 $w_{t+u} - w_t \sim N(0, u I)$ 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 $\eta$ 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 $\theta^*$ 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 $\theta_t$ 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 $\theta_k$ 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 and 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 $\theta^*$.

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

### 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, Jastrz˛ebski 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 \(H^*\).

It is known that Assumption 2 holds around a critical point \(\theta^*\) \cite{Jas2017,Zhu2019}. It is also empirically shown that Assumption 2 can approximately hold even for randomly chosen \(\theta\) (see Section 2 of \cite{Xie2020}). We further investigate a variant of Assumption 2 in Section 5.

3 Quasi-potential Theory

We introduce the basic notions of the quasi-potential theory. We start with defining a notion of steepness:

Although there are several trajectories from \(\theta\), we restrict our attention to the system (3), the same discussion is applicable to a general class of diffusion processes with Markov perturbations (For details, see section 5.7 in \cite{Dembo2010} or Section 6.5 in \cite{Freidlin2012}).

Although there are several trajectories from \(\theta^*\) 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

\[
V(\theta) := \inf_{\theta > 0} \inf_{\varphi : (\phi_0, \varphi_T) = (\theta^*, \theta)} S_{OT}(\varphi).
\]

Similar to steepness, quasi-potential can be seen as the minimum effort the system (3) needs to climb from \(\theta^*\) up to \(\theta\) on \(L(\theta)\). (For more details, see Section 5.3 of \cite{Freidlin2012}).
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 \(\theta^*\).

**Assumption 3** (\(\theta^*\) is asymptotically stable). \(\text{For any neighborhood } U \text{ that contains } \theta^*, \text{there exists a small neighborhood } V \text{ of } \theta^* \text{ such that gradient flow with any initial value } \theta_0 \in V \text{ does not leave } U \text{ for } t \geq 0 \text{ and } \lim_{t \to \infty} \theta_t = \theta^*.\)

**Assumption 4** (\(D\) is attracted to \(\theta^*\)). \(\forall \theta_0 \in D, \text{gradient flow with initial value } \theta_0 \text{ converges to } \theta^* \text{ without leaving } D \text{ as } t \to \infty.\)

where “gradient flow” means a continuous gradient descent defined as \(\dot{\theta} = -\nabla L(\theta)\).

**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 [3] is known to be equivalent to the local minimality of \(\theta^*\) under the condition that \(L(\theta)\) is real analytic around \(\theta^*\) \([Absil and Kurdyka, 2006]\). Also, by definition of asymptotic stability in Assumption [3], we can always find a region \(D\) that satisfies Assumption [4]. The more detailed properties of stability can be found, such as in Section 6.5 of \([Teschl, 2000]\).

Assumption [3] and [4] 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 5.** \(L(\theta^*) = 0\)

Assumption [5] 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, \(V = \frac{1}{2} \nabla V(\theta)^T C(\theta)^{1/2} \nabla V(\theta) - \nabla L(\theta)^T \nabla V(\theta) = 0.\) (4)

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

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

(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 \([Freidlin and Wentzell, 2012]\)), 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 \(\mathbb{E}[\tau_{SGD}]\) be the mean exit time of the continuous SGD, and let \(\mathbb{E}[\tau_{isoSGD}]\) be the mean exit time of an isotropic continuous SGD whose \(C(\theta)\) is set to \(I\).

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

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

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. \(\lambda\). 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 \(\Delta L\). 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 \(\exp [\lambda^2]\).

Our theory can be extended to the discrete case. By \(\mathbb{E}[\nu_{SGD}]\), we denote the discrete mean exit time of the discrete SGD, and by \(\mathbb{E}[\nu_{isoSGD}]\) 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 \(\text{Gobet and Menozzi}\).
We investigate a variation of Assumption 2 on the covariance matrix. Although Assumption 2 is commonly used, we describe a proof of Theorem 3. We begin with the isotropic case and then investigate the non-isotropic case.

Assumption 6 (Variant of Assumption 2)

**Anisotropic case**

\[
\mathbb{E}[\nu_{\text{isoSGD}}] = \exp \left[ \frac{2B}{\eta} r^2 \lambda \right] + O\left( \sqrt{\eta} \right),
\]

which immediately prove the following theorem:

**Theorem 2** (Discrete SGD). Given, Assumptions 1, 2, 3, 4, 5, and 6 for sufficiently small \( \eta \).

\[
\mathbb{E}[\nu_{\text{isoSGD}}] = \exp \left[ \frac{2B}{\eta} r^2 \lambda \right] + O\left( \sqrt{\eta} \right),
\]

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 for escape analysis.

4.3 Proof for Theorem 1

We describe a proof of Theorem 1. We begin with the isotropic case and then investigate the non-isotropic case.

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

\[
\frac{1}{2} \nabla V(\theta) \nabla V(\theta) - L(\theta) \nabla V(\theta) = 0.
\]

We have \( \nabla V(\theta) = 2L(\theta) \) as a solution of (6). Given that \( V(\theta) = 0 \) by the definition of steepness and \( L(\theta) = 0 \) by Assumption 5, we obtain \( V(\theta) = 2L(\theta) \) for \( \theta \in D \). Therefore, we have

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

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

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

\[
\frac{1}{2} \nabla V(\theta)^T C(\theta) \nabla V(\theta) - L(\theta) \nabla V(\theta) = 0.
\]

\[
\nabla V(\theta) = 2C(\theta)^{-\frac{1}{2}} \nabla L(\theta) \nabla V(\theta)\]

is a solution of (7). Then, by Assumptions 1 and 2, \( \nabla V(\theta) \) is simply written as

\[
\nabla V(\theta) = 2H^{\ast -\frac{1}{2}} \nabla (\theta^H H^* \theta) = 2H^{\ast -\frac{1}{2}} 2H^* \theta = 4H^{\ast -\frac{1}{2}} \theta.
\]

Given that \( V(\theta^*) = 0 \), we obtain \( V(\theta) = \theta^H H^{\ast -\frac{1}{2}} \theta \) for \( \theta \in D \). Then, we rewrite \( V_0 \) as

\[
V_0 = 2r^2 \lambda_{\text{min}}(H^{\ast -\frac{1}{2}}).
\]

Combined with (5), or Theorem 5 in Appendix B, we finish the proof of Theorem 1.

5 Further Investigation on Covariance Matrix

We investigate a variation of Assumption 2 on the covariance matrix. Although Assumption 2 is commonly used in [Jastrzȩbski et al., 2017] and [Xie et al., 2020], relaxing this is important for a flexible modeling of SGD.

**Assumption 6** (Variant of Assumption 2). \( H^{\ast -1/2} \) is a positive matrix and there exist constants \( 0 < c_1 \leq c_2 < \infty \) such that for \( \theta \in D \), \( C(\theta) = H^* G \) holds with a positive invertible matrix \( G \) as \( 0 < c_1 \leq \lambda_{\text{min}}(G) \leq \lambda_{\text{max}}(G) \leq c_2 \).

Under this different setup, we obtain the following results.

**Theorem 3.** Suppose that Assumptions 4, 5, and 6 hold. Then, for sufficiently small \( \eta \),

\[
\exp \left[ \frac{2B}{\eta} r^2 \lambda \right] \leq \mathbb{E}[\tau_{\text{isoSGD}}] \leq \mathbb{E}[\tau_{\text{SGD}}] \leq \exp \left[ \frac{2B}{\eta} r^2 \lambda \right]
\]

It suggests that replacing Assumption 2 by Assumption 6 does not affect the isotropic case, but has a constant effect in the anisotropic case.
Proof for Theorem 3. The result of $\mathbb{E}[\tau_{\text{isoSGD}}]$ is obtained in the same way as Theorem 1. For the proof of $\mathbb{E}[\tau_{\text{SGD}}]$, the following lemmas are useful, whose proofs are provided in Appendix C.

Lemma 3. For positive invertible matrices $A$ and $B$, the following inequality holds,

$$\lambda_{\min}(A^{-1}B^{-1}) \leq \lambda_{\max}(A^{-1})\lambda_{\min}(B^{-1})$$

Lemma 4. For invertible matrices $A$ and $B$, the following inequality holds

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

Similar to the proof of Theorem 1, we obtain $\nabla V(\theta) = 2C(\theta)^{-\frac{1}{2}}\nabla L(\theta)$. Then, $\nabla V(\theta)$ is simply written as

$$\nabla V(\theta) = 2C(\theta)^{-\frac{1}{2}}\nabla L(\theta) = 2C(\theta)^{-\frac{1}{2}}\nabla (\theta^T H^* \theta) = 2C(\theta)^{-\frac{1}{2}}2H^*\theta = 2(H^*G)^{-\frac{1}{2}}2H^*\theta = 4G^{-\frac{1}{2}}H^*\theta = 4G^{-\frac{1}{2}}H^*\frac{1}{2}\theta.$$

The second equation follows Assumption 1 and 5, and the fourth equation follows Assumption 6. Given that $V(\theta^*) = 0$, we obtain $V(\theta) = \theta^T G^{-\frac{1}{2}}H^*\frac{1}{2}\theta$ for $\theta \in D$. Then, we rewrite $V_0$ as

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

By Lemma 3, we develop an upper bound on $\lambda_{\min}(G^{-1/2})\lambda_{\min}(H^{1/2})$ as $\lambda_{\min}(G^{-1/2})\lambda_{\min}(H^{1/2}) \leq c_1^{-1/2}\lambda^{1/2}$. Similarly, Lemma 4 gives us the following lower bound $\lambda_{\min}(G^{-1/2})\lambda_{\min}(H^{1/2}) \geq \lambda_{\min}(G^{-1/2})\lambda_{\min}(H^{1/2}) \geq c_2^{-1/2}\lambda^{1/2}$. We substitute the two inequalities into the solution (9), 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 3.

6 Experiment

![Figure 2](image)

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

We conduct an experiment to validate our result of discrete setup (Theorem 2), using a neural network and real-world datasets. We use a multi-layer perceptron and the AVILA dataset [De Stefano et al., 2011] to observe that the discrete mean exit time of SGD has exponential dependence on eigenvalue $\lambda$, radius $r$ and a ratio of the learning rate and the batch size $\eta/B$.

In order for our essential assumptions to hold, we use the mean square loss with $\ell_2$ regularizer for $L(\theta)$ (Assumption 1) and train the model with the gradient descent for a sufficiently long time to obtain $\theta_0$ near $\theta^*$. We set the $r$-neighborhood of $\theta_0$ as $D$ (Assumption 3 and 4). To measure the discrete mean exit time, we repeatedly execute a vanilla SGD from $\theta_0$ for 1000 times and take an average number of steps at which SGD exit from $D$ (i.e. when the distance from $\theta_0$ becomes farther than $r$).

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

We show the results in Figure 2. As Theorem 2 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.
7 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 and 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 Suzuki 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. [2019a] 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 and 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 and 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], Simsekli 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.

8 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 3) and the escape problem of a discrete SGD (Theorem 2). We believe our theory provides a solid insight for SGD dynamics and also flexible theory for further studies.

References

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

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

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, pages 196–201. ieeexplore.ieee.org, December 2011.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
Quasi-potential theory for escape problem

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

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

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.

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.

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

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

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

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

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.

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.

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

Zeke Xie, Isssei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Machine Learning*, pages 7654–7663. PMLR, 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, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 2017.
Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In Maria Florina Balcan and Kilian Q Weinberger, editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 354–363, New York, New York, USA, 2016. PMLR.

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, 2019b.

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, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 5827–5837. PMLR, 2019.

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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.

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.

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

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

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. arXiv preprint arXiv:1710.06451, 2017.
We develop a parameterization for the term in (11). For a trajectory
\[ \mathcal{V}(\theta) \] satisfies the following Jacobi equation,
\[ \frac{1}{2} \nabla \mathcal{V}(\theta)^\top C(\theta)^{1/2} \nabla \mathcal{V}(\theta) - \nabla L(\theta)^\top \nabla \mathcal{V}(\theta) = 0 \]

**Proof of Theorem 4** For \( u, v \in \mathbb{R}^d \), we introduce an inner product and a norm regarding a point \( \theta \in D \) as \( \langle u, v \rangle_\theta := u^\top C(\theta)^{-1/2} v \) and \( \| u \|_\theta := \sqrt{\langle u, u \rangle_\theta} \). With these definitions, the \( S_{OT}(\varphi) \) is written as follows:
\[ S_{OT}(\varphi) = \frac{1}{2} \int_0^T \| \dot{\varphi}_t + \nabla L(\varphi_t) \|_{\varphi_t}^2 \, dt. \tag{10} \]

Note that \( \varphi_t \in D \) holds for any \( t \in [0, T] \) by the definition of trajectories. We rewrite the integrand of (10) as follows:
\begin{align*}
\| \dot{\varphi}_t + \nabla L(\varphi_t) \|_{\varphi_t}^2 &= \| \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} \\
&= \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} \\
&\geq 2 \langle \dot{\varphi}_t, \nabla L(\varphi_t) \rangle_{\varphi_t}.
\end{align*} \tag{11}

We develop a parameterization for the term in (11). For a trajectory \( \varphi \), we select an bijective function \( f : [0, 1] \to [0, T] \) as satisfying the follows: for each \( t \in [0, T] \) and \( t^* \in [0, 1] \) as \( t = f(t^*) \), a parameterized trajectory \( \varphi_t = \varphi_{f(t^*)} \) satisfies
\[ \| \dot{\varphi}_{f(t^*)} \|_{\varphi_{f(t^*)}} = \| \nabla L(\varphi_{f(t^*)}) \|_{\varphi_{f(t^*)}}. \tag{12} \]

This parameterization reduces the quasi-potential to the minimum of the following quantity:
\[ (S_{OT}(\varphi) \geq S_{OT}(f(T))(\varphi^*) = \int_0^{f(T)} \| \dot{\varphi}_{f(t^*)} \|_{\varphi_{f(t^*)}} \| \nabla L(\varphi_{f(t^*)}) \|_{\varphi_{f(t^*)}} + \langle \dot{\varphi}_{f(t^*)}, \nabla L(\varphi_{f(t^*)}) \rangle_{\varphi_{f(t^*)}} dt^* \tag{13} \]
subject to the constraint (12). Since the integrand of (13) includes the first order derivative regarding \( t^* \), (13) holds over different parameterizations \( f \). For convenience, we use another bijective parameterization function \( g : [0, R] \to [0, 1] \) as \( t^* = g(r) \) with \( R > 0 \) and \( r \in [0, R] \) such that \( \dot{\varphi}_r := \varphi_{g(r)} \) satisfies
\[ \| \dot{\varphi}_r \|_{\varphi_r} = 1. \tag{14} \]
Then, the quasi-potential is reduced to the following formula,

\[
V(\theta) = \inf_{r \in [0,R]} \int_0^R \left( \|\dot{\phi}_r\|_{\phi_r} \|\nabla L(\dot{\phi}_r)\|_{\phi_r} + \langle \dot{\phi}_r, \nabla L(\dot{\phi}_r) \rangle_{\phi_r} \right) dr, \tag{15}
\]

where \(\dot{\phi}_R = \theta\). By the Bellman equation-type optimality, we expand the right hand side of (15) into the following form:

\[
V(\theta) = \inf_{r \in [0,R]} \int_{R-\delta}^R \left( \|\dot{\phi}_r\|_{\phi_r} \|\nabla L(\dot{\phi}_r)\|_{\phi_r} + \langle \dot{\phi}_r, \nabla L(\dot{\phi}_r) \rangle_{\phi_r} \right) dr + V(\dot{\phi}_{R-\delta}) \tag{16}
\]

with a width value \(\delta > 0\). The Taylor expansion around \(r = R\) gives

\[
\int_{R-\delta}^R \left( \|\dot{\phi}_r\|_{\phi_r} \|\nabla L(\dot{\phi}_r)\|_{\phi_r} + \langle \dot{\phi}_r, \nabla L(\dot{\phi}_r) \rangle_{\phi_r} \right) dr + V(\dot{\phi}_{R-\delta}) = \delta \left( \|\nabla L(\dot{\phi}_R)\|_{\phi_R} + \langle \nabla L(\dot{\phi}_R), \dot{\phi}_R \rangle_{\phi_R} - \dot{\phi}_R^T \nabla V(\dot{\phi}_R) \right) + V(\dot{\phi}_R) + O(\delta^2)
\]

Taking \(\delta \to 0\) and noticing \(\dot{\phi}_R = \theta\), (16) can be simplified to the following equation:

\[
0 = \inf_{r \in [0,R]} \left( \|\nabla L(\theta)\|_{\theta} + \langle \nabla L(x)^T, \dot{\phi}_R \rangle_{\theta} - \dot{\phi}_R \nabla V(\theta) \right) \tag{17}
\]

It remains to select \(\dot{\phi}\) which solves the minimization problem (17). Since the following equality holds,

\[
\langle \nabla L(\theta)^T, \dot{\phi}_R \rangle_{\theta} - \dot{\phi}_R \nabla V(\theta) = \langle \nabla L(\theta)^T - \nabla V(\theta)^T C(\theta)^{-1/2}, \dot{\phi}_R \rangle_{\theta}, \tag{18}
\]

it is easy to see that a trajectory \(\dot{\phi}^*\) such that

\[
\dot{\phi}^*_R = -\frac{\nabla L(x)^T - \nabla V(\theta)^T C(\theta)^{-1/2}}{\|\nabla L(\theta)^T - \nabla V(\theta)^T C(\theta)^{-1/2}\|_{\theta}}
\]

minimizes (18). With this \(\dot{\phi}^*\), (17) simplifies to

\[
\|\nabla L(\theta)\|_{\theta} = \|\nabla L(\theta)^T - \nabla V(\theta)^T C(\theta)^{-1/2}\|_{\theta}.
\]

Taking the square of both sides, we get the statement.

\[
\Box
\]

### B Theorem on Exit Time

In this section, we develop the following fundamental theorem on the notion of exit time. For simplicity, we use \(\varepsilon\) to denote \(\sqrt{\eta/B}\).

**Theorem 5.** If \(\varepsilon\) is sufficiently small,

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

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

To prove this result, we provide the proof for an upper bound (Lemma 6) and a lower bound (Lemma 7). Throughout the proofs, we use \(C_{\theta T, \theta_0}\), \(P_{\theta_0}\), instead of \(C_{\theta T}\) or \(P\), to clearly indicate which trajectory we are referring to.

We introduce several notions. For \(\delta > 0\) and \(\theta \in \mathbb{R}^d\), let \(B_\delta(\theta)\) denote an \(\delta\)-neighbourhood of \(\theta\), that is, \(B_\delta(\theta) := \{\theta' \in \mathbb{R}^d \mid \|\theta' - \theta\| \leq \delta\}\). Further, for a set \(\Theta \subset \mathbb{R}^d\), \(B_\delta(\Theta) := \cup_{\theta \in \Theta} B_\delta(\theta)\).

The following lemma provides preliminary facts for proofs.

**Lemma 5.** For any \(c > 0\), there exist \(\mu_1, \mu_2, T_1, T_2 > 0\) such that the following hold:

1. \(\forall \theta \in D\), there exists a trajectory \(\phi^1\) such that \(\phi^1_0 = \theta, \phi^1_T \in B_{\mu_1}(\theta^*)\), \(0 < T \leq T_1\) and \(S_{\theta T}(\phi^1) = 0\).
2. \(\forall \theta \in B_{\mu_2}(\theta^*)\), there exists a trajectory \(\phi^2\) such that \(\phi^2_0 = \theta, \phi^2_T \in \partial B_{\mu_2}(D)\), \(0 < T \leq T_2\) and \(S_{\theta T}(\phi^2) < V_0 + \frac{c}{2}\).
We define the terminating time of \( \theta \) as
\[
\theta = \min_{t \in \partial D} V(\theta).
\]

The illustration can be found in Fig. 3.

**Proof of Lemma 6.** The first statement immediately holds by the fact that \( D \) is attracted to an asymptotically stable equilibrium position \( \theta^* \) (Assumption 3 and 4).

For the second statement, since \( V_0 := \min_{\theta \in \partial D} V(\theta) \), there exists a trajectory \( \varphi^0 \) such that \( \varphi^0_0 = \theta^*, \varphi^0_T \in \partial D \) and \( S_{UT}(\varphi^0) = V_0 \), where \( T = \infty \) be finite by Lemma 2.2 (a) in Freidlin and Wentzell [2012]. We cut off the first portion of \( \varphi^0 \) up until the first intersecting point with \( B_{\mu_1}(\theta^*) \) and define it as \( \varphi^0_{\mu_2} \). This means \( \varphi^0_0 \in B_{\mu_1}(\theta^*), \varphi^0_T \in \partial D \) and \( S_{UT}(\varphi^0) < V_0 \) hold. By Lemma 2.3 in Freidlin and Wentzell [2012], there exists a trajectory from \( \varphi^0_{\mu_2} \) to a point \( \theta_{\mu_2} \) in \( \partial B_{\mu_2}(D) \) such that the steepness is less than \( K|\theta_{\mu_2} - \varphi^0_T| \) with a constant \( K > 0 \). Then, if we take a small enough \( \mu_2 \), we can obtain \( \varphi^{\mu_2} \) such \( \varphi^{\mu_2}_0 = \varphi^{\mu_2}_{\mu_2}, \varphi^{\mu_2}_T \in \partial B_{\mu_2}(D) \) and \( S_{UT}(\varphi^{\mu_2}) < \frac{1}{2} \). By connecting \( \varphi^0 \) and \( \varphi^{\mu_2} \), we obtain an appropriate \( \varphi^0 \).

**Lemma 6.** If \( \varepsilon > 0 \) is sufficiently small,
\[
\mathbb{E}[\tau] = O\left( \exp\left[ \varepsilon^{-2}V_0 \right] \right)
\]
holds, where \( V_0 := \min_{\theta \in \partial D} V(\theta) \).

**Proof of Lemma 6.** We show that for any constant \( c > 0 \), there exists a small \( \varepsilon_0 \) such that \( \forall \varepsilon \leq \varepsilon_0 \) such that \( \mathbb{E}[\tau] \leq \exp\left[ \varepsilon^{-2}(V_0 + c) \right] \). To the aim, we split the dynamical system \( (1) \) of our interest into the first half and the second half, \( \{\theta^1_t\}_t \) and \( \{\theta^2_t\}_t \). \( \{\theta^1_t\}_t \) starts with \( \theta^1_0 = \theta_0 \in D \) and terminates when it first reaches \( B_{\mu_1/2}(\theta^*) \). We define the terminating time of \( \{\theta^1_t\}_t \) as \( t_1 := \min\{t > 0 : \theta^1_t \in B_{\mu_1/2}(\theta^*) \} \). On the other hand, \( \{\theta^2_t\}_t \) starts with \( \theta^2_0 = \theta_{\tau_1} \in B_{\mu_1/2}(\theta^*) \) and terminates when it first reaches \( \partial D \). We define the terminating time of \( \{\theta^2_t\}_t \) as \( t_2 := \min\{t > 0 : \theta^2_t \in \partial D \} \). Clearly, the exit time \( \tau = t_1 + t_2 \).

Regarding \( t_1 \) and \( t_2 \), we show the following two independent facts with sufficiently small \( \varepsilon > 0 \).

Fact 1 : \( t_1 \) is no more than \( T_1 \) with probability at least 1/2.

Fact 2 : \( t_2 \) is no more than \( T_2 \) with probability at least \( \exp\left\{ -\varepsilon^{-2}(V_0 + c) \right\} \).

\[\text{[1]}\text{One might think that if we parametrize as above (14), the equality condition for (11) is violated. Indeed,}
\]
\[
S_{UT}(\dot{\varphi}) \neq \int_0^T \| \dot{\varphi}_t \| \dot{\varphi}_t \| \nabla L(\dot{\varphi}_t) \| \dot{\varphi}_t \ast \dot{\varphi}_t, \nabla L(\dot{\varphi}_t) \| \dot{\varphi}_t, \ dt
\]
\[
\text{for } \dot{\varphi}. \text{ However, } \dot{\varphi} \text{ is introduced just for the simple calculation of } S_{UT}(\varphi^*). \text{ Although } \dot{\varphi} \text{ frequently appears in the proof, our attention is still on } \varphi^* \text{ and } S_{UT}(\varphi^*), \text{ not on } S_{UT}(\dot{\varphi}).\]
Fact 1: Given the trajectory $\varphi_1$ provided by Lemma 5, Lemma 1 gives us that if $\varepsilon < \varepsilon_{\text{stp1}}(\varphi_1, \mu_1/2, 1)$, the following inequality holds
\[
P_{\theta_0^1}\left\{ \varphi' \in C_{0T_1, \theta_0^1}(\mathbb{R}^d) \mid \rho(\varphi', \varphi_1) < \mu_1/2 \right\} \geq \exp\left\{ -\varepsilon^{-2} \right\}.
\]
Therefore, if we take $\varepsilon < \min\{\sqrt{1/\ln 2}, \varepsilon_{\text{stp1}}(\varphi_1, \mu_1/2, 1)\}$, we have
\[
P_{\theta_0^1}\left\{ \varphi' \in C_{0T_1, \theta_0^1}(\mathbb{R}^d) \mid \rho(\varphi', \varphi_1) < \mu_1/2 \right\} \geq \frac{1}{2}.
\]
Since the event of $\{\varphi' \in C_{0T_1, \theta_0^1}(\mathbb{R}^d) \mid \rho(\varphi', \varphi_1) < \mu_1/2\}$ means that $\{\theta_1^1\}_t$ reaches $\mathcal{B}_{\mu_1}(\theta^*)$ in no later than $T_1$, we obtain the following which provides Fact 1.
\[
P_{\theta_0^1}\left\{ \tau_1 < T_1 \right\} \geq \frac{1}{2}.
\] (19)

Fact 2: Given the trajectory $\varphi_2$ provided by Lemma 5, Lemma 1 tells us that if $\varepsilon < \varepsilon_{\text{stp1}}(\varphi_2, \mu_2, c/2)$, the following inequality holds
\[
P_{\theta_0^1}\left\{ \varphi' \in C_{0T_2, \theta_0^1}(\mathbb{R}^d) \mid \rho(\varphi', \varphi_2) < \mu_2 \right\} \geq \exp\left\{ -\varepsilon^{-2} \left( S_{0T_2} (\varphi_2) + C \right) \right\}.
\]
\{\varphi' \in C_{0T_2, \theta_0^1}(\mathbb{R}^d) \mid \rho(\varphi', \varphi_2) < \mu_2\} is the event that $\{\theta_2^1\}_t$ goes out of $D$ in no more than the time $T_2$. Also, we know that $S_{0T_2} (\varphi_2) < V_0 + \frac{c}{2}$ by Lemma 5. Hence, we can conclude the following for Fact 2:
\[
P_{\theta_0^1}\left\{ \tau_2 < T_2 \right\} \geq \exp\left\{ -\varepsilon^{-2} (V_0 + c) \right\}.
\] (20)

Combining (19) and (20), we can obtain
\[
P_{\theta_0}\left\{ \tau < T_1 + T_2 \right\} \geq \frac{1}{2} \exp\left\{ -\varepsilon^{-2} (V_0 + c) \right\}.
\] (21)

Since this is a simple exponential distribution, we can obtain the following expectation:
\[
\mathbb{E}[\tau] \leq 2 (T_1 + T_2) \exp\left\{ -\varepsilon^{-2} (V_0 + c) \right\}
\]

By setting
\[
\varepsilon < \frac{1}{\sqrt{\ln 2 (T_1 + T_2)}} \min\left\{ \frac{1}{\ln 2}, \varepsilon_{\text{stp1}}(\varphi_1, \mu_1/2, 1), \varepsilon_{\text{stp1}}(\varphi_2, \mu_2, c/2) \right\},
\]
we can get
\[
\mathbb{E}[\tau] \leq \exp\left\{ -\varepsilon^{-2} (V_0 + c) \right\}.
\]

Then, we obtain the statement. \hfill \square

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

Lemma 7. If $\varepsilon > 0$ is sufficiently small,
\[
\mathbb{E}[\tau] = \Omega\left( \exp\left[ -\varepsilon^{-2} V_0 \right] \right)
\]
holds, where $V_0 := \min_{\theta' \in \partial D} V(\theta')$.

Proof of Lemma 7. We show for any positive constant $c > 0$, there exists a small $\varepsilon_0$ such that $\forall \varepsilon \leq \varepsilon_0$, $\mathbb{E}[\tau] \geq \exp\left[ -\varepsilon^{-2} (V_0 - c) \right]$ holds.

We consider a specific case where the initial value of $\varphi_1$ is in $\partial B_{\mu_1/2}(\theta^*)$, which can be trivially extended to general cases. Consider a Markov chain $Z_k$ ($k \in \mathbb{N}$) as a discretization of $\theta_t$ as $t = \tau_k$ with a $k$-th time grid $\tau_k$. It is formally defined as follows:

1. $\tau_0 = 0$,
2. $\sigma_k = \inf\{t > \tau_k : \theta_t \in \partial B_{\mu_1}(\theta^*)\},$

(15)
which follows from the definition of $\kappa$.

This implies that

$$E.$$ 

This can be further evaluated as

$$\frac{\kappa}{\kappa - 1} \in \mathcal{B}_{\mu_{1/2}}(\theta^*) \cup \partial D.$$ 

By introducing $Z_k$, we can reduce the continuous process $\{\theta_t\}_t$ to a discrete Markov chain transiting between $\partial \mathcal{B}_{\mu_{1/2}}(\theta^*)$ and $\partial D$. The illustration can be found in Fig. 4.

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

$$E[\tau] = \sum_{k=0}^{\infty} \left( P_{\theta_0} \{\kappa \geq k\} - P_{\theta_0} \{\kappa \geq k + 1\} \right) \tau_k = \sum_{k=1}^{\infty} P_{\theta_0} \{\kappa \geq n\} (\tau_k - \tau_{k-1}) / \kappa.$$ 

This can be further evaluated as

$$E[\tau] > \sum_{k=1}^{\infty} P_{\theta_0} \{\kappa \geq k\} (\tau_k - \sigma_{k-1}) > \sum_{k=1}^{\infty} P_{\theta_0} \{\kappa \geq n\} \left( \inf_{\theta_0 \in \partial \mathcal{B}_{\mu_{1/2}}(\theta^*)} E[\tau_1] \right),$$

which follows $\tau_{k-1} < \sigma_{k-1}$. Since $\mathcal{B}_{\mu_{1/2}}(\theta^*)$ is a strict subset of $\mathcal{B}_{\mu_{1}}(\theta^*)$, and $\mathcal{B}_{\mu_{1}}(\theta^*)$ is a strict subset of $D$, it takes a positive amount of time to transit from $\partial \mathcal{B}_{\mu_{1}}(\theta^*)$ to either $\partial \mathcal{B}_{\mu_{1/2}}(\theta^*)$ or $\partial D$, and there exists a positive lower bound $t_1$ for $\inf_{\theta_0 \in \mathcal{B}_{\mu_{1}}(\theta^*)} E[\tau_1]$ that is independent of $\varepsilon$. Thus we get

$$E[\tau] > t_1 \sum_{k=1}^{\infty} P_{\theta_0} \{\kappa \geq k\}.$$ 

By Lemma 8, we immediately get $P_{\theta_0} \{\kappa > k\} \geq [1 - \exp\{-\varepsilon^{-2}(V_0 - c)\}]^{k-1}$, hence we have

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

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

**Lemma 8.** We obtain

$$P(Z_{k+1} \in \partial D \mid Z_k \in \mathcal{B}_{\mu_{1/2}}(\theta^*)) \leq \exp\{-\varepsilon^{-2}(V_0 - c)\}.$$ 

**Proof of Lemma 8.** First, we decompose $P(Z_{k+1} \in \partial D \mid Z_k \in \partial \mathcal{B}_{\mu_{1/2}}(\theta^*))$ into two parts in the following way:

$$P_{\theta_0}(Z_{k+1} \in \partial D \mid Z_k \in \partial \mathcal{B}_{\mu_{1/2}}(\theta^*)) \leq \max_{\theta_0 \in \partial \mathcal{B}_{\mu_{1/2}}(\theta^*)} P_{\theta_0}(\tau_1 = \tau)$$

$$= \max_{\theta_0 \in \partial \mathcal{B}_{\mu_{1/2}}(\theta^*)} \left[ P_{\theta_0}(\tau = \tau_1 < T) + P_{\theta_0}(\tau = \tau_1 \geq T) \right]$$

(22)
This holds for arbitrary \( T \), so we pick \( T = T' \) large enough so that this inequality holds for the first term:

\[
\mathbf{P}_{\theta_0'} \{ \tau = \tau_1 \geq T' \} \leq \frac{1}{2} \exp \{-\varepsilon^{-2} (V_0 - c) \}
\]

(23)

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

Lemma 9 (Lemma 2.2 (b) in Freidlin and Wentzell [2012]). For any \( \alpha > 0 \), there exists positive constants \( c \) and \( T_0 \), such that for all sufficiently small \( \varepsilon > 0 \) and any \( \theta_0 \in D \cup \partial D \backslash B_\alpha(\theta^*) \) we have the inequality

\[
\mathbf{P}_{\theta_0} \{ \zeta_\alpha > T \} \leq \exp \{-\varepsilon^{-2}c(T - T_0) \},
\]

where \( \zeta_\alpha = \inf \{ t : \theta_t \notin D \backslash B_\alpha(\theta^*) \} \).

Given a constant \( T' \), we consider bounding \( \mathbf{P}_{\theta_0'} \{ \tau = \tau_1 < T' \} \). Consider the following set of trajectories:

\[
\Phi(V_0 - c/2) := \{ \varphi : \varphi_0 = \theta_0', S_{\alpha T}(\varphi) \leq V_0 - c/2 \}.
\]

Since it takes at least \( V_0 \) to reach \( \partial D \) from \( \theta^* \), the following inequality holds:

\[
\mathbf{P}_{\theta_0'} \{ \tau = \tau_1 < T' \} \leq \mathbf{P}_{\theta_0'} \{ \varphi_\gamma \notin \Phi(V_0 - c/2) \}.
\]

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

\[
\mathbf{P}_{\varphi'} \{ \varphi' \in \mathbf{C}_{\alpha T, \theta_0'}(\mathbb{R}^d) \mid \rho((\varphi' - \Phi(V_0 - c/2)) \geq \delta \} \leq \exp\{-\varepsilon^{-2}((V_0 - c/2) - c/2))
\]

\[
= \exp\{-\varepsilon^{-2}(V_0 - c)\}
\]

Since \( \delta \) can be arbitrarily small, the event of \( \{ \varphi_\gamma \notin \Phi(V_0 - c/2) \} \) is equal to the event of \( \{ \varphi' \in \mathbf{C}_{\alpha T, \theta_0'}(\mathbb{R}^d) \mid \rho((\varphi' - \Phi(V_0 - c/2)) \geq \delta \} \). Hence, we obtain

\[
\mathbf{P}_{\theta_0'} \{ \tau = \tau_1 < T \} < \exp\{-\varepsilon^{-2}(V_0 - c)\}.
\]

(24)

If we set \( \varepsilon \leq \frac{1}{\sqrt{\ln 2}} \varepsilon_{\text{step}}(V_0 - c/2, \delta, c/2) \), we conclude

\[
\mathbf{P}_{\theta_0'} \{ \tau = \tau_1 < T \} < \frac{1}{2} \exp\{-\varepsilon^{-2}(V_0 - c)\}.
\]

(25)

Combining (22), (23), and (25), we prove the statement.

\[\square\]

C  Proofs for Lemma 3 and Lemma 4

Proof of Lemma 3. We introduce several definitions only for this proof. Supposing that we have a vector \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \), \( x^+ \) and \( x^- \) denote vectors whose coordinates are obtained by rearranging the numbers \( x_j \) in decreasing order and in increasing order respectively, that is, \( x^+ = (x^+_1, \ldots, x^+_n) \) and \( x^- = (x^-_1, \ldots, x^-_n) \), where \( x^+_1 \geq \cdots \geq x^+_n \) and \( x^-_1 \leq \cdots \leq x^-_n \). Also, given \( x, y \in \mathbb{R}^n \), we denote \( x \prec y \), if

\[
\sum_{j=1}^n x^+_j = \sum_{j=1}^n y^+_j, \quad \text{and} \quad \sum_{j=1}^k x^+_j \leq \sum_{j=1}^k y^+_j \quad \text{for} \quad 1 \leq k \leq n.
\]

For an \( n \times n \) matrix \( A, \lambda(A) = (\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)) \). Also, we define its elementwise logarithm \( \log \lambda(A) = (\log \lambda_1(A), \log \lambda_2(A), \ldots, \log \lambda_n(A)) \).

Corollary III.4.6 in [Bhatia, 1997] claims that given two positive matrices \( A \) and \( B \), we have

\[
\log \lambda^+(A) + \log \lambda^+(B) \prec \log \lambda(A B) \prec \log \lambda^+(A) + \log \lambda^+(B)
\]

Following the definition of \( \prec \), the left part, \( \log \lambda^+(A) + \log \lambda^+(B) \prec \log \lambda(A B) \), can be equivalently restated as

\[
\sum_{j=1}^k \{ \log \lambda^+(A) + \log \lambda^+(B) \}^+_j \leq \sum_{j=1}^k \{ \log \lambda(A B) \}^+_j \quad \text{for} \quad 1 \leq k \leq n
\]

17
For the case $k = 1$, we have

\[
\{ \log \lambda^+ (A) + \log \lambda^+ (B) \}^+ \leq \{ \log \lambda (AB) \}^+
\]
\[
\iff \max_{1 \leq i \leq n} \{ \log \lambda_{n-i+1} (A) + \log \lambda_i (B) \} \leq \log \lambda_{\text{max}} (AB)
\]

Since $\max_{1 \leq i \leq n} \{ \log \lambda_{n-i+1} (A) + \log \lambda_i (B) \} \geq \log \lambda_n (A) + \log \lambda_1 (B) = \log \lambda_{\text{min}} (A) + \log \lambda_{\text{max}} (B)$ holds, we have

\[
\log \lambda_{\text{min}} (A) + \log \lambda_{\text{max}} (B) \leq \log \lambda_{\text{max}} (AB), \quad \text{or equivalently}
\]

\[
\lambda_{\text{min}} (A) \lambda_{\text{max}} (B) \leq \lambda_{\text{max}} (AB).
\]

Since $\lambda_{\text{max}} (A) = 1/\lambda_{\text{min}} (A^{-1})$, we have

\[
\frac{1}{\lambda_{\text{max}} (A^{-1})} \frac{1}{\lambda_{\text{min}} (B^{-1})} \leq \frac{1}{\lambda_{\text{min}} ((AB)^{-1})}
\]
\[
\iff \lambda_{\text{min}} ((AB)^{-1}) \leq \lambda_{\text{max}} (A^{-1}) \lambda_{\text{min}} (B^{-1}).
\]

Since commuted matrices share eigenvalues, $\lambda_{\text{min}} ((AB)^{-1}) = \lambda_{\text{min}} (B^{-1} A^{-1}) = \lambda_{\text{min}} (A^{-1} B^{-1})$, we have

\[
\lambda_{\text{min}} (A^{-1} B^{-1}) \leq \lambda_{\text{max}} (A^{-1}) \lambda_{\text{min}} (B^{-1}).
\]

\[\square\]

**Proof of Lemma** \[2\] $\lambda_{\text{max}} (A)$ is equal to the spectral norm of $A$. By the sub-multiplicative property of spectral norm, we obtain

\[
\lambda_{\text{max}} (A) \lambda_{\text{max}} (B) \geq \lambda_{\text{max}} (AB),
\]

Therefore, since $\lambda_{\text{max}} (A) = 1/\lambda_{\text{min}} (A^{-1})$,

\[
\lambda_{\text{max}} (A^{-1}) \lambda_{\text{max}} (B^{-1}) \geq \lambda_{\text{max}} (A^{-1} B^{-1})
\]
\[
\iff \lambda_{\text{max}} (A^{-1}) \lambda_{\text{max}} (B^{-1}) \geq \lambda_{\text{max}} ((AB)^{-1})
\]
\[
\iff \frac{1}{\lambda_{\text{min}} (A)} \frac{1}{\lambda_{\text{min}} (B)} \geq \frac{1}{\lambda_{\text{min}} (AB)}
\]
\[
\iff \lambda_{\text{min}} (AB) \geq \lambda_{\text{min}} (A) \lambda_{\text{min}} (B).
\]

Then, we obtain the statement. \[\square\]