When does SGD favor flat minima? A quantitative characterization via linear stability

Anonymous Author(s)
Affiliation
Address
email

Abstract

The observation that stochastic gradient descent (SGD) tends to select flat minima has played a fundamental role in understanding implicit regularization of SGD and guiding the tuning of hyperparameters. In this paper, we provide a quantitative explanation of this phenomenon by relating the particular noise structure of SGD to its *linear stability* (Wu et al., 2018). Specifically, we consider training over-parameterized models with square loss. We prove that if a global minimum θ^* is linearly stable for SGD, then it must satisfy $\|H(\theta^*)\|_F \leqslant O(\sqrt{B}/\eta)$, where $\|H(\theta^*)\|_F, B, \eta$ denote the Frobenius norm of Hessian at θ^*, batch size, and learning rate, respectively. Otherwise, SGD will escape from that minimum *exponentially* fast. Hence, for minima accessible to SGD, the flatness—as measured by the Frobenius norm of the Hessian—is bounded independently of the model size and sample size. The key to obtaining these results is exploiting the particular geometry awareness of SGD noise: 1) the noise magnitude is proportional to loss value; 2) the noise directions concentrate in the sharp directions of local landscape. This property of SGD noise provably holds for linear networks and random feature models (RFMs) and is empirically verified for nonlinear networks. Moreover, the validity and practical relevance of our theoretical findings are justified by extensive numerical experiments.

1 Introduction

Modern machine learning (ML) models are often operated with far more unknown parameters than training examples, a regime referred to as over-parameterization. In this regime, there is a large set of global minima, all of which have zero training loss but the test performance can be significantly different [45]. Fortunately, SGD always converges to those good ones even without needing any explicit regularizations [47], suggesting there must exist certain "implicit regularization" [31] mechanism at work.

More mysteriously, SGD solutions often generalize better than gradient descent (GD) solutions [20]. Therefore, the SGD noise must play a critical role in implicit regularization. The most popular explanation is that SGD converges to flatter minima [20] than GD and flatter minima generalize better [14]. This flat-minima principle has been extensively and successfully adopted in practice to tune the hyperparameters of SGD [42, 20] and to design new optimizers [16, 10, 41] for improving generalization. Therefore, understanding how SGD noise biases SGD towards flatter minima is of paramount importance, which is the main focus of this paper.

The works [42, 9, 46] show that SGD noise is highly anisotropic; [29, 39] find the magnitude of SGD noise to be loss dependent. Both structures are shown to be critical for SGD picking flat minima. However, these works all make unrealistic (even wrong) over-simplifications of SGD noise (see the related work section for more details) in their analysis. Moreover, instead of studying SGD, they all

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

consider the continuous-time stochastic differential equation (SDE), which is a good modeling of SGD only in finite time and when the learning rate (LR) is sufficiently small [24]. It is generally unclear how much the SDE modeling is relevant for understanding SGD with a large LR—a regime preferred in practice. Consequently, these works only provide intuitive and empirical analyses, lacking a quantitative characterization of when and how SGD favors flat minima.

Another line of works [44, 28] relate the selection bias of SGD to the *dynamical stability* of global minima. In over-parameterized case, all global minima are fixed points of SGD but their dynamical stabilities can be very different. At unstable minima, a small perturbation will drive SGD to leave away, whereas, for stable minima, SGD can stay around and even converge back after initial perturbations. Thus SGD prefers stable minima over unstable ones. Specifically, [44, 28] analyze the linear stability [1] of SGD, showing that a linearly stable minimum should be flat and uniform. Different from the SDE-based analysis, this stability analysis is relevant for large-LR SGD and is even empirically accurate in predicting the properties of minima selected by SGD [44, 19, 7].

In this work, we follow the linear stability analysis in [44, 28] but take the particular geometry-aware structure of SGD noise into consideration. We establish a direct connection between the linear stability and the flatness, which allows us to obtain a quantitative characterization of the flatness of minima accessible to SGD. In contrast, [44, 28] have to introduce the another quantity: *non-uniformity* together with flatness to characterize the linear stability because of neglecting the noise structure.

Setup. Let $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ be the training set and $f(\cdot;\theta)$ with $\theta \in \mathbb{R}^p$ be our model. Here p denotes the model size. We focus on training the model with *square loss*. Let $L_i(\theta) = \frac{1}{2}|f(x_i;\theta) - y_i|^2$ be the fitting error at the i-th sample and $L(\theta) = \frac{1}{n}\sum_{i=1}^n L_i(\theta)$ be the empirical risk. To minimize $L(\cdot)$, we consider the mini-batch SGD:

$$\theta_{t+1} = \theta_t - \frac{\eta}{B}\sum_{i\in I_t} \nabla L_i(\theta_t), \tag{1}$$

where η and B are the learning rate and batch size, respectively. This SGD can be rewritten as $\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t)$, where the noise ξ_t satisfies $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t\xi_t^T] = \Sigma(\theta_t)/B$ with $\Sigma(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla L_i(\theta)\nabla L_i(\theta)^T - \nabla L(\theta)L(\theta)^T$. To characterize the local geometry of loss landscape, we use the Gram matrix: $G(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla f(x_i;\theta)\nabla f(x_i;\theta)^T$ and the Hessian matrix: $H(\theta) = G(\theta) + \frac{1}{n}\sum_{i=1}^n (f(x_i;\theta) - y_i)\nabla^2 f(x_i;\theta)$. We focus on the over-parameterized regime, for which if θ^* is a global minimum, $L(\theta^*) = 0$ and $H(\theta^*) = G(\theta^*)$. In addition, throughout this paper, we denote by $\{\lambda_1(A)\}_{i\geqslant 1}$ the eigenvalues of matrix A in the non-decreasing order.

Our contributions are summarized as follows.

- We first show that for many ML models, the SGD noise is geometry aware: 1) the noise magnitude is proportional to the loss value; 2) the noise covariance aligns with the Gram matrix. Hence SGD noise is non-degenerate and concentrates along the sharp directions of local landscape. Specifically, a loss-scaled alignment factor $\mu(\theta)$ is defined to quantify the alignment strength. We prove that there exists a size-independent positive constant μ_0 such that $\mu(\theta) \geqslant \mu_0$ for RFMs (Proposition 2.4) and linear networks (Theorem 2.1).
- We then turn to analyze the linear stability of SGD with noise satisfying the above alignment property. We prove in Theorem 3.3 that if a global minimum θ^* is linearly stable, then $\|H(\theta^*)\|_F \leqslant \eta^{-1}\sqrt{B/\mu_0}$. Here the constant μ_0 quantifies the alignment strength of SGD noise. Hence, for minima accessible to SGD, the Hessian's Frobenius norm—the flatness perceived by SGD—is bounded independently of the model size and sample size. On the other hand, if a minimum is too sharp, violating the preceding stability condition, SGD will escape from it *exponentially* fast (Theorem 3.6). Together, we obtain a quantitative characterization of when and how much SGD likes flat minima.
- We provide extensive numerical experiments to justify the validity of our theoretical findings for a variety of nonlinear models. In particular, the practical relevance is demonstrated in Section 4 by considering the classification of full CIFAR-10 dataset with VGG networks and ResNets.

1.1 Related work

Noise structures. [50, 18, 26] consider the Hessian-based approximation: $\Sigma(\theta) \approx \sigma^2 H(\theta)$, where σ is a small constant. [51] proposes an improved version: $\Sigma(\theta) \approx 2L(\theta)H(\theta)$. But these approxima-

tions in general cannot be accurate since Hessian is not semi-positive definite (SPD) in non-convex region. More recently, [29] and [39] study SGD by assuming $\Sigma(\theta) = 2L(\theta)H(\theta^*)$ (θ^* is a minimum of interest) and $\Sigma(\theta) = \sigma^2 L(\theta) I_p$, respectively. These assumptions completely ignore the state-dependence of noise direction. In contrast, we only assume $\Sigma(\theta) = 2L(\theta)C(\theta)$ with $C(\theta)$ having a nontrivial alignment with the Gram matrix $G(\theta)$. Note that we do not impose any explicit structural assumption on $C(\theta)$. As a result, our assumption is much weaker and can be rigorously justified for many ML models. More importantly, we also show that this weak alignment property is sufficient for studying the linear stability of SGD.

Escape from sharp minima. This escape behavior of SGD was initiated studied in [50, 44], as an indicator of how much SGD dislikes sharp minima. It was shown that the escape happens in an unreasonably efficient way. However, the theoretical analysis there assumes the noise to be state-independent, and consequently, the derived escape time depends polynomially on the loss barrier. Later [46, 29] attempt to study this issue using the classical diffusion-based framework [11]. However, the obtained escape rates depend on the loss barrier exponentially, which cannot explain the unreasonable escape efficiency at all. See also [29] for an improved analysis. [36] argues that the SGD noise is heavy-tailed and thus SGD should be modeled as Lévy-SDE instead of Itô-SDE. Moreover, [36, 49] shows that the heavy-tailedness can ensure the escape rate depends on the basin volume instead of the loss barrier. However, the volume in high dimensions always scales exponentially with the dimension. Moreover, whether SGD noise is really heavy-tailed and whether the heavy-tailedness is really important for generalization are still debatable for neural networks [38, 25]. In contrast, we show that the unreasonable escape efficiency comes from the particular geometry-aware structure of SGD noise, regardless of whether the noise is heavy- or light-tailed.

Notions of flatness. In the literature, a variety of flatness notions have been proposed, such as the largest eigenvalue of Hessian [20], the trace of Hessian [8, 5], and many other variants [27, 37] that incorporate norm information into definitions. However, it is unclear if these flatnesses are perceivable to SGD? In our understanding, the "flatness" perceived by SGD is related to the stability, which further depends on the noise structure. Hence, different noises induce different flatnesses. Specifically, we show that for many popular ML models, due to the particular noise structure, SGD can perceive the Frobenius norm of Hessian—a flatness that characterizes the linear stability of SGD. As a comparison, the flatness perceived by GD is only the largest eigenvalue of Hessian [44, 7].

Lastly, we particularly mention the work [32], which provides a rigorous analysis of the implicit regularization of training two-layer diagonal linear networks. This work is related to ours since we both consider the magnitude and direction structure of SGD noise simultaneously. However, the analysis in [32] is limited to the specific toy model but ours is relevant for general models.

2 The geometry-aware structure of SGD noise

To gain some intuition, we first consider the *decoupling approximation* [29]:

$$\Sigma(\theta) = \frac{2}{n}\sum_{i=1}^{n} L_i(\theta)\nabla f(x_i;\theta)\nabla f(x_i;\theta)^T - \nabla L(\theta)\nabla L(\theta)^T \approx \frac{2}{n}\sum_{i=1}^{n} L_i(\theta)\nabla f(x_i;\theta)\nabla f(x_i;\theta)^T$$
$$\approx 2\big(\frac{1}{n}\sum_{i=1}^{n} L_i(\theta)\big)\frac{1}{n}\sum_{i=1}^{n}\nabla f(x_i;\theta)\nabla f(x_i;\theta)^T = 2L(\theta)G(\theta). \tag{2}$$

The second step assumes that the influence of $\nabla L(\theta)$ can be neglected; the third step assumes that $\{\nabla f(x_i;\theta)\}_i$ and $\{L_i(\theta)\}_i$ are nearly decoupled. This approximation is very crude (cannot be accurate in general) but tells us two critical properties of SGD noise: 1) The noise magnitude is proportional to the loss value; 2) the noise covariance aligns with the Gram matrix.

To utilize the above structure in a rigorous and quantitative way, we define

$$\alpha(\theta) = \frac{\mathrm{Tr}(G(\theta)\Sigma(\theta))}{\|G(\theta)\|_F\|\Sigma(\theta)\|_F},\ \beta(\theta) = \frac{\|\Sigma(\theta)\|_F}{2L(\theta)\|G(\theta)\|_F},\ \mu(\theta) = \frac{\mathrm{Tr}(\Sigma(\theta)G(\theta))}{2L(\theta)\|G(\theta)\|_F^2}. \tag{3}$$

Here $\alpha(\theta)$ quantifies the strength of standard alignment between the noise covariance and Gram matrix, where the latter characterizes the local geometry of landscape. The more $\alpha(\theta)$ is close to 1, the more the noise concentrates in sharp directions of local landscape. $\beta(\theta)$ characterizes the relative magnitude of noise with respect to the loss value. When $\beta(\theta)$ is bounded below, the influence of

noise is non-degenerate even if its absolute magnitude is small. Lastly, $\mu(\theta) = \alpha(\theta)\beta(\theta)$ can be viewed as a loss-scaled alignment factor, combining the direction and magnitude structure of noise together. When the decoupling approximation holds, we have $\alpha(\theta) = \beta(\theta) = \mu(\theta) = 1$.

The specific form of $\mu(\theta)$ defined above is inspired by our linear stability analysis in Section 3. In this paper, the noise is said to satisfy the (μ-)**alignment** if there exist a positive constant μ_0 such that

$$\mu(\theta) \geqslant \mu_0. \tag{4}$$

This alignment property ensures that in sharp directions, SGD noise is non-degenerate.

A relaxed alignment. Let $\Sigma_1(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla L_i(\theta)\nabla L_i(\theta)^T, \Sigma_2(\theta) = \nabla L(\theta)\nabla L(\theta)^T$. Then $\Sigma(\theta) = \Sigma_1(\theta) - \Sigma_2(\theta)$. It is often believed that during the late phase of training, the full-batch gradient ∇L is relatively small compared to the sample gradients $\{\nabla L_i\}_i$. As a result, the influence of $\Sigma_2(\theta)$ should be negligible compared to $\Sigma_1(\theta)$. To disentangle the influences of them, we define

$$\mu_1(\theta) = \mathrm{Tr}(\Sigma_1(\theta)G(\theta))/(2L(\theta)\|G(\theta)\|_F^2), \mu_2(\theta) = \mathrm{Tr}(\Sigma_2(\theta)G(\theta))/(2L(\theta)\|G(\theta)\|_F^2).$$

Then $\mu(\theta) = \mu_1(\theta) - \mu_2(\theta)$. Our linear stability analysis in Section 3 show that $\mu_1(\theta) \geqslant \mu_1 > 0$, a condition we refer to as **μ_1-alignment**, is sufficient to ensure that SGD only selects flat minima.

2.1 Over-parameterized linear models

Consider an over-parameterized linear model (OLM): $f(x;\theta) = F(\theta)^T x$, where $F : \Omega \mapsto \mathbb{R}^d$ denotes a general re-parameterization function. $f(\cdot;\theta)$ only represents linear functions but the corresponding landscape can be highly non-convex. Typical examples include the (deep) linear network: $F(\theta) = W_1 W_2 \cdots W_L$ and the quadratically parameterized linear model: $F(\theta) = (\alpha_1^2 - \beta_1^2, \ldots, \alpha_d^2 - \beta_d^2)^T$. The latter has attracted a lot of attention recently in analyzing the implicit regularization of GD and SGD [40, 32, 12, 3]. The following theorem provides a precise characterization of (online)-SGD noise for OLM models, whose proof is deferred to Appendix C.

Theorem 2.1. *Denote by $\mathcal{N}(0, S)$ the Gaussian distribution with mean zero and covariance matrix S. Suppose $f(\cdot;\theta)$ is a general OLM and $x \sim \mathcal{N}(0, S)$. Consider the online SGD setting, i.e., $n = \infty$. Then, $\Sigma(\theta) = \nabla L(\theta)\nabla L(\theta)^T + 2L(\theta)G(\theta)$, and $\mu_1(\theta) \geqslant \mu(\theta) \geqslant 1$*

This theorem shows that the alignments hold in the entire parameter space. Moreover, the alignment strengths are independent of model size. More importantly, this theorem provides a closed-form expression of the noise covariance, which might be useful for analyzing other properties of SGD beyond the linear stability considered in Section 3. It should be stressed that the alignment is only proved for the infinite-sample case. Similar results should hold for finite-sample cases by concentration inequalities as long as n is relatively large but this straightforward extension does not bring new insights. It is more interesting to consider the low-sample regime (i.e., $n < d$) and Figure 1 in Section 2 shows that the alignments indeed hold in this regime, at least in typical regions explored by SGD. We leave the comprehensive analysis of this regime to further work.

2.2 Feature-based models

We now turn to the feature-based model $f(x;\theta) = \sum_{j=1}^m \theta_j \varphi_j(x) = \langle \theta, \Phi(x)\rangle$. Here $\Phi(x) \in \mathbb{R}^m$ denotes the feature of x. In this case, the Hessian and Gram matrix are both constant but the SGD noise is still state-dependent. Let $g_i = \Phi(x_i)$ be the sample feature of x_i. Then,

$$\Sigma(\theta) = \frac{2}{n}\sum_{i=1}^n L_i(\theta) g_i g_i^T - \nabla L(\theta)\nabla L(\theta)^T, \quad G(\theta) = H(\theta) = \frac{1}{n}\sum_{i=1}^n g_i g_i^T.$$

Our aim is to identify conditions on the features $\{g_i\}_{i=1}^n$ such that the alignments hold for the noise.

Remark 2.2. *Although feature-based models are linear, the analysis provided below is also relevant for understanding nonlinear models. For instance, wide neural networks trained by SGD are close to RFMs [17]. A general nonlinear model may locally behave like the linearized model: $f_{\mathrm{lin}}(x;\theta) := f(x;\theta^*) + \langle \theta - \theta^*, \nabla f(x;\theta^*)\rangle$ with the tangent feature $\nabla f(x;\theta^*)$ learned from data. Hence, the analysis here at least can explain the alignment property of SGD noise in a local region, which is sufficient for characterizing the linear stability of SGD (see Section 3).*

Assumption 1. *Let $\chi_i = g_i^T G g_i, \bar{\chi} = \frac{1}{n}\sum_{i=1}^n \chi_i$. Assume $\gamma := \min_{i\in[n]} \chi_i/\bar{\chi} > 0$.*

This assumption means that the feature norms (weighted by G) are uniformly bounded below, which can ensure that the corresponding SGD noise satisfies the alignments (at least the μ_1-alignment).

Lemma 2.3. *Suppose Assumption 1 holds. Let $\lambda_1(G) \geqslant \lambda_2(G) \geqslant \cdots \geqslant \lambda_n(G)$ be the eigenvalues of G. Then $\mu_1(\theta) \geqslant \gamma, \mu_2(\theta) \leqslant \tau(G) := \lambda_1^2(G)/\sum_j \lambda_j^2(G)$, and $\mu(\theta) \geqslant \gamma - \tau(G)$.*

Proof. Noticing $\bar{\chi} = \frac{1}{n}\sum_{i=1}^n g_i^T G g_i = \|G\|_F^2$, then $\mathrm{Tr}(\Sigma_1(\theta)G) = \frac{1}{n}\sum_{i=1}^n |g_i^T\theta|^2 \,\mathrm{Tr}(g_i g_i^T G) = \frac{1}{n}\sum_{i=1}^n |g_i^T\theta|^2 \chi_i \geqslant \frac{\gamma\bar{\chi}}{n}\sum_{i=1}^n |g_i^T\theta|^2 = 2\gamma L(\theta)\|G\|_F^2$. Thus $\mu_1(\theta) \geqslant \gamma$. In addition, $\mu_2(\theta) = \frac{\nabla L(\theta)^T G \nabla L(\theta)}{2L(\theta)\|G\|_F^2} = \frac{\theta^T G^3 \theta}{\theta^T G \theta \|G\|_F^2} \leqslant \frac{\lambda_1^2(G)}{\|G\|_F^2}$, where the last step follows from Lemma E.1. □

We prove below that the uniform boundedness of feature norms can be rigorously verified for the ReLU random feature model. As a result, the μ_1- and μ-alignment provably hold with a high probability as stated in the following proposition. The proof is presented in Appendix D.3, where similar results for general RFMs are provided (see Proposition D.7).

Proposition 2.4. *Let $\varphi_j(x) = \mathrm{ReLU}(w_j^T x)$ with $w_j \overset{iid}{\sim} \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$ and $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. For any $\delta \in (0,1)$, assume $n \gtrsim d^5 \log(1/\delta)$, then w.p. at least $1-\delta$, $\mu_1(\theta) \geqslant 1$ and $\mu(\theta) \gtrsim d^{-1}$.*

2.3 Empirical validations

Reported in Figure 1a are the values of $\alpha(\theta_t), \beta(\theta_t), \mu(\theta_t)$ during the SGD training of four types of models, including the RFM and linear networks analyzed above. First, one can see that t $\alpha(\theta_t)$'s are quite close to 1 during the whole training, suggesting the concentration of SGD noise on sharp directions of local landscape. Second, $\beta(\theta_t)$'s are bounded below, implying the noise magnitudes are sufficiently large with respect to the training loss. As a result, we see clearly that the loss-scaled alignment factor $\mu(\theta_t)$ is significantly positive for all models examined.

Figure 1b further examines how the extent of over-paramaterization affects the alignment strength. One can see clearly that for linear networks and RFMs, $\mu(\theta)$'s are independent of the model size, which confirms our theoretical results proved above. In addition, we also observe that for nonlinear networks the alignment strength is also (nearly) independent of the model size.

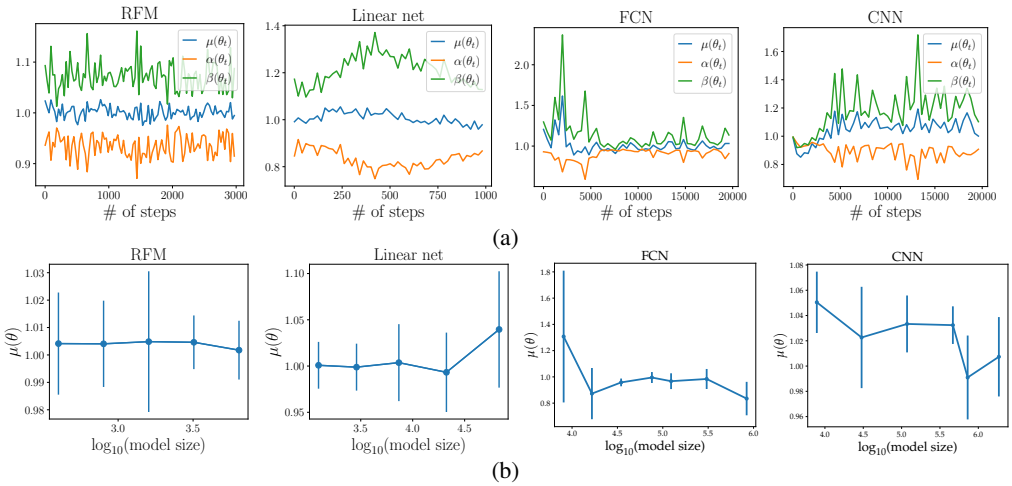
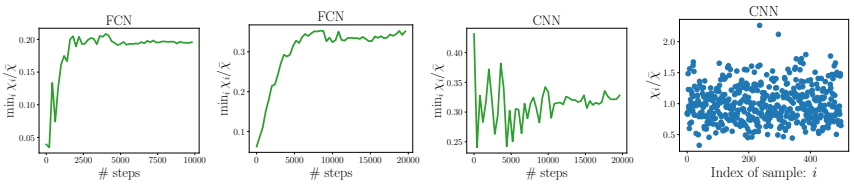


Figure 1: **The alignment property of SGD noise**. Four types of models, including the RFM, linear network, fully-connected network (FCN), and convolutional neural network (CNN), are examined. We refer to Appendix A.1 for the details of experimental setup. We particularly mention that the linear network is trained in a low-sample regime ($n = 100, d = 50$). (a) The alignment factors during the training. (b) How the alignment strength changes with the over-parameterization. Here $\mu(\theta)$'s at convergent solutions are reported; the error bar corresponds to the standard deviation over 5 runs.

In Figure 2, we plot the (relative) feature norms of three nonlinear models. Here we compute χ_i with $g_i = \nabla f(x_i;\theta)$, which corresponds to the features of linearized model $f_{\mathrm{lin}}(\cdot;\theta)$ at θ. It is shown that Assumption 1, i.e., the uniform boundedness of feature norms, holds during the SGD training. According to Lemma 2.3, this implies that the SGD noise of the local linearized model satisfies the μ_1-alignment, which is sufficient for the linear stability analysis in Section 3.



(a) (b)

Figure 2: **The feature norms are uniformly bounded below** (i.e., Assumption 1 holds) for the linearized model. (a) The values of $\gamma = \min_i \chi_i/\bar{\chi}$ during the training. (b) The feature norms of each sample at the convergent solutio of CNN; similar results of other models are provided in Appendix A.2.

We proved that the alignment of SGD noise holds in the whole parameter space for the simplified models, but this does not necessarily generalize to arbitrarily models. The experiments provided above only suggest that for nonlinear networks, the alignment property is satisfied in typical regions explored by SGD, including the random initialization and the convergent solutions. Similar observations for larger-scale models can be found in Section 4.

3 The linear stability analysis

Let θ^* be a global minimum of $L(\cdot)$. When θ_t is close to θ^*, the local dynamical behavior of SGD can be characterized by linearizing the dynamics around θ^*:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \frac{\eta}{B} \sum_{i \in I_t} \nabla^2 L_i(\theta^*)(\tilde{\theta}_t - \theta^*), \tag{5}$$

where $\nabla^2 L_i(\theta^*) = \nabla f(x_i; \theta^*) \nabla f(x_i; \theta^*)^T$. This corresponds to the local quadratic approximation of $L(\cdot)$ or the local linearization of the model around θ^*:

$$f_{\text{lin}}(x; \theta) = f(x; \theta^*) + \langle \theta - \theta^*, \nabla f(x; \theta^*) \rangle, \tag{6}$$

where $\nabla f(x; \theta^*)$ is the tangent feature [17]. Note that (5) is exactly the SGD of training (6).

Definition 3.1 (Linear stability)**.** *A global minimum θ^* is said to be linearly stable if there exists a $C > 0$ such that it holds for the linearized dynamics* (5) *that $\mathbb{E}[L(\tilde{\theta}_t)] \leqslant C\, \mathbb{E}[L(\tilde{\theta}_0)], \forall t \geqslant 0$.*

If the local quadratic approximation is non-degenerate around θ^*, then θ_t is close to $\tilde{\theta}_t$ when θ_t is close to θ^*. In other words, the local behavior of the original SGD around θ^* can be characterized by the linearized SGD. However, in the over-parameterized case, the local quadratic approximation is degenerate in flat directions. Consequently, one may be concerned about the relevance of linear stability. Fortunately, the stability is measured with the change of loss value. Thus, the instability mostly comes from noise perturbations in sharp directions and the noise also concentrates in these directions (guaranteed by the alignment property). In sharp directions, the local quadratic approximation is always validated. This explains why linear stability is still relevant in over-parameterized case. The rigorous formulation of this intuition is left to future work. In current paper, we instead resort to numerical experiments to demonstrate the validity.

For simplicity, we will use θ_t to denote $\tilde{\theta}_t$; let $\theta^* = 0$ and $g_i = \nabla f(x_i; \theta^*)$. For the linearized model $f_{\text{lin}}(\cdot; \theta)$, we have $L(\theta) = \frac{1}{2n} \sum_{i=1}^n |\theta^T g_i|^2 = \frac{1}{2} \theta^T H \theta, G = H = \frac{1}{n} \sum_{i=1}^n g_i g_i^T$, where we omit the dependence on θ^* for simplicity. Note that the Gram and Hessian matrix are constant but the noise covariance $\Sigma(\theta) = \frac{1}{n} \sum_{i=1}^n |g_i^T \theta|^2 g_i g_i^T - H\theta\theta^T H$ is still state-dependent.

Before considering the specific linearized SGD (5), we first have a general result.

Lemma 3.2. *Consider a general SGD: $\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t)$ for the linearized model* (6)*, where $(\xi_t)_{t \geqslant 1}$ are any noises satisfying $\mathbb{E}[\xi_t] = 0, \mathbb{E}[\xi_t \xi_t^T] = S(\theta_t)$. Then we have*

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[r(\theta_t) L(\theta_t) + \eta^2 v(\theta_t)], \tag{7}$$

where $\nu(\theta) = \operatorname{Tr}(HS(\theta))/2$ and $r(\theta) \geqslant 0$. Moreover, if $\eta \leqslant 2/\lambda_1(H)$, then $r(\theta) \leqslant 1$.

Proof. The update of expected loss can be written as

$$\begin{aligned} \mathbb{E}[L(\theta_{t+1})] &= \mathbb{E}[\frac{1}{2}(\theta_t - \eta \nabla L(\theta_t) + \eta \xi_t)^T H (\theta_t - \eta \nabla L(\theta_t) + \eta \xi_t)] \\ &= \mathbb{E}[L(\theta_t) - \eta \nabla L(\theta)^T H \theta_t + \frac{\eta^2}{2} \nabla L(\theta_t)^T H \nabla L(\theta_t) + \frac{\eta^2}{2} \mathbb{E}[HS(\theta_t)]], \end{aligned} \tag{8}$$

where we use $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}[\xi_t \xi_t^T] = S(\theta_t)$. Substituting $\nabla L(\theta) = H\theta$ yields $r(\theta) = 1 - 2\eta \frac{\theta^T H^2 \theta}{\theta^T H \theta} + \eta^2 \frac{\theta^T H^3 \theta}{\theta^T H \theta}$. By Lemma E.2, $r(\theta) \geqslant 0$ and if $\eta \leqslant 2/\lambda_1(H)$, then $r(\theta) \leqslant 1$. □

The two terms $r(\theta_t)L(\theta_t)$ and $\eta^2 v(\theta_t)$ denote the contributions from the full-batch gradient $\nabla L(\theta_t)$ and the noise ξ_t, respectively. The stability is affected by both terms simultaneously. It is well-known that if θ^* is linearly stable for GD, then $\lambda_1(H(\theta^*)) \leqslant 2/\eta$ (see, e.g., [44, 30]). This also holds for SGD but SGD imposes a stricter condition because of the extra $\eta^2 \nu(\theta_t)$ term. Specifically,

$$\mathbb{E}[L(\theta_{t+1})] = \mathbb{E}[r(\theta_t)L(\theta_t) + \eta^2 \nu(\theta_t)] \geqslant \eta^2 \, \mathbb{E}[\nu(\theta_t)] = 0.5\eta^2 \operatorname{Tr}(HS(\theta)). \tag{9}$$

It is implied that *the more $S(\theta)$ aligns with H, the more unstable that minimum is*.

3.1 The linear stability imposes size-independent flatness constraints

For the mini-batch SGD, the following theorem characterizes how the batch size and learning rate affect the flatness—as measured by the Frobenius norm of Hessian—of minima accessible to SGD.

Theorem 3.3. *Let θ^* be a global minimum that is linearly stable. Denote by $\mu(\theta)$ the alignment factors for the linearized SGD (5) and model (6). If $\mu(\theta) \geqslant \mu_0$, then $\|H(\theta^*)\|_F \leqslant \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}$.*

Proof. By (9) and the definition of $\mu(\theta)$, we have

$$\mathbb{E}[L(\theta_{t+1})] \geqslant \frac{\eta^2}{2B} \mathbb{E}[\operatorname{Tr}(H\Sigma(\theta_t))] \geqslant \frac{\eta^2 \|H\|_F^2}{B} \mathbb{E}[\mu(\theta_t)L(\theta_t)] \geqslant \frac{\mu_0 \eta^2 \|H\|_F^2}{B} \mathbb{E}[L(\theta_t)]. \tag{10}$$

To ensure the stability, we must have $\mu_0 \eta^2 \|H\|_F^2 / B \leqslant 1$, leading to $\|H\|_F \leqslant \sqrt{B/\mu_0}/\eta$. □

We have shown in Section 2 that μ_0 is (nearly) size-independent, and thus the obtained upper bound of flatness is also (nearly) size-independent. As a comparison, for GD, the linear stability can only ensure $\lambda_1(H(\theta^*)) \leqslant 2/\eta$. This gives a bound of the Hessian's Frobenius norm: $\|H(\theta^*)\|_F \leqslant 2\sqrt{p}/\eta$, depending on the model size explicitly. The comparison of two bounds partially explains why SGD tends to select flatter minima than GD.

We show below that the μ-alignment can be further relaxed to the μ_1-alignment. The proof is similar to the one of Theorem 3.3 and deferred to Appendix E.

Proposition 3.4. *Under the setting of Theorem 3.3, if $\mu_1(\theta) \geqslant \mu_1$, then $\|H(\theta^*)\|_F \leqslant \min\left(\frac{B}{\sqrt{(B-1)\mu_1}}, \frac{2B}{\mu_1}\right)\frac{1}{\eta}$.*

When $B \gg 1$, the bound becomes $B/(\eta\sqrt{(B-1)\mu_1}) \approx \sqrt{B/\mu_1}/\eta$, which is the same as the case of μ-alignment. Thus the influence of ∇L is indeed negligible compared with $\{\nabla L_i\}_i$.

Remark 3.5. *The linear stability is local in nature and hence our analysis essentially only needs the alignment properties of SGD noise to hold locally around minima of interest.*

Numerical validations. Figure 3 numerically shows that the actual flatness (not only the upper bound) is also nearly independent of the extent of over-parameterization. For instance, for the FCNs, the flatness only oscillates around 2.4 when increasing the model size. For the CNNs, one can see that the flatness only increases roughly from 7.6 to 8.6 as increasing the model size from 10^4 to 10^6, an increase of two orders of magnitude. Similar results for larger-scale models can be found in Section 4. These further justify our findings: the learning rate, batch size, and the alignment strength together control the complexities of solutions reachable to SGD, regardless of the model size.

3.2 SGD escapes from sharp minima exponentially fast

The following theorem shows that the pure noise-driven escape from a sharp minimum is *exponentially fast*, whose proof follows trivially from (10).

Theorem 3.6. *Under the setting of Theorem 3.3, if $\|H(\theta^*)\|_F > \frac{1}{\eta}\sqrt{\frac{B}{\mu_0}}$, then $\mathbb{E}[L(\theta_t)] \geqslant \gamma_0^t \, \mathbb{E}[L(\theta_0)]$ with $\gamma_0 = \frac{\eta^2 \mu_0}{B}\|H(\theta^*)\|_F^2 > 1$.*

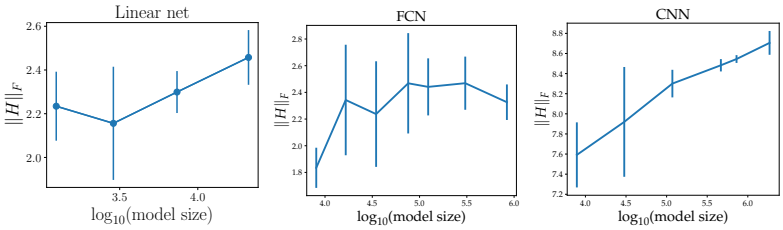


Figure 3: **The flatness of minima selected by SGD is (nearly) size-independent.** The error bar corresponds to the standard deviation estimated over 5 runs. It is shown that the flatness of SGD solutions (nearly) does not increase as increasing the model size. We refer to Appendix A.1 for the experiment details.

Hence, SGD takes roughly $\log_{\gamma_0}(1/\varepsilon)$ steps to escape from a $O(\varepsilon)$-loss region to a $O(1)$-loss region. The escape time depends on the loss barrier only logarithmically and is independent of the dimension of parameter space. This explains the unreasonable escape efficiency of SGD for training big models. In contrast, the escape rates of existing works [46, 50, 49, 29] are either exponentially slow with respect to the loss barrier or suffer from the curse of dimensionality.

Figure 4 shows the trajectories of SGD escaping from sharp minima. It is demonstrated that the escape is indeed exponentially fast and specifically, 10 steps are enough for SGD escaping to a high-loss region for all the models examined. In addition, we observe that the escape is still exponentially fast in the high-loss region, although our analysis only applies to a local region. How can we explain this nonlocal escape behavior? We leave the study of this interesting phenomenon to future work.

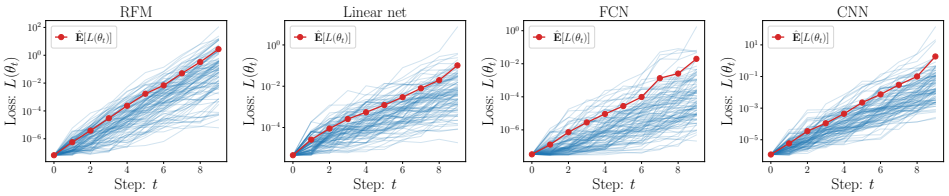


Figure 4: **The exponentially fast escape from sharp minima.** The blue curves are 200 trajectories of SGD; the orange curve corresponds to the average. The sharp minimum is found by GD. When GD nearly converge, we switch to SGD with the same learning rate. This choice ensures that the minimum is stable for GD, and thus the escape is purely driven by SGD noise. For more experimental details, we refer to Appendix A.1.

3.3 The importance of the noise structure

The magnitude structure. Theorem 3.6 together with its proof suggests that the loss dependence of noise magnitude is critical for obtaining the exponentially fast escape. The intuition is as follows. When θ_t is perturbed by noise to θ_{t+1} where $L(\theta_{t+1}) > L(\theta_t)$, the noise magnitude becomes larger there and thus θ_{t+1} is easier to be perturbed to a larger-loss region. This positive feedback drives SGD to leave exponentially fast. On the contrary, the following lemma shows that if the noise is uniformly bounded, the noise-driven escape is at most linear in time.

Lemma 3.7. *Under the setting of Lemma 3.2, assume* $\eta \leqslant 2/\lambda_1(H)$ *and* $\mathbb{E}[HS(\theta)] \leqslant 2\sigma^2$. *Then* $\mathbb{E}[L(\theta_t) - L(\theta_0)] \leqslant \eta^2\sigma^2 t$.

Here we set $\eta \leqslant 2/\lambda_1(H)$ to avoid the exponential escape caused by the GD part.

Proof. By Lemma E.2, when $\eta \leqslant 2/\lambda_1(H)$, $r(\theta) \leqslant 1$. Thus Lemma 3.2 implies $\mathbb{E}[L(\theta_{t+1})] \leqslant \mathbb{E}[L(\theta_t)] + \eta^2\sigma^2$, which implies $\mathbb{E}[L(\theta_t)] \leqslant \mathbb{E}[L(\theta_0)] + \eta^2\sigma^2 t$. □

The direction structure. We now turn to consider the impact of direction structure. Consider general SGDs: $\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \xi_t)$ with $\mathbb{E}[\xi_t\xi_t] = S(\theta_t)/B$ for the linearized model (6). We compare two type of (unrealistic) noises:

- *Geometry-aware noise:* $S_1(\theta) = 2L(\theta)H$.
- *Isotropic noise:* $S_2(\theta) = 2\sigma^2 L(\theta) I_p$ with $\sigma^2 = \mathrm{Tr}(H)/p$. Here, the value of σ^2 is chosen to ensure two types of noises have the same total variance for a fair comparison [50].

Note that p denotes the model size. Analogous to Theorem 3.3, for the second isotropic noise,

$$\mathbb{E}[L(\theta_{t+1})] \geqslant \frac{\eta^2}{2B}\operatorname{Tr}(HS(\theta)) \geqslant \mathbb{E}[L(\theta_t)]\frac{\sigma^2\eta^2}{B}\operatorname{Tr}(H) = \mathbb{E}[L(\theta_t)]\frac{\eta^2}{pB}\operatorname{Tr}(H)^2.$$

Hence, the instability decreases with the parameter-space dimension and the resulting flatness constraint is $\operatorname{Tr}(H(\theta^*)) \leqslant \sqrt{pB}/\eta$, which depends on the model size explicitly. In contrast, for the fist noise, Theorem 3.3 implies $\|H(\theta^*)\|_F \leqslant \sqrt{B}/\eta$, independent of the model size. This difference can be intuitively explained as follows. The isotropic noise wastes most energy in perturbing SGD along flat directions, which barely affects the instability. In contrast, the geometry-aware noise focuses more energy on perturbing SGD along sharp directions, causing much more instability.

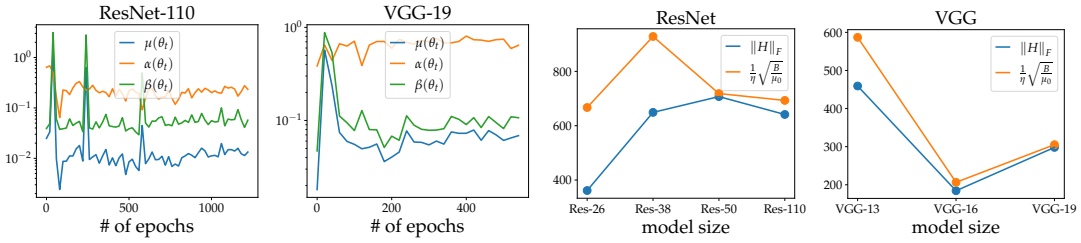
4 Larger-scale experiments

We have provided small-scale experiments to justify the validity of our theoretical findings for a variety of nonlinear models. Here we turn to demonstrate the relevance of our theory for practical models and datasets. Specifically, we consider classifying the CIFAR-10 dataset [21] with VGG networks [35] and ResNets [13] of a variety of depths. In training, all explicit regularizations are removed to keep consistent with our theoretical analysis. We follow [15] to use square loss for the classification problem. More details of the experimental setup can be found in Appendix A.1.

Figure 5a shows that the alignment factors are significantly positive during the training. We again observe that the alignment strength is (nearly) independent of the model size. In addition, Figure 9 in the appendix shows that the escape from sharp minimum is exponential for both ResNets and VGG networks. These further confirm our theoretical findings.

Figure 5b reports the flatness of convergent solution and the corresponding theoretical upper bound of different models. We again observe that the flatness is nearly independent of the model size for both ResNets and VGG networks. Moreover, we surprisingly discover that the upper bound obtained from linear stability is rather tight in predicting the flatness of convergent solution.

Lastly, we mention that the specific values of $\mu(\theta)$ in Figure 5a are on the order of 0.01, smaller than those in small-scale experiments. To understand the origin of this smallness, we conduct the same experiments but for classifying a two-class subset of CIFAR-10 and the result is reported in Figure 8 of the appendix. For this binary classification, $\mu(\theta)$'s are much larger, specifically on the order of 0.1, for all the models examined. This comparison suggests that the alignment strength significantly depends on the intrinsic complexity of the problem, (nearly) independent of the model size.



(a) (b)

Figure 5: (a) The alignment factors during the training. Here we only report the results of ResNet-110 and VGG-19. Similar results for ResNet-26, ResNet-38, ResNet-50, VGG-13, and VGG-16 are provided in Figure 7 of the appendix. (b) The flatness of SGD solution and the corresponding theoretical upper bound (Theorem 3.3).

5 Concluding remark

We provide a stability-based explanation of why SGD tends to select flat minima. Our current understanding is as follows. 1) For popular ML models, the SGD noise is geometry aware—non-degenerate in the sharp directions of local landscape. 2) This geometry awareness together with linear stability ensures that the flatness of minima accessible to SGD is size-independent. Moreover, this understanding is made rigorous and quantitative by introducing a loss-scaled alignment factor to characterize the particular noise structure.

Obviously, many questions remains open. For example, can we understand what roles the stability plays in the whole dynamic process of SGD instead of only around global minima? Can we establish the connection between the Hessian's Frobenius norm and generalization? This is important for understanding implicit regularization of SGD.

References

[1] Vladimir Igorevich Arnold. *Geometrical methods in the theory of ordinary differential equations*, volume 250. Springer Science & Business Media, 2012.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[3] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

[4] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[5] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[6] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 342–350, 2009.

[7] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

[8] Alex Damian, Tengyu Ma, and Jason Lee. Label noise SGD provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

[9] Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

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

[11] Crispin Gardiner. *Stochastic methods*, volume 4. springer Berlin, 2009.

[12] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[15] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2020.

[16] P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018.

[17] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

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

[19] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2019.

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

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer Science & Business Media, 2013.

[24] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2101–2110. PMLR, Aug 2017.

[25] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). volume 34, 2021.

[26] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

[27] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 888–896. PMLR, 2019.

[28] Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Logarithmic landscape and power-law escape rate of SGD. *arXiv preprint arXiv:2105.09557*, 2021.

[30] Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[32] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[33] Dino Sejdinovic and Arthur Gretton. What is an RKHS? *Lecture Notes*, 2012.

[34] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[36] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

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

[38] Xingyu Wang, Sewoong Oh, and Chang-Han Rhee. Eliminating sharp minima from SGD with truncated heavy-tailed noise. In *International Conference on Learning Representations*, 2022.

[39] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part II: Continuous time analysis. *arXiv preprint arXiv:2106.02588*, 2021.

[40] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[41] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

[42] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *International Conference on Machine Learning*, pages 10367–10376. PMLR, 2020.

[43] Lei Wu and Jihao Long. A spectral-based analysis of the separation between two-layer neural networks and linear methods. *Journal of Machine Learning Research*, 23(119):1–34, 2022.

[44] Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018.

[45] Lei Wu, Zhanxing Zhu, and Weinan E. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

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

[47] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[48] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019.

[49] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[51] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Local analysis
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix A.1
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Estimated over 5 runs
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]