Power-Law Escape Rate of SGD

Takashi Mori [1] **Liu Ziyin** [2] **Kangqiao Liu** [2] **Masahito Ueda** [1 2 3]

Abstract

Stochastic gradient descent (SGD) undergoes complicated multiplicative noise for the mean-square loss. We use this property of SGD noise to derive a stochastic differential equation (SDE) with simpler additive noise by performing a random time change. Using this formalism, we show that the log loss barrier $\Delta \log L = \log[L(\theta^s)/L(\theta^*)]$ between a local minimum θ^* and a saddle θ^s determines the escape rate of SGD from the local minimum, contrary to the previous results borrowing from physics that the linear loss barrier $\Delta L = L(\theta^s) - L(\theta^*)$ decides the escape rate. Our escape-rate formula strongly depends on the typical magnitude h^* and the number n of the outlier eigenvalues of the Hessian. This result explains an empirical fact that SGD prefers flat minima with low effective dimensions, giving an insight into implicit biases of SGD.

1. Introduction

Deep learning has achieved breakthroughs in various applications in artificial intelligence such as image classification (Krizhevsky et al., 2012; LeCun et al., 2015), speech recognition (Hinton et al., 2012), natural language processing (Collobert & Weston, 2008), and natural sciences (Iten et al., 2020; Bapst et al., 2020; Seif et al., 2021). Such unparalleled success of deep learning hinges crucially on stochastic gradient descent (SGD) and its variants as an efficient training algorithm.

Although the loss landscape is highly nonconvex, the SGD often succeeds in finding a global minimum. It has been argued that the SGD noise plays a key role in escaping from local minima (Jastrzębski et al., 2017; Wu et al., 2018;

2020; Zhu et al., 2019; Meng et al., 2020; Xie et al., 2021; Liu et al., 2021). It has also been suggested that SGD has an implicit bias that is beneficial for generalization. For example, SGD may help the network to find *flat minima*, which are considered to imply good generalization (Keskar et al., 2017; Hoffer et al., 2017; Wu et al., 2018; Pittorino et al., 2022). How and why does SGD help the network escape from bad local minima and find flat minima? These questions have been addressed in several works, and it is now recognized that the SGD noise strength and structure significantly affect the efficiency of escape from local minima. Our work follows this line of research and adds new theoretical perspectives.

In physics and chemistry, escape from a local minimum of the energy landscape due to thermal noise at temperature T is a thoroughly studied fundamental problem (Kramers, 1940; Eyring, 1935). When the energy barrier is given by ΔE, the escape rate is proportional to $e^{-\Delta E/T}$, which is known as the Arrhenius law. By analogy, in machine learning, escape from a local minimum of the loss function is considered to be determined by the loss barrier height $\Delta L = L(\theta^s) - L(\theta^*)$, where $L(\theta)$ denotes the loss function at the network parameters θ, θ^* stands for a local minimum of $L(\theta)$, and θ^s denotes a saddle point that separates θ^* from other minima. Indeed, if we assume that the SGD noise is uniform and isotropic, which is often assumed in machine-learning literature (Jastrzębski et al., 2017), the escape rate is proportional to $e^{-\Delta L/D}$, where D denotes the SGD noise strength.

In this paper, we show that inhomogeneity of the SGD noise strength in the parameter space brings about drastic modification for the mean-square loss. We show that the escape rate is determined by the *logarithmic* loss barrier height $\Delta \log L = \log L(\theta^s) - \log L(\theta^*) = \log[L(\theta^s)/L(\theta^*)]$. In other words, the escape rate is determined not by the difference but by the ratio of $L(\theta^s)$ and $L(\theta^*)$. This result means that even if the loss barrier height ΔL is the same, minima with smaller values of $L(\theta^*)$ are more stable.

Moreover, given the fact that the eigenvalue spectrum of the Hessian at a minimum consists of a bulk of almost zero eigenvalues and outliers (Sagun et al., 2017; Papyan, 2019), our escape-rate formula implies that *SGD prefers flat minima with a low effective dimension*, where the effective

dimension is defined as the number of outliers (MacKay, 1992) and flatness is measured by a typical magnitude of outlier eigenvalues (Keskar et al., 2017). The previous theories (Jastrzębski et al., 2017; Wu et al., 2018; Zhu et al., 2019; Meng et al., 2020; Xie et al., 2021; Liu et al., 2021) have also successfully explained the fact that SGD prefers flat minima, but not the preference of small effective dimensions. The logarithmic loss naturally explains the latter, and sheds light on implicit biases of SGD.

Main contributions: We obtain the following main results:

- We derive an equation for approximating the SGD noise in Eq. (6). Remarkably, the SGD noise strength in the mean-square loss is shown to be proportional to the loss function, which is experimentally verified in Section 5.2. A key ingredient in deriving Eq. (6) is the *decoupling approximation* given in Eq. (7). This is a novel approximate method introduced in our analysis, and hence we experimentally verify it in Section 5.1.

- We derive a novel stochastic differential equation (SDE) in Eq. (14) via a random time change introduced in Eq. (13). Although the original SDE (4) has a multiplicative noise, the transformed SDE (14) has a simple additive noise with the gradient of the logarithmic loss. This shows the relevance of the logarithmic loss landscape for understanding SGD.

- We derive a novel form of SGD escape rate from a local minimum in Eq. (15). Remarkably, the escape rate takes the power-law form with respect to the ratio between $L(\theta^*)$ and $L(\theta^s)$. In Section 5.3, we experimentally test the validity of this result for a linear regression and a neural network.

- We show that the escape rate of SGD crucially depends on the flatness and the effective dimension, which implies that SGD has implicit biases towards flat minima with low effective dimension. We also show in Eq. (16) that a local minimum with an effective dimension n greater than a certain critical value n_c becomes unstable.

Related works: The role of the SGD noise structure has been discussed in some previous works (Zhu et al., 2019; Xie et al., 2021; Liu et al., 2021; Meng et al., 2020; Wojtowytsch, 2021). It was pointed out that the anisotropic nature of the SGD noise is important: the SGD noise covariance matrix is aligned with the Hessian of the loss function, which is beneficial for escape from sharp minima (Zhu et al., 2019; Xie et al., 2021; Liu et al., 2021). These previous works, however, do not take the parameter dependence of the SGD noise strength into account, and consequently, escape rates derived there depend exponentially on the loss barrier height, which differs from our formula.

Compared with the anisotropy of the SGD noise, the inhomogeneity of the SGD noise strength has been less explored. In (Meng et al., 2020; Wojtowytsch, 2021), the SGD dynamics under a state-dependent noise is discussed. However, in these previous works, the connection between the noise strength and the loss function was not theoretically established, and the logarithmic loss landscape was not discussed. The instability due to large effective dimensions was also not shown. Another recent work (Pesme et al., 2021) observed that the noise is proportional to the loss for specific simple models. In our paper, such a result is derived for more generic models. Gürbüzbalaban et al. (2021) showed that SGD will converge to a heavy-tailed stationary distribution due to a multiplicative nature of the SGD noise in a simple linear regression problem. Our paper strengthens this result: we argue that such a heavy-tailed distribution generically appears for the mean-square loss.

2. Background

2.1. Setup

We consider supervised learning. Let $\mathcal{D} = \{(x^{(\mu)}, y^{(\mu)}) : \mu = 1, 2, \dots, N\}$ be the training dataset, where $x^{(\mu)} \in \mathbb{R}^d$ denotes a data vector and $y^{(\mu)} \in \mathbb{R}$ be its label. The network output for a given input x is denoted by $f(\theta, x) \in \mathbb{R}$, where $\theta \in \mathbb{R}^P$ stands for a set of trainable parameters with P being the number of trainable parameters. In this work, we focus on the mean-square loss

$$L(\theta) = \frac{1}{2N} \sum_{\mu=1}^{N} \left[f(\theta, x^{(\mu)}) - y^{(\mu)} \right]^2 =: \frac{1}{N} \sum_{\mu=1}^{N} \ell_\mu(\theta). \tag{1}$$

The training proceeds through optimization of $L(\theta)$. In most machine-learning applications, the optimization is done via SGD or its variants. In SGD, the parameter θ_{k+1} at the time step $k+1$ is determined by

$$\theta_{k+1} = \theta_k - \eta \nabla L_{B_k}(\theta_k), \quad L_{B_k}(\theta) = \frac{1}{B} \sum_{\mu \in B_k} \ell_\mu(\theta), \tag{2}$$

where $\eta > 0$ is the learning rate, $B_k \subset \{1, 2, \dots, N\}$ with $|B_k| = B$ is a mini-batch used at the kth time step, and L_{B_k} denotes the mini-batch loss. Since the training dataset $\mathcal{D}$ is randomly divided into mini-batches, the dynamics defined by Eq. (2) is stochastic. When $B = N$, the full training data samples are used for every iteration. In this case, the dynamics is deterministic and called gradient descent (GD). SGD is interpreted as GD with stochastic noise. By introducing the SGD noise $\xi_k = -[\nabla L_{B_k}(\theta_k) - \nabla L(\theta_k)]$, Eq. (2) is rewritten as

$$\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k) + \eta \xi_k. \tag{3}$$

Obviously, $\langle \xi_k \rangle = 0$, where the brackets denote the average over possible choices of mini-batches. The noise covariance

matrix is defined as $\Sigma(\theta_k) := \langle \xi_k \xi_k^{\mathrm{T}} \rangle$. The covariance structure of the SGD noise is important in analyzing the SGD dynamics, which will be discussed in Section 3.1.

2.2. Stochastic differential equation for SGD

When the parameter update for each iteration is small, which is typically the case when the learning rate η is sufficiently small, the continuous-time approximation can be used (Li et al., 2017; Smith & Le, 2018). By introducing a continuous time variable $t \in \mathbb{R}$ and regarding η as an infinitesimal time step dt, we have a SDE

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\eta \Sigma(\theta_t)} \cdot dW_t, \qquad (4)$$

where $dW_t \sim \mathcal{N}(0, I_P dt)$ with I_n being the n-by-n identity matrix, and the multiplicative noise $\sqrt{\eta \Sigma(\theta_t)} \cdot dW_t$ is interpreted as Itô since the noise ξ_k in Eq. (3) should not depend on θ_{k+1}. Throughout this work, we consider the continuous-time approximation (4) with Gaussian noise.

In machine learning, the gradient Langevin dynamics (GLD) is also considered, in which the isotropic and uniform Gaussian noise is injected into the GD as

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2D}dW_t, \qquad (5)$$

where $D > 0$ corresponds to the noise strength (it is also called the diffusion coefficient) (Sato & Nakagawa, 2014; Zhang et al., 2017b; Zhu et al., 2019). The GLD yields an escape rate proportional to $e^{-\Delta L/D}$ (Eyring, 1935; Kramers, 1940). We will see in Section 4 that the SGD noise structure, which is characterized by $\Sigma(\theta)$, drastically alters the escape rate from a local minimum.

3. Dynamics of SGD

The fluctuation of SGD and the local landscape are the two most important factors that influence the escaping behavior of learning. In Sections 3.1 and 3.2, we establish the form of the noise of SGD and show that under this type of noise, the relevant landscape for consideration should be $\log L$ instead of L. We then use these two results to derive the escape rate of SGD in Section 4.

3.1. Structure of the SGD noise covariance

The SGD noise covariance matrix $\Sigma(\theta)$ significantly affects the dynamics (Jastrzębski et al., 2017; Smith & Le, 2018; Zhu et al., 2019; Ziyin et al., 2021). In this section, under some approximations, we derive the following expression of $\Sigma(\theta)$ for the mean-square loss near a local minimum θ^*:

$$\Sigma(\theta) \approx \frac{2L(\theta)}{B}H(\theta^*), \qquad (6)$$

where $H(\theta) = \nabla^2 L(\theta)$ is the Hessian. It should be noted that $\nabla L(\theta^*) = 0$ and $H(\theta^*)$ is positive semidefinite at any

local minima θ^*. We give a derivation below and the list of the approximations and their justifications in Appendix A. In particular, the following *decoupling approximation* is a key assumption in the derivation.

Assumption 3.1 (decoupling approximation)**.** The quantities ℓ_μ and $C_f^{(\mu)}(\theta) := \nabla f(\theta, x^{(\mu)}) \nabla f(\theta, x^{(\mu)})^{\mathrm{T}}$ are uncorrelated, which implies

$$\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu C_f^{(\mu)}(\theta) = \left(\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu\right) \cdot \left(\frac{1}{N}\sum_{\mu=1}^{N} C_f^{(\mu)}\right)$$

$$= L(\theta)\frac{1}{N}\sum_{\mu=1}^{N} C_f^{(\mu)}(\theta). \qquad (7)$$

This approximation seems justified for large networks in which ∇f behaves as a random vector. In Section 5.1, we experimentally verify the decoupling approximation for the entire training dynamics.

Our formula (6) possesses two important properties. First, the noise is aligned with the Hessian, which is well known and has been pointed out in the literature (Jastrzębski et al., 2017; Zhu et al., 2019; Xie et al., 2021; Liu et al., 2021). If the loss landscape has flat directions, which correspond to the directions of the Hessian eigenvectors belonging to vanishingly small eigenvalues, the SGD noise does not arise along these directions. Consequently, the SGD dynamics is frozen along the flat directions, which effectively reduces the dimension of the parameter space explored by the SGD dynamics. This reduction plays an important role in the escape efficiency. Indeed, we will see that the escape rate crucially depends on the effective dimension of a given local minimum.

Second, the noise is proportional to the loss function, which is indeed experimentally confirmed in Section 5.2. This property has not been pointed out and not been taken into account in previous studies (Jastrzębski et al., 2017; Zhu et al., 2019; Xie et al., 2021; Liu et al., 2021) and therefore gives new insights into the SGD dynamics. Indeed, this property allows us to formulate the Langevin equation on the logarithmic loss landscape with simple additive noise as discussed in Section 3.2. This new formalism yields the power-law escape rate summarized as Theorem 4.7, and the importance of the effective dimension of local minima for their stability.

It should be noted that although Eq. (6) is derived for the mean-square loss, the proportionality between the noise strength and the loss function seems to hold for more general loss functions such as the cross-entropy loss (see Appendix D for the detail).

Derivation of Eq. (6): We start from an analytic expression

of $\Sigma(\theta)$, which reads

$$\Sigma(\theta) = \frac{1}{B}\frac{N-B}{N-1}\left(\frac{1}{N}\sum_{\mu=1}^{N}\nabla\ell_\mu\nabla\ell_\mu^{\mathrm{T}} - \nabla L\nabla L^{\mathrm{T}}\right)$$

$$\simeq \frac{1}{B}\left(\frac{1}{N}\sum_{\mu=1}^{N}\nabla\ell_\mu\nabla\ell_\mu^{\mathrm{T}} - \nabla L\nabla L^{\mathrm{T}}\right), \qquad (8)$$

where $B \ll N$ is assumed in the second equality. The derivation of Eq. (8) is found in Jastrzębski et al. (2017); Smith & Le (2018). Usually, the gradient noise variance dominates the square of the gradient noise mean, and hence the term $\nabla L\nabla L^{\mathrm{T}}$ in Eq. (8) is negligible (Shwartz-Ziv & Tishby, 2017; Zhu et al., 2019).

For the mean-square loss, we have $\nabla\ell_\mu = [f(\theta, x^{(\mu)}) - y^{(\mu)}]\nabla f(\theta, x^{(\mu)})$, and hence

$$\Sigma(\theta) \approx \frac{2}{BN}\sum_{\mu=1}^{N}\ell_\mu\nabla f(\theta, x^{(\mu)})\nabla f(\theta, x^{(\mu)})^{\mathrm{T}}$$

$$= \frac{2}{BN}\sum_{\mu=1}^{N}\ell_\mu C_f^{(\mu)}(\theta) \approx \frac{2L(\theta)}{NB}\sum_{\mu=1}^{N}C_f^{(\mu)}(\theta), \quad (9)$$

where the decoupling approximation (7) is used in the last equality. Equation (9) is directly related to the Hessian of the loss function near a (local or global) minimum. The Hessian $H(\theta) = \nabla^2 L(\theta)$ is written as

$$H(\theta) = \frac{1}{N}\sum_{\mu=1}^{N}\left\{C_f^{(\mu)}(\theta) + \left[f(\theta, x^{(\mu)}) - y^{(\mu)}\right]\nabla^2 f(\theta, x^{(\mu)})\right\}. \quad (10)$$

It is shown by Papyan (2018) that the last term of Eq. (10) does not contribute to outliers (i.e. large eigenvalues) of the Hessian. The dynamics near a local minimum is governed by outliers, and hence we can ignore this term. At $\theta = \theta^*$, we therefore obtain

$$H(\theta^*) \approx \frac{1}{N}\sum_{\mu=1}^{N}C_f^{(\mu)}(\theta^*). \qquad (11)$$

Let us assume $C_f^{(\mu)}(\theta) \approx C_f^{(\mu)}(\theta^*)$ for θ within the valley of a local minimum θ^*. We then obtain the desired expression (6) by substituting it into Eq. (9).

3.2. Logarithmized loss landscape

Let us consider the Itô SDE (4) with the SGD noise covariance (6) near a local minimum θ^*, which is written as

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\frac{2\eta L(\theta_t)}{B}H(\theta^*)}\cdot dW_t. \quad (12)$$

Let us consider a stochastic time $t(\tau)$ for $\tau \geq 0$ as

$$\tau = \int_0^{t(\tau)} dt'\, L(\theta_{t'}), \qquad (13)$$

and perform a random time change from t to τ (Øksendal, 1998). Correspondingly, we introduce the Wiener process $d\tilde{W}_\tau \sim \mathcal{N}(0, I_P d\tau)$. Since $d\tau = L(\theta_t)dt$, we have $d\tilde{W}_\tau = \sqrt{L(\theta_t)}\cdot dW_t$. In terms of the notation $\tilde{\theta}_\tau = \theta_t$, Eq. (12) is expressed as

$$d\tilde{\theta}_\tau = -\frac{1}{L(\tilde{\theta}_\tau)}\nabla L(\tilde{\theta}_\tau)d\tau + \sqrt{\frac{2\eta H(\theta^*)}{B}}d\tilde{W}_\tau$$

$$= -\left[\nabla\log L(\tilde{\theta}_\tau)\right]d\tau + \sqrt{\frac{2\eta H(\theta^*)}{B}}d\tilde{W}_\tau. \quad (14)$$

We should note that at a global minimum with $L(\theta) = 0$, which is realized in an overparameterized regime (Zhang et al., 2017a), the random time change through Eq. (13) is ill-defined since τ is frozen at a finite value once the model reaches a global minimum. We can overcome this difficulty by adding an infinitesimal constant $\epsilon > 0$ to the loss, which makes the loss function positive without changing the finite-time dynamics like the escape from a local minimum θ^* with $L(\theta^*) > 0$.

In this way, the Langevin equation on the loss landscape $L(\theta)$ with multiplicative noise is transformed to that on the logarithmic loss landscape $U(\theta) = \log L(\theta)$ with simpler additive noise. This formulation indicates the importance of considering the logarithmic loss landscape $U(\theta) = \log L(\theta)$. In the following, we use Eq. (14) to discuss the escape efficiency from local minima.

4. Escape rate from local minima

Now we present our main result on the escape from a local minimum θ^*. First, we formally define some key concepts regarding the escape problem.

Definition 4.1 (basin of attraction). For a given local minimum θ^*, its basin of attraction $\mathcal{A}_{\theta^*}$ (or a valley of θ^*) is defined as the set of all the starting points θ_0 such that $\theta_t \to \theta^*$ as $t \to \infty$ when there is no noise.

Definition 4.2 (escape rate). We denote by P^* the total probability within $\mathcal{A}_{\theta^*}$ and by $\mathcal{J}$ the total flux of the probability current across the boundary of $\mathcal{A}_{\theta^*}$.[1] The escape rate κ is then defined as $\mathcal{J}/P^*$ (Kramers, 1940).

---

[1]By defining the probability current density $J(\theta, t)$ via the continuity equation $\partial P(\theta, t)/\partial t = -\nabla J(\theta, t)$ for the probability distribution $P(\theta, t)$ of θ at time t, $\mathcal{J}$ is explicitly given by $\mathcal{J} = \int_{\mathcal{A}_{\theta^*}} d\theta\, \nabla\cdot J(\theta, t) = \int_{\partial\mathcal{A}_{\theta^*}} da\cdot J(\theta, t)$, where the second equality follows from Gauss' law. Here, $\partial\mathcal{A}_{\theta^*}$ denotes the boundary of $\mathcal{A}_{\theta^*}$ and da is a vector representing an infinitesimal element of area of the surface.

We now calculate the asymptotic behavior of the escape rate in the weak-noise limit $\eta/B \to +0$, which has been investigated in many fields, including physics (Kramers, 1940; Langer, 1969), chemistry (Eyring, 1935), stochastic processes (Bovier et al., 2004; Berglund, 2013), and machine learning (Jastrzębski et al., 2017; Xie et al., 2021). In these previous studies, the following two basic assumptions are commonly used, which are expected to hold in the weak-noise limit. Indeed, the Kramers formula (Kramers, 1940), which is obtained by using Assumptions 4.3 and 4.4 below, is rigorously justified (Bovier et al., 2004).

Assumption 4.3 (quasi-stationary approximation). *The parameters θ obey the quasi-stationary distribution near θ^*, which is identified as the stationary distribution restricted to $\mathcal{A}_{\theta^*}$ (Bianchi & Gaudillière, 2016).*

Assumption 4.4 (most probable escape path). *The escape from θ^* is dominated by the most probable escape path (MPEP) (Freidlin & Wentzell, 1998; Maier & Stein, 1993). It is known that the direction of the MPEP at one point must be the direction of one eigenvector of the Hessian at that point, and the MPEP passes a saddle point θ^s at the boundary of $\mathcal{A}_{\theta^*}$.*

Now we introduce further key assumptions in our analysis for $P > 1$.

Assumption 4.5 (Hessian outliers). *The eigenvalue spectrum of the Hessian $H(\theta^*)$ at θ^* has n nonzero eigenvalues, which are called "outliers". The remaining $P - n$ eigenvalues are vanishingly small.*

Assumption 4.6 (low-dimensional subspace of SGD). *Outlier eigenvectors of the Hessian $v_1, v_2, \ldots, v_n \in \mathbb{R}^P$ do not change within $\mathcal{A}_{\theta^*}$, and the SGD dynamics is restricted to the n-dimensional subspace spanned by those outlier eigenvectors (n is called the effective dimension of the local minimum).*

Assumption 4.5 is empirically confirmed (Sagun et al., 2017; Papyan, 2019). Since the SGD dynamics is frozen along flat directions as we pointed out in Section 3.1, the restriction to the n-dimensional outlier subspace in Assumption 4.6 is justified. Indeed, it is empirically observed that the SGD dynamics is actually restricted to a low-dimensional subspace spanned by top eigenvectors of the Hessian (Gur-Ari et al., 2018).

Now we are ready to state our main result in the form of the following theorem.

Theorem 4.7. *Let the model parameter θ_t evolve according to the SDE in Eq. (12). Under Assumptions 4.3 to 4.6, the escape rate κ asymptotically behaves as*

$$\kappa \sim \frac{\sqrt{h_e^* |h_e^s|}}{2\pi} \left[\frac{L(\theta^s)}{L(\theta^*)} \right]^{-\left(\frac{B}{\eta h_e^*} + 1 - \frac{n}{2} \right)} \quad (15)$$

as $\eta/B \to +0$[2], where θ^s is the saddle on the MPEP, h_e^* is the eigenvalue of the Hessian at θ^* along the MPEP, and h_e^s is the negative eigenvalue of the Hessian at θ^s along the MPEP.

Proof. The proof is given in Appendix B. □

From Eq. (15) we can obtain some implications. The factor $[L(\theta^s)/L(\theta^*)]^{-\left(\frac{B}{\eta h_e^*} + 1 - \frac{n}{2} \right)}$ increases with h_e^* and n, which indicates that sharp minima (i.e. minima with large h_e^*) or minima with large n are unstable. This fact explains why SGD finds flat minima with a low effective dimension n. Equation (15) also implies that the effective dimension of any stable minima must satisfy $n < 2(B/(\eta h_e^*) + 1)$. This condition guarantees the stability along the MPEP, but any minimum must be stable along all the directions. The stability condition along the direction of the ith eigenvector of the Hessian is obtained by replacing h_e^* by h_i. Therefore, the maximum eigenvalue $h_{\max}$ gives the most stringent stability condition:

$$n < n_c := 2 \left(\frac{B}{\eta h_{\max}} + 1 \right). \quad (16)$$

The instability due to a large effective dimension is a new insight naturally explained by the picture of the logarithmic loss landscape.

Let us summarize novel aspects of the escape rate derived here, which are not found in previous studies (Jastrzębski et al., 2017; Wu et al., 2018; Zhu et al., 2019; Meng et al., 2020; Xie et al., 2021; Liu et al., 2021). First, the escape rate in Eq. (15) takes a power-law form with respect to the ratio between $L(\theta^*)$ and $L(\theta^s)$, which differs from a conventional exponential form with respect to the difference $\Delta L = L(\theta^s) - L(\theta^*)$. Intuitively, escaping behavior must be a result of noise in the gradient, and if there is no noise, gradient descent cannot escape any basin of attraction. Our result agrees with this intuition: at the global minimum where $L(\theta^*) = 0$, there is no noise, and the SGD must be stuck in where it is with no escaping behavior, which is directly reflected by our formula. However, the standard exponential escape rate formula predicts a non-zero escape rate even when $L(\theta^*) = 0$, which cannot be correct in principle. Second, Eqs. (15) and (16) imply that the SGD is biased to small effective dimensions n. As we mentioned, it is empirically confirmed that the SGD dynamics is restricted to a low-dimensional subspace spanned by top eigenvectors of the Hessian (Gur-Ari et al., 2018). Since the restriction to a low-dimensional subspace greatly reduces the capacity of the network (Li et al., 2018), Eq. (16) supports the presence of an implicit regularization via the SGD.

---

[2]Here, the notation "$f(x) \sim g(x)$ as $x \to +0$" means $\lim_{x \to +0}(f(x)/g(x)) = 1$.

The formula in Eq. (15) is tested in Section 5.3. Since it is difficult in practice to identify when the barrier crossing occurs, we instead consider the mean first passage time, which imitates the escape time for a non-convex loss landscape. Let us fix a threshold value of the loss function. The first passage time t_p is defined as the shortest time at which the loss exceeds the threshold value. We identify the threshold value as $L(\theta^s)$, i.e., the loss at the saddle in the escape problem. It is expected that t_p is similar to the escape time and proportional to κ^{-1}. Indeed, at least when noise is isotropic and uniform as in the GLD (5), the inverse of the mean first passage time $1/\langle t_p \rangle$ is logarithmically equivalent to κ, i.e., $\log \kappa \sim -\log \langle t_p \rangle$, in the weak-noise asymptotic limit (Freidlin & Wentzell, 1998; Berglund, 2013).

5. Experiments

Our key theoretical observation is that the SGD noise strength is proportional to the loss function, which is obtained as a result of the decoupling approximation. This property leads us to the Langevin equation (14) with the logarithmic loss gradient and an additive noise through a random time change (13). Equation (14) implies the escape rate (15).

In Section 5.1, we show that the decoupling approximation is valid during the entire training dynamics. In Section 5.2, we measure the SGD noise strength and confirm that it is indeed proportional to the loss function near a minimum. In Section 5.3, we experimentally test the validity of Eq. (15) for the escape rate. We will see that numerical results for a linear regression and for a non-linear neural network agree with our theoretical results.

5.1. Experimental verification of the decoupling approximation

Let us compare the eigenvalue distribution of the exact matrix $(1/N)\sum_{\mu=1}^{N} \ell^{(\mu)} C_f^{(\mu)}$ with that of the decoupled one $L(\theta) \cdot (1/N)\sum_{\mu=1}^{N} C_f^{(\mu)}$ with $C_f^{(\mu)} = \nabla f(\theta, x^{(\mu)})\nabla f(\theta, x^{(\mu)})^{\mathrm{T}}$. We consider a binary classification problem using the first 10^4 samples of the MNIST dataset such that we classify each image into even (its label is $y = +1$) or odd number (its label is $y = -1$). The network has two hidden layers, each of which has 100 units and the ReLU activation, followed by the output layer of a single unit with no activation. Starting from the Glorot initialization, the training is performed via SGD with the mean-square loss, where we fix $\eta = 0.01$ and $B = 100$.

Figure 1 shows histograms of their eigenvalues at different stages of the training: (a) at initialization, (b) after 50 epochs, and (c) after 500 epochs. We see that the exact matrix and the approximate one have statistically similar eigenvalue distributions except for exponentially small eigenvalues during the training dynamics. This shows that the decoupling approximation holds during the entire training dynamics in this experiment.

5.2. Measurements of the SGD noise strength

Now we test whether the SGD noise strength is actually proportional to the loss function, which is predicted by Eq. (6), an essential theoretical result of ours. As a measure of the SGD noise strength, let us consider the norm of the noise vector ξ given by $\langle \xi^{\mathrm{T}}\xi \rangle = \mathrm{Tr}\,\Sigma \approx \mathcal{N}/B$, where $\mathcal{N} := (1/N)\sum_{\mu=1}^{N} \nabla\ell_\mu^{\mathrm{T}}\nabla\ell_\mu - \nabla L^{\mathrm{T}}\nabla L$. Here we present experimental results for two architectures and datasets. First, we consider training of the Fashion-MNIST dataset by using a fully connected network with three hidden layers, each of which has 2×10^3 units and the ReLU activation, followed by the output layer of 10 units with no activation (classification labels are given in the one-hot representation). Second, we consider training of the CIFAR-10 dataset by using a convolutional neural network. Following Keskar et al. (2017), let us denote a stack of n convolutional layers of a filters and a kernel size of $b \times c$ with the stride length of d by $n \times [a, b, c, d]$. We use the configuration: $3 \times [64, 3, 3, 1]$, $3 \times [128, 3, 3, 1]$, $3 \times [256, 3, 3, 1]$, where a MaxPool(2) is applied after each stack. To all layers, the ReLU activation is applied. Finally, an output layer consists of 10 units with no activation.

Starting from the Glorot initialization, the network is trained by SGD of the mini-batch size $B = 100$ and $\eta = 0.1$ for the mean-square loss. During the training, we measure the training loss and the noise strength $\mathcal{N}$ for every epoch. Numerical results are given in Fig. 2. We see that roughly $\mathcal{N} \propto L$ at a later stage of the training, which agrees with our theoretical prediction.

Although $\mathcal{N}$ is not proportional to L at an early stage of training, it does not mean that Eq. (9) is invalid there. Since the decoupling approximation is valid for the entire training dynamics, Eq. (9) always holds. The reason why the SGD noise strength does not decrease with the loss function in the early-stage dynamics is that $\mathcal{N} \approx 2L(\theta) \times (1/N)\sum_{\mu=1}^{N} \nabla f(\theta, x^{(\mu)})^{\mathrm{T}}\nabla f(\theta, x^{(\mu)})$, but the quantity $(1/N)\sum_{\mu=1}^{N} \nabla f(\theta, x^{(\mu)})^{\mathrm{T}}\nabla f(\theta, x^{(\mu)})$ also changes during training.

Although Eq. (9) is derived for the mean-square loss, the relation $\mathcal{N} \propto L(\theta)$ holds in more general loss functions; see Appendix D for general argument and experiments on the cross entropy loss.

5.3. Experimental test of the escape rate formula

We now experimentally verify our escape-rate formula (15). Below, we first present numerical results for a simple linear regression problem, and then for a nonlinear model, i.e., a
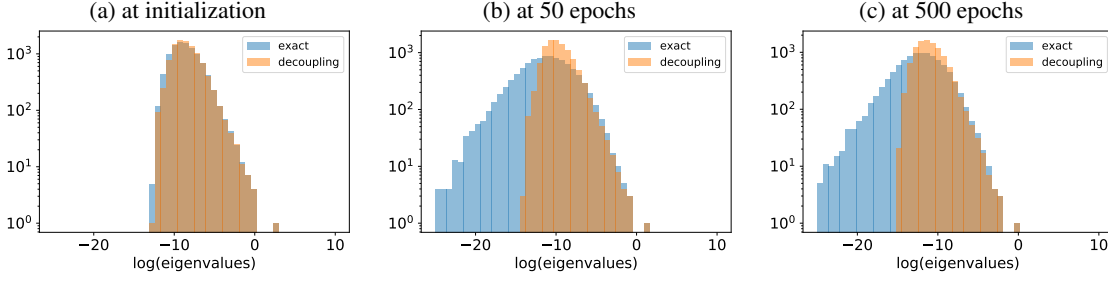
Figure 1. Comparison of the eigenvalue distributions of the left-hand side (exact expression) and the right-hand side (decoupled one) of Equation (7). They agree with each other except for exponentially small eigenvalues during the entire training dynamics. For the detail, see the description in Section 5.1.
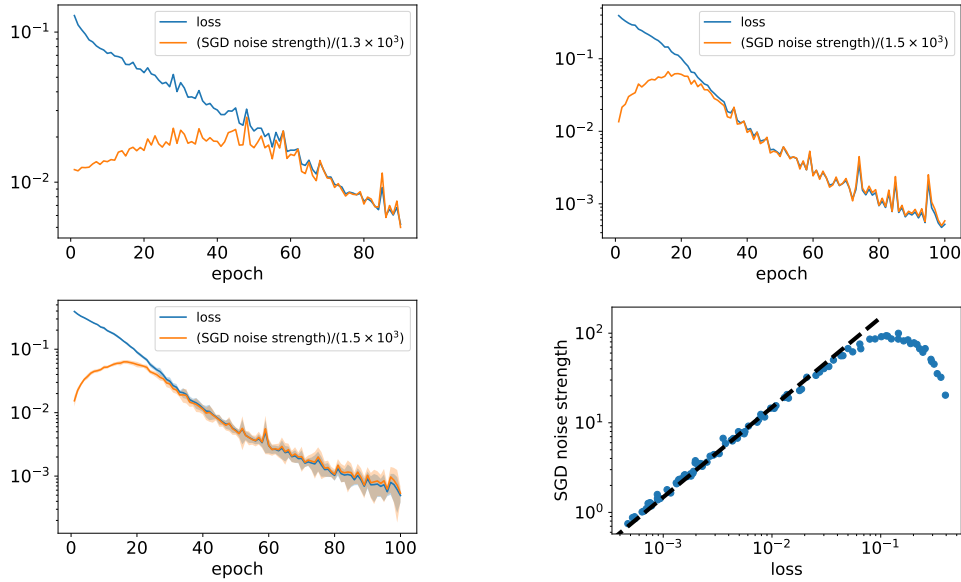

Figure 2. Training dynamics of the loss and the SGD noise strength $\mathcal{N}$. In the figure, we multiply $\mathcal{N}$ by a numerical factor to emphasize that $\mathcal{N}$ is proportional to the loss in a later stage of the training. (Top left) Fully connected network trained by the Fashion-MNIST dataset. (Top right) Convolutional network trained by the CIFAR-10 dataset. (Bottom left) Mean values with standard deviations (shaded regions) taken over 10 independent runs with different initializations in convolutional network trained by the CIFAR-10 dataset. (Bottom right) Loss vs $\mathcal{N}$ in the training of the convolutional network. The dashed line is a straight line of slope 1, which implies $\mathcal{N} \propto L(\theta)$.

neural network.

5.3.1. LINEAR REGRESSION

Let us start with the following linear regression problem: each entry of $x^{(\mu)} \in \mathbb{R}^d$ and its label $y^{(\mu)} \in \mathbb{R}$ are i.i.d. Gaussian random variables of zero mean and unit variance. We focus on the case of $d \ll N$. The output for an input x is given by $f(\theta, x) = \theta^{\mathrm{T}} x$, where $\theta \in \mathbb{R}^d$ is the trainable network parameter. We optimize θ via SGD. The mean-square loss $L(\theta) = (1/2N) \sum_{\mu=1}^{N} \left(\theta x^{(\mu)} - y^{(\mu)} \right)^2$ is quadratic and has a unique minimum at $\theta \approx 0$. the Hessian $H = (1/N) \sum_{\mu=1}^{N} x^{(\mu)} x^{(\mu)\mathrm{T}}$ has d nonzero eigenvalues, all of which are close to unity. We can therefore identify $h_e^* = h^* = 1$ and $n = d$.

Before investigating the escape rate, we remark that the stationary distribution of θ in this case is theoretically obtained because of the isotropy of SGD noise. Indeed, $H(\theta^*)$ in Eq. (14) can be replaced by $h^* I_d$, where I_d is the d-dimensional identity matrix, and then the stationary distribution of $\tilde{\theta}_\tau$ is simply given by the Gibbs distribution under the potential $U(\theta) = \log L(\theta)$: $\tilde{P}_{\mathrm{s}}(\theta) \propto e^{-U(\theta)/T} = L(\theta)^{-1/T}$, where the "temperature" T is given by $T = \eta h^*/B$. In Appendix C, it is shown that the stationary distribution $P_{\mathrm{s}}(\theta)$ of θ_t under Eq. (12) is related to $\tilde{P}_{\mathrm{s}}(\theta)$ via $P_{\mathrm{s}}(\theta) \propto L(\theta)^{-1} \tilde{P}_{\mathrm{s}}(\theta)$. We therefore have

$$P_{\mathrm{s}}(\theta) \propto L(\theta)^{-\phi}, \quad \phi = 1 + \frac{B}{\eta h^*}. \tag{17}$$

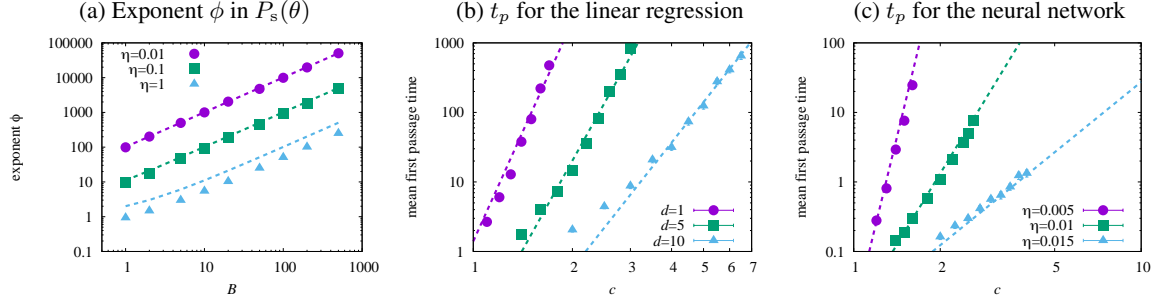Remarkably, it depends on $L(\theta)$ polynomially rather than

Figure 3. (a) Exponent ϕ for the stationary distribution $P_{\mathrm{s}}(\theta) \propto L(\theta)^{-\phi}$ for $d = 1$ in the linear regression. Dashed lines show the theoretical prediction $\phi = 1 + B/(\eta h^*)$. (b) Log-log plot of the mean first-passage time t_p vs $c = L(\theta^s)/L(\theta^*)$ for $B = 1$ and $\eta = 0.1$ in the linear regression. Dashed lines show the theoretical prediction, $t_p \propto \kappa^{-1} \propto c^{\phi-d/2}$. (c) Log-log plot of t_p vs c for $B = 10$ and various η in the neural network. Dashed lines show the theoretical prediction, $t_p \propto \kappa^{-1} \propto c^{B/(\eta h_e^*)+1-n/2}$ with $h_e^* = 95.6$ and $n = 9$.

exponentially as in the standard GLD (Jastrzębski et al., 2017; Sato & Nakagawa, 2014; Zhang et al., 2017b). This property is closely related to the power-law form of the escape rate.

First, we test Eq. (17), i.e. the stationary distribution, for $d = 1$ and $N = 10^5$. We sampled the value of θ_k at every 100 iterations ($k = j \times 100$, $j = 1, 2, \ldots, 10^4$) and made a histogram. We then fit the histogram to the form $P_{\mathrm{s}}(\theta) \propto L(\theta)^{-\phi}$ and determine the exponent ϕ. Our theory predicts $\phi = 1 + B/(\eta h^*)$. Numerical results for the exponent ϕ are presented in Fig. 3 (a) against B for three fixed learning rates η. In the same figure, theoretical values of ϕ are plotted in dashed lines. The agreement between theory and experiment is fairly well. For a large learning rate $\eta = 1$, the exponent slightly deviates from its theoretical value. This is due to the effect of a finite learning rate (recall that η is assumed to be small in deriving the continuous-time stochastic differential equation).

Next, we test our formula on the escape rate, Eq. (15). Although the mean-square loss is quadratic and no barrier crossing occurs, we can measure the first passage time, which imitates the escape time for a non-convex loss landscape as is mentioned in Section 4.

The mean first passage time over 100 independent runs is measured for varying threshold values which are specified by $c = L(\theta^s)/L(\theta^*) > 1$. Experimental results for $N = 10^4$ are presented in Fig. 3 (b). Dashed straight lines have slope $B/(\eta h^*) + 1 - n/2$. Experiments show that the first passage time behaves as $t_p \propto [L(\theta^s)/L(\theta^*)]^{B/(\eta h^*)+1-n/2}$, which agrees with our theoretical evaluation of κ^{-1} [see Eq. (15)]. We conclude that the escape rate crucially depends on the effective dimension n, which is not explained by the previous results (Zhu et al., 2019; Xie et al., 2021; Liu et al., 2021; Meng et al., 2020).

5.3.2. NEURAL NETWORK

The escape-rate formula (15) is also verified in a non-linear model. As in Section 5.1, we consider the binary classification problem using the first 10^4 samples of MNIST dataset such that we classify each image into even or odd. The network has one hidden layer with 10 units activated by ReLU, followed by the output layer of a single unit with no activation. We always use the mean-square loss. Starting from the Glorot initialization, the network is pre-trained via SGD with $\eta = 0.01$ and $B = 100$ for 10^5 iterations. We find that after pre-training, the loss becomes almost stationary around at $L(\theta) \approx 0.035$. We assume that the pre-trained network is near a local minimum. We then further train the pre-trained network via SGD with a new choice of η and B (here we fix $B = 10$), and measure the first passage time t_p. The mean first-passage time over 100 independent runs is measured for varying threshold values which are specified by $c = L(\theta^s)/L(\theta^*) > 1$. Experimental results are presented in Fig. 3 (c). We see the power-law behavior, which is consistent with our theory.

To further verify our theoretical formula (15), we also compare the power-law exponent for the mean first-passage time with our theoretical prediction $B/(\eta h_e^*) + 1 - n/2$. Here, h_e^* and n are estimated by the Hessian eigenvalues. In Appendix E, we present a numerical result for eigenvalues of the approximate Hessian given by the right-hand side of Eq. (11). By identifying the largest eigenvalue of the Hessian as h_e^*, we have $h_e^* \approx 95.6$. On the other hand, it is difficult to precisely determine the effective dimension n, but it seems reasonable to estimate $n \sim 10$. It turns out that theory and experiment agree with each other by choosing $n = 9$. Dashed lines in Fig. 3 (c) correspond to our theoretical prediction $\kappa^{-1} \propto c^{B/(\eta h_e^*)+1-n/2}$ with $h_e^* = 95.6$ and $n = 9$. This excellent agreement shows that our theoretical formula (15) is also valid in non-linear models.

5.4. Stability condition and minima selection

As we have argued in Section 4, any stable solution found by SGD should satisfy Equation (16). For a fixed value of B, Equation (16) implies the existence of the critical learning rate

$$\eta_c = \frac{2}{n-2} \frac{B}{h_{\max}}, \qquad (18)$$

above which a minimum with a certain value of $h_{\max}^*$ and n is unstable. Moreover, Equation (16) indicates that increasing η/B, i.e. making the SGD noise stronger, leads to a flatter minimum with a lower effective dimension.

We numerically test these statements in the following. We conducted binary classification of 5,000 samples of MNIST dataset (classify each image into even or odd number) by using a fully connected network with three hidden layers of width 100. The number of trainable parameters is about 10^5. First, we perform pre-training with fixed values of the learning rate η_{ini} and the batch size B_{ini}. Here we consider the two cases $(\eta_{\mathrm{ini}}, B_{\mathrm{ini}}) = (0.1, 50)$ and $(0.05, 100)$, the former of which corresponds to stronger SGD noise compared with the latter. In both cases, we find that the loss becomes almost stationary at a small value after 10^4 iterations. We then measure the largest Hessian eigenvalue $h_{\max}$ and the effective dimension n by the method in Appendix E, and determine η_c for $B = 8$. Next, we fix $B = 8$ and gradually increase the learning rate from a small value $\eta = 0.001$ and measure the loss. When the minimum becomes unstable, the loss will rapidly increase and fluctuate. Such behavior is actually observed in Figure 4. We find that our theoretical prediction of η_c (vertical dashed lines in Figure 4) is relatively close to the actual transition point. In this way, Equation (16) is relevant in learning with a neural network.

Furthermore, by comparing $h_{\max}$ and n in Figure 4 (a) and (b), we find that increasing the SGD noise results in a flatter minimum with a lower effective dimension. It shows that SGD has a bias towards flatter minima with lower effective dimensions.

6. Conclusion

In this work, we have investigated the SGD dynamics via a Langevin approach. With several approximations listed in Appendix A, we have derived Eq. (6), which shows that the SGD noise strength is proportional to the loss function. This SGD noise covariance structure yields the stochastic differential equation (14) with additive noise near a minimum via a random time change (13). The original multiplicative noise is reduced to simpler additive noise, but instead the gradient of the loss function is replaced by that of the logarithmic loss function $U(\theta) = \log L(\theta)$. This new formalism yields the power-law escape rate formula (15) whose exponent depends on η, B, h_e^*, and n.
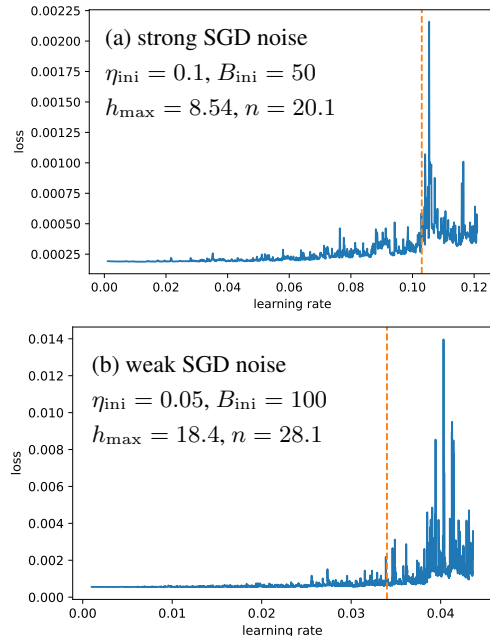

Figure 4. Stability of a minimum found by SGD with $\eta = \eta_{\mathrm{ini}}$ and $B = B_{\mathrm{ini}}$. After finding a minimum, we fix $B = 8$ and gradually increase the learning rate η. When η exceeds a threshold value, the minimum becomes unstable and the loss rapidly increases and fluctuates. Vertical dashed lines correspond to the critical learning rate η_c estimated by Equation (16).

The escape-rate formula in Eq. (15) explains an empirical fact that SGD favors flat minima with low effective dimensions. The effective dimension of a minimum must satisfy Eq. (16) for its stability. This result as well as the formulation of the SGD dynamics using the logarithmic loss landscape should help understand more deeply the SGD dynamics and its implicit biases in machine learning problems.

Although the present work focuses on the Gaussian noise, the non-Gaussianity can also play an important role. For example, Şimşekli et al. (2019) approximated SGD as a Lévy-driven SDE, which explains why SGD finds wide minima. It would be an interesting future problem to take the non-Gaussian effect into account.

Acknowledgements

References

Bapst, V., Keck, T., Grabska-Barwińska, A., Donner, C., Cubuk, E. D., Schoenholz, S. S., Obika, A., Nelson, A. W., Back, T., Hassabis, D., and Kohli, P. Unveiling the predictive power of static structure in glassy systems. *Nature Physics*, 16(4):448–454, 2020.

Berglund, N. Kramers' Law: Validity , Derivations and Generalisations. *Markov Processes and Related Fields*, 19:459–490, 2013.

Bianchi, A. and Gaudillière, A. Metastable states, quasi-stationary distributions and soft measures. *Stochastic Processes and their Applications*, 126:1622–1680, 2016.

Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M. Metastability in reversible diffusion processes I. Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6:399–424, 2004.

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning*, 2008.

Eyring, H. The Activated Complex in Chemical Reactions. *Journal of Chemical Physics*, 3:63–71, 1935. ISSN 00219606.

Freidlin, M. I. and Wentzell, A. D. *Random Perturbations of Dynamical Systems*. Springer, 1998.

Gur-Ari, G., Roberts, D. A., and Dyer, E. Gradient Descent Happens in a Tiny Subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. The Heavy-Tail Phenomenon in SGD. In *International Conference on Machine Learning*, 2021.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Others. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29 (6):82–97, 2012.

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

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. Discovering Physical Concepts with Neural Networks. *Physical Review Letters*, 124(1):10508, 2020.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three Factors Influencing Minima in SGD. *arXiv:1711.04623*, 2017.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Langer, J. S. Statistical theory of the decay of metastable states. *Annals of Physics*, 54(2):258–275, 1969.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the Intrinsic Dimension of Objective Landscapes. *International Conference on Learning Representations*, 2018.

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, 2017.

Liu, K., Ziyin, L., and Ueda, M. Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent. In *International Conference on Machine Learning*, 2021.

MacKay, D. Bayesian model comparison and backprop nets. In *Advances in Neural Information Processing Systems*, 1992.

Maier, R. S. and Stein, D. L. Escape problem for irreversible systems. *Physical Review E*, 48:931–938, 1993.

Meng, Q., Gong, S., Chen, W., Ma, Z. M., and Liu, T. Y. Dynamic of Stochastic Gradient Descent with State-Dependent Noise. *arXiv:2006.13719*, 2020.

Øksendal, B. *Stochastic differential equations: an introduction with applications*. Springer, Berlin, 1998.

Papyan, V. The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size. *arXiv:1811.07062*, 2018.

Papyan, V. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In *International Conference on Machine Learning*, 2019.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. *arXiv:2106.09524*, 2021.

Pittorino, F., Ferraro, A., Perugini, G., Feinauer, C., Baldassi, C., and Zecchina, R. Deep Networks on Toroids: Removing Symmetries Reveals the Structure of Flat Regions in the Landscape Geometry. *arXiv:2202.03038*, 2022.

Sagun, L., Evci, U., Güney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv:1706.04454*, 2017.

Sato, I. and Nakagawa, H. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, 2014.

Seif, A., Hafezi, M., and Jarzynski, C. Machine learning the thermodynamic arrow of time. *Nature Physics*, 17: 105–113, 2021.

Shwartz-Ziv, R. and Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810*, 2017.

Şimşekli, U., Sagun, L., and Giirbiizbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Smith, S. L. and Le, Q. V. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Wojtowytsch, S. Stochastic gradient descent with noise of machine learning type. Part II: Continuous time analysis. *arXiv:2106.02588*, 2021.

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V., and Zhu, Z. On the Noisy Gradient Descent that Generalizes as SGD. In *International Conference on Machine Learning*, 2020.

Wu, L., Ma, C., and Weinan, E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, 2018.

Xie, Z., Sato, I., and Sugiyama, M. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. In *International Conference on Learning Representations*, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding Deep Learning Requires Rethinking of Generalization. In *International Conference on Learning Representations*, 2017a.

Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient Langevin dynamics. In *Proceedings of Machine Learning Research*, 2017b.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, 2019.

Ziyin, L., Liu, K., Mori, T., and Ueda, M. On Minibatch Noise: Discrete-Time SGD, Overparametrization, and Bayes. *arXiv:2102.05375*, 2021.

A. List of approximations and their justifications

In Section 3.1, we made several approximations to derive the result (6). For clarity, we list the approximations made and their justifications below.

- In Eq. (8), we make the approximation of $B \ll N$, which simplifies the expression but is not essential. The main conclusion is not affected by this approximation.

- We ignore $\nabla L \nabla L^{\mathrm{T}}$ in Eq. (8), which is justified near a local minimum.

- In Eq. (9), we make the decoupling approximation (7), which is one of the key heuristic approximations in our work. This approximation is verified experimentally in Section 5.1.

- We ignore the last term of Eq. (10), which is a common approximation (Sagun et al., 2017). This approximation is justified if we are only interested in the outliers of the Hessian eigenvalues. Indeed, outliers of the Hessian eigenvalues, which play dominant roles in escape from local minima, are known to be attributable to the first term of the right-hand side of Eq. (10) (Papyan, 2018).

- In Eq. (11), we assume that the matrix $(1/N) \sum_{\mu=1}^{N} \nabla f(\theta, x^{(\mu)}) \nabla f(\theta, x^{(\mu)})^T$ does not change so much in a valley with a given local minimum. This approximation is indirectly verified in Fig. 2 (the proportionality between the loss and the noise strength implies this matrix is actually constant).

B. Proof of Theorem 4.7

We now prove Theorem 4.7. Under Assumption 4.6, dynamics of θ is restricted to the n-dimensional subspace spanned by the outlier eigenvectors $v_1, v_2, \ldots, v_n \in \mathbb{R}^P$ of the Hessian. Consequently, θ is parametrized by n variables $z_1, z_2, \ldots, z_n \in \mathbb{R}$ as

$$\theta = \theta^* + \sum_{i=1}^{n} z_i v_i. \tag{19}$$

The Hessian is a diagonal matrix in the basis of $(v_1, v_2, \ldots, v_n)$ within $\mathcal{A}_{\theta^*}$ because of Assumption 4.6. Therefore, $\partial^2 U / \partial z_i \partial z_j = 0$ for any $i \neq j$, where $U = \log L$ is the logarithmic loss introduced in Eq. (14). It implies that $\partial U / \partial z_i$ is a function of z_i, and hence U is written in the form

$$U(\theta) = \log L(\theta^*) + \sum_{i=1}^{n} U_i(z_i). \tag{20}$$

It should be noted that

$$U_i(0) = 0, \quad U_i(z_i) > 0 \text{ for any } z_i \neq 0, \quad \frac{\partial U_i}{\partial z_i}(0) = 0, \quad \frac{\partial^2 U_i}{\partial z_i^2}(0) = \frac{h_i^*}{L(\theta^*)}, \tag{21}$$

where the last equality is obtained by putting $U = \log L$ and $\partial^2 L / \partial z_i^2 \big|_{z=0} = h_i^*$.

The stochastic differential equation for the n-dimensional vector $z = (z_1, z_2, \ldots, z_n)^{\mathrm{T}}$ in the rescaled time variable τ is given by $dz = -\nabla_z U d\tau + \sqrt{2\eta \hat{H}(\theta^*)/B} d\tilde{W}_\tau$ [see Eq. (14), which is equivalent to the following Fokker-Planck equation for the distribution function $P(z, \tau)$ of z at τ:

$$\frac{\partial P(z, \tau)}{\partial \tau} = \sum_{i=1}^{n} \left[\frac{\partial}{\partial z_i} \left(\frac{\partial U_i}{\partial z_i} P \right) + \frac{\eta h_i^*}{B} \frac{\partial^2}{\partial z_i^2} P \right] =: -\nabla_z J(z), \tag{22}$$

where the probability current density $J(z) = (J_1(z), J_2(z), \ldots, J_n(z))^{\mathrm{T}}$ is explicitly given by

$$J_i(z) = -\frac{\partial U_i(z_i)}{\partial z_i} P(z) - \frac{\eta h_i^*}{B} \frac{\partial}{\partial z_i} P(z). \tag{23}$$

Under Assumption 4.4, the escape rate is evaluated by considering the MPEP. It is known that the MPEP aligns with one of the eigenvectors of the Hessian (Xie et al., 2021). Without loss of generality, let us assume that the direction of eth

eigenvector v_e corresponds to the direction of the MPEP. We denote by $z_\perp \in \mathbb{R}^{n-1}$ the displacement perpendicular to the escape direction, and write $z = (z_e, z_\perp)$. At the saddle z^s, $h_e^s < 0$ and $h_i^s > 0$ for all $i \neq e$, where $\{h_i^s\}$ is the set of eigenvalues of the Hessian at z^s, and

$$\frac{\partial U_i}{\partial z_i}(z_i^s) = 0, \quad \frac{\partial^2 U_i}{\partial z_i^2}(z_i^s) = \frac{h_i^s}{L(\theta^s)}. \tag{24}$$

We now derive the escape rate formula for asymptotically weak noise $\eta/B \to +0$ following Kramers (1940). The steady current $J \in \mathbb{R}^n$ is aligned to the escape direction, and hence $J_e \neq 0$ and $J_\perp = 0$ along the MPEP. Let us denote by P^* the total probability within the valley $\mathcal{A}_{\theta^*}$ and by $\mathcal{J}_\tau$ the total current per unit change of τ flowing to the outside of the valley through the saddle $\theta^s = \theta^* + z^s$. From Assumption 4.3, z is assumed to be in the quasi-stationary distribution. Since the quasi-stationary distribution is concentrated around θ^* in the weak-noise limit, $L(\theta_t)$ is approximately equal to $L(\theta^*)$ before the escape. Consequently, the rescaled time variable τ in Eq. (13) is approximately equal to $L(\theta^*)t$, and hence the total probability current $\mathcal{J}_t$ per unit change of t is given by $L(\theta^*)\mathcal{J}_\tau$. The escape rate κ per unit change of t is therefore given by

$$\kappa = \frac{\mathcal{J}_t}{P^*} \sim L(\theta^*)\frac{\mathcal{J}_\tau}{P^*} \quad \text{as } \frac{\eta}{B} \to +0.^3 \tag{25}$$

We now evaluate P^* and $\mathcal{J}_\tau$. First, to obtain P^*, we must know about the quasi-stationary distribution $P_s(z)$ near θ^*. It is obtained by putting $\partial P(z,\tau)/\partial \tau = 0$ in Eq. (22), which implies $J(z) = 0$. We therefore have

$$\frac{\partial P_s(z)}{\partial z_i} = -\frac{B}{\eta h_i^*}\frac{\partial U_i(z_i)}{\partial z_i}P_s(z). \tag{26}$$

By solving this equation, we obtain for $\theta = \theta^* + z \in \mathcal{A}_{\theta^*}$

$$P_s(z) = P(\theta^*) \exp\left[-\sum_{i=1}^n \frac{B}{\eta h_i^*}U_i(z_i)\right]. \tag{27}$$

For convenience, we put $P_s(z) = 0$ for $z \notin \mathcal{A}_{\theta^*}$. Then, P^*, which is the total probability within $\mathcal{A}_{\theta^*}$, is given by

$$P^* = \int_{-\infty}^\infty dz_1 \int_{-\infty}^\infty dz_2 \cdots \int_{-\infty}^\infty dz_n\, P_s(z) = P(\theta^*) \prod_{i=1}^n \int_{-\infty}^\infty dz_i\, e^{-\frac{B}{\eta h_i^*}U(z_i)}. \tag{28}$$

In the weak-noise limit $\eta/B \to +0$, we can evaluate the above integral by using the saddle-point method, which is also called the Laplace method. By using Eq. (21), the saddle-point method yields

$$\int_{-\infty}^\infty dz_i\, e^{-\frac{B}{\eta h_i^*}U(z_i)} \sim \int_{-\infty}^\infty dz_i\, e^{-\frac{B}{2\eta L(\theta^*)}z_i^2} = \left[\frac{2\pi\eta L(\theta^*)}{B}\right]^{1/2} \tag{29}$$

as $\eta/B \to +0$, and therefore

$$P^* \sim P(\theta^*)\left[\frac{2\pi\eta L(\theta^*)}{B}\right]^{n/2}. \tag{30}$$

Next, let us evaluate $\mathcal{J}_\tau$. Along the MPEP ($z_\perp = 0$), the current vector $J_\perp$ perpendicular to the escape direction is zero, and hence the probability distribution near the saddle is evaluated in a similar way as Eq. (27):

$$P(z_e = z_e^s, z_\perp) = P(z_e = z_e^s, z_\perp = 0) \exp\left[-\sum_{i(\neq e)}\frac{B}{\eta h_i^*}U_i(z_i)\right]. \tag{31}$$

By substituting it into

$$J_e = -\frac{\partial U_e}{\partial z_e}P - \frac{\eta h_e^*}{B}\frac{\partial}{\partial z_e}P, \tag{32}$$

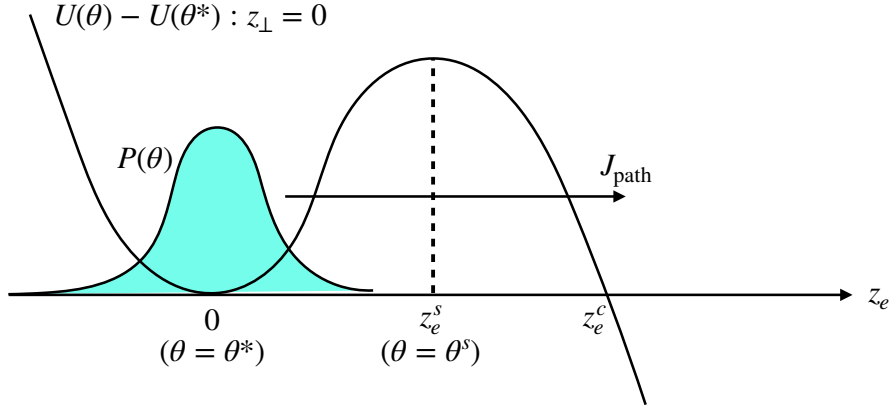³Recall that the notation "$f(x) \sim g(x)$ as $x \to +0$" means $\lim_{x\to+0}(f(x)/g(x)) = 1$.

Figure 5. Schematic illustration of the escape from a potential barrier.

we obtain

$$J_e(z_e^s, z_\perp) = J_{\text{path}} \exp\left[-\sum_{i(\neq e)} \frac{B}{\eta h_i^*} U_i(z_i)\right], \tag{33}$$

where the current along the MPEP ($z_\perp = 0$) is denoted by $J_{\text{path}} := J(z_e^s, z_\perp = 0)$. The total current through the saddle is then evaluated by using the saddle-point method:

$$\mathcal{J}_\tau = \int dz_\perp \, J_e(z_e^s, z_\perp) = J_{\text{path}} \prod_{i(\neq e)} \int_{-\infty}^{\infty} dz_\perp \, e^{-\frac{B}{\eta h_i^*} U_i(z_i)} \sim J_{\text{path}} \left[\frac{2\pi\eta L(\theta^s)}{B}\right]^{(n-1)/2} \tag{34}$$

as $\eta/B \to +0$. When the distribution function is almost stationary, Eq. (22) yields $\partial J_e(z_e, z_\perp = 0)/\partial z_e = 0$, and hence the current along the MPEP is constant $J_e(z_e, z_\perp = 0) = J_{\text{path}}$. By putting $z_\perp = 0$ in Eq. (33), we have

$$J_{\text{path}} = -\frac{\partial U}{\partial z_e} P(z_e, z_\perp = 0) - \frac{\eta h_e^*}{B} \frac{\partial P}{\partial z_e}(z_e, z_\perp = 0) = -\frac{\eta h_e^*}{B} e^{-\frac{B}{\eta h_e^*} U} \frac{\partial}{\partial z_e}\left[e^{\frac{B}{\eta h_e^*} U} P(z_e, z_\perp = 0)\right]. \tag{35}$$

By multiplying $e^{\frac{B}{\eta h_e^*} U}$ in both sides and integrating over z_e from 0 to z_e^c, where z_e^c defined as $U_e(\theta^* + z_e^c v_e) = U_e(\theta^*)$ (see Fig. 5), we obtain

$$J_{\text{path}} \int_0^{z_e^c} dz_e \, e^{\frac{B}{\eta h_e^*} U} \sim \frac{\eta h_e^*}{B} e^{\frac{B}{\eta h_e^*} U(\theta^*)} P(\theta^*) \tag{36}$$

as $\eta/B \to +0$, where we used the fact that the probability at z_e^c is negligible in the weak-noise limit. By using the saddle-point method, the integral in the left-hand side of Eq. (36) is evaluated as

$$\int_0^{z_e^c} dz_e \, e^{\frac{B}{\eta h_e^*} U} \sim \int_{-\infty}^{\infty} dz_e \, \exp\left[\frac{B}{\eta h_e^*}\left(U(\theta^s) + \frac{h_e^s}{2L(\theta^s)}(z_e - z_e^s)^2\right)\right]$$

$$= \left(\frac{2\pi\eta h_e^*}{B|h_e^s|L(\theta^s)}\right)^{1/2} e^{\frac{B}{\eta h_e^*} U(\theta^s)}. \tag{37}$$

By substituting this result in Eq. (36), we obtain

$$J_{\text{path}} \sim \left(\frac{\eta h_e^* |h_e^s|}{2\pi B L(\theta^s)}\right)^{1/2} e^{-\frac{B}{\eta h_e^*} \Delta U} P(\theta^*), \tag{38}$$

where $\Delta U = U(\theta^s) - U(\theta^*)$. The total current $\mathcal{J}_\tau$ in Eq. (34) is then expressed as

$$\mathcal{J}_\tau \sim \frac{\sqrt{h_e^* |h_e^s|}}{2\pi L(\theta^s)} \left(\frac{2\pi\eta L(\theta^s)}{B}\right)^{n/2} e^{-\frac{B}{\eta h_e^*} \Delta U} P(\theta^*). \tag{39}$$

By using Eqs. (30) and (39), the escape rate κ in Eq. (25) is evaluated as

$$\kappa \sim L(\theta^*)\frac{\mathcal{J}_\tau}{P^*} = \frac{\sqrt{h_e^*|h_e^s|}}{2\pi}\left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{\frac{n}{2}-1}e^{-\frac{B}{\eta h_e*}\Delta U}. \tag{40}$$

Since $\Delta U = \log[L(\theta^s)/L(\theta^*)]$, we finally obtain

$$\kappa \sim \frac{\sqrt{h_e^*|h_e^s|}}{2\pi}\left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{-\left(\frac{B}{\eta h_e^*}+1-\frac{n}{2}\right)}, \tag{41}$$

which is exactly identical to the escape rate formula given in Theorem 4.7.

C. Stationary distribution

Suppose that $\tilde{\theta}_\tau$ obeys a simple Langevin equation

$$d\tilde{\theta}_\tau = -U'(\tilde{\theta}_\tau) + \sqrt{2T}d\tilde{W}_\tau. \tag{42}$$

The stationary distribution of $\tilde{\theta}_\tau$ is then given by the Gibbs distribution $\tilde{P}_s(\theta) \propto e^{-U(\theta)/T}$. On the other hand, what we want is the stationary distribution $P_s(\theta)$ of θ_t, where $\theta_t = \tilde{\theta}_\tau$ with $\tau = \int_0^t dt'\, L(\theta_{t'})$. In this section, we show the relation between the two distributions: $P_s(\theta) \propto L(\theta)^{-1}\tilde{P}_s(\theta)$.

We express the stationary distributions in terms of the long-time average of the delta function:

$$P_s(\theta) = \lim_{s\to\infty}\frac{1}{s}\int_0^s dt\,\delta(\theta_t - \theta), \quad \tilde{P}_s(\theta) = \lim_{s\to\infty}\frac{1}{s}\int_0^s d\tau\,\delta(\tilde{\theta}_\tau - \theta). \tag{43}$$

By using the relation $\tau = \int_0^t dt'\, L(\theta_{t'})$, we have $d\tau = L(\theta_t)dt$. For a sufficiently large t, we also obtain $\tau \sim t\bar{L}$, where $\bar{L} := \lim_{s\to\infty}(1/s)\int_0^s dt'\, L(\theta_{t'})$ denotes the long-time average of $L(\theta_t)$. By using them, $P_s(\theta)$ is rewritten as

$$\begin{aligned}
P_s(\theta) &\approx \lim_{s\to\infty}\frac{1}{s}\int_0^{s\bar{L}} d\tau\,\frac{\delta(\tilde{\theta}_\tau - \theta)}{L(\tilde{\theta}_\tau)} \\
&= \frac{1}{L(\theta)}\lim_{s\to\infty}\frac{\bar{L}}{s\bar{L}}\int_0^{s\bar{L}} d\tau\,\delta(\tilde{\theta}_\tau - \theta) \\
&= \frac{\bar{L}}{L(\theta)}\tilde{P}_s(\theta).
\end{aligned} \tag{44}$$

We thus obtain the desired relation, $P_s(\theta) \propto L(\theta)^{-1}\tilde{P}_s(\theta)$.

D. Other loss functions

In our paper, we mainly focus on the mean-square loss, for which we can analytically derive the relation between the loss $L(\theta)$ and the SGD noise covariance $\Sigma(\theta)$. An important observation is that the SGD noise strength $\mathcal{N}$ is proportional to the loss, i.e., $\mathcal{N} \propto L(\theta)$ (see Section 5.2 for the definition of $\mathcal{N}$).

Here, we argue that the relation $\mathcal{N} \propto L(\theta)$ also holds in more general situations. During the training, the value of ℓ_μ will fluctuate from sample to sample. At a certain time step of SGD, let us suppose that $N - M$ samples in the training dataset are already fit correctly and hence $\ell_\mu \approx 0$, whereas the other M samples are not and hence $\ell_\mu = \mathcal{O}(1)$. The loss function is then given by $L(\theta) = (1/N)\sum_{\mu=1}^N \ell_\mu \propto M/N$. When ℓ_μ is small, $\nabla\ell_\mu$ will also be small. Therefore, for $N - M$ samples with $\ell_\mu \approx 0$, $\nabla\ell_\mu \approx 0$ also holds. The other M samples will have non-small gradients: $\|\nabla\ell_\mu\|^2 = \mathcal{O}(1)$. We thus estimate $\mathcal{N}$ as $\mathcal{N} \approx (1/N)\sum_{\mu=1}^N \nabla\ell_\mu^{\mathrm{T}}\nabla\ell_\mu \propto M/N \propto L(\theta)$. In this way, $\mathcal{N} \propto L(\theta)$ will hold, irrespective of the loss function.

However, we emphasize that this is a crude argument. In particular, the above argument will not hold near a global minimum because all the samples are correctly fit there, which implies that ℓ_μ is small for all μ, in contrast to the above argument relying on the existence of M samples with ℓ_μ and $\|\nabla\ell_\mu\|^2$ of $\mathcal{O}(1)$.
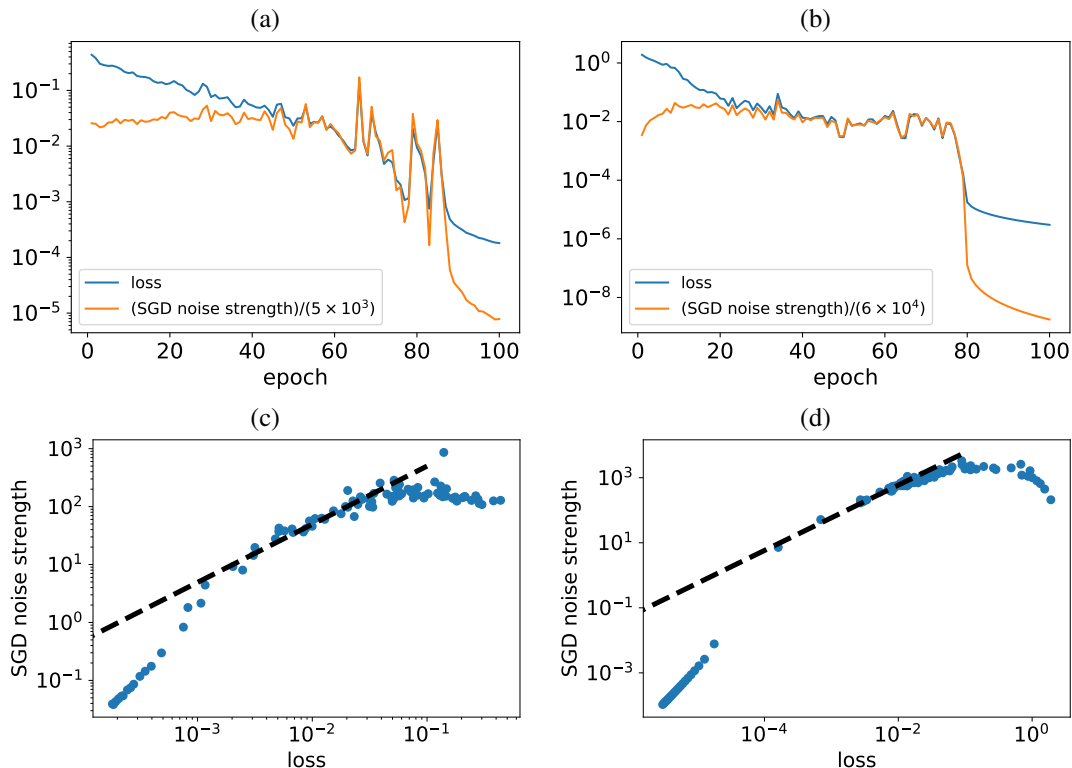
Figure 6. Training dynamics of the cross-entropy loss and the SGD noise strength $\mathcal{N}$ for (a) a fully connected network trained by the Fashion-MNIST dataset and (b) a convolutional network trained by the CIFAR-10 dataset. In the figure, we multiplied $\mathcal{N}$ by a numerical factor to emphasize that $\mathcal{N}$ is actually proportional to the loss in an intermediate stage of the training. Loss vs $\mathcal{N}$ for (c) a fully connected network trained by the Fashion-MNIST and (d) a convolutional network trained by CIFAR-10. Dashed lines in (c) and (d) are straight lines of slope 1, which imply $\mathcal{N} \propto L(\theta)$.

We now experimentally test the relation $\mathcal{N} \propto L(\theta)$ for the cross-entropy loss. We consider the same architectures and datasets in Section 5.2: a fully connected neural network trained by Fashion-MNIST and a convolutional neural network trained by CIFAR-10 (see Section 5.2 for the detail). We fix $B = 100$ in both cases, and $\eta = 0.1$ for the fully connected network and $\eta = 0.05$ for the convolutional network. Experimental results are presented in Fig. 6. We find that the relation $\mathcal{N} \propto L(\theta)$ seems to hold true at an intermediate stage of the training dynamics, although the proportionality is less clear compared with Fig. 2 in the main text for the mean-square loss.

We also find that for sufficiently small values of the loss, $\mathcal{N} \propto L(\theta)^2$ [see Fig. 6 (c) and (d)], whose implications should merit further investigation in future studies.

E. Hessian eigenvalues for a neural network in Section 5.3

We present numerical results on Hessian eigenvalues in a pre-trained neural network studied in Section 5.3. Instead of the exact Hessian, we consider an approximate Hessian given on the right-hand side of Eq. (11), i.e.,

$$H(\theta^*) \approx \frac{1}{N} \sum_{\mu=1}^{N} \nabla f(\theta, x^{(\mu)}) \nabla f(\theta, x^{(\mu)})^{\mathrm{T}}. \tag{45}$$

Eigenvalues $\{h_i\}$ are arranged in descending order as $h_1 \geq h_2 \geq \cdots \geq h_P$ (in our model $P = 7861$).

A histogram of the Hessian eigenvalues is presented in Fig. 7 (a). We see that most eigenvalues are close to zero, which corresponds to the bulk, but there are some large eigenvalues, which correspond to the outliers. The largest eigenvalue is $\lambda_1 = 95.6$, which is identified as h^* in our theoretical formula (15).
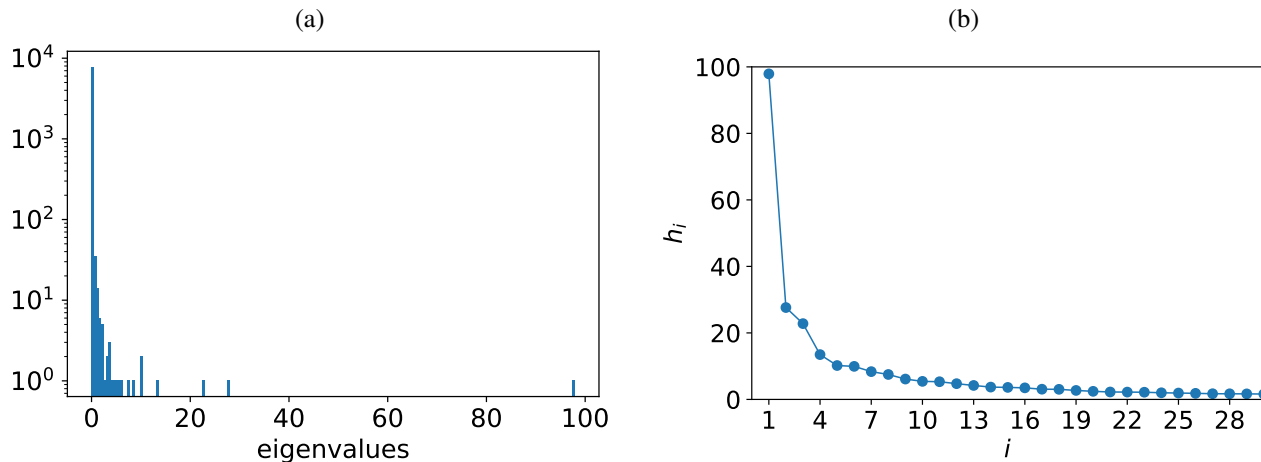
Figure 7. Hessian eigenvalues for a pre-trained neural network studied in Section 5.3. (a) The histogram of the Hessian eigenvalues and (b) top 30 eigenvalues. Most eigenvalues are close to zero, but there are some large eigenvalues, which correspond to outliers.

Another important quantity is the effective dimension n corresnding to the number of outliers. Since the outliers and the bulk are not sharply separated, it is difficult to precisely determine n. In Fig. 7 (b), we plot h_i up to $i = 30$. From this figure, it seems reasonable to estimate $n \approx 10$.

As a heuristic method of determining n, we can consider the following identification: first we define the weight p_i for ith eigenmode as $p_i = h_i / \sum_{j=1}^{P} h_j$. We then determine n as

$$n = \left(\sum_{i=1}^{P} p_i^2 \right)^{-1}, \tag{46}$$

which gives $n = 12.7$ in our case.

In Section 5.3, our formula (15) with $n = 9$ explains numerical results on the first-passage time.