Stochastic Gradient Descent Introduces an Effective Landscape-Dependent Regularization Favoring Flat Solutions

Ning Yang,[1] Chao Tang,[1,2,3] and Yuhai Tu[4]
[1]Peking-Tsinghua Center for Life Science, Peking University, Beijing 100871, China
[2]Center for Quantitative Biology, Peking University, Beijing 100871, China
[3]School of Physics, Peking University, Beijing 100871, China
[4]IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA

(Received 2 June 2022; revised 15 December 2022; accepted 27 April 2023; published 7 June 2023)

Generalization is one of the most important problems in deep learning, where there exist many low-loss solutions due to overparametrization. Previous empirical studies showed a strong correlation between flatness of the loss landscape at a solution and its generalizability, and stochastic gradient descent (SGD) is crucial in finding the flat solutions. To understand the effects of SGD, we construct a simple model whose overall loss landscape has a continuous set of degenerate (or near-degenerate) minima and the loss landscape for a minibatch is approximated by a random shift of the overall loss function. By direct simulations of the stochastic learning dynamics and solving the underlying Fokker-Planck equation, we show that due to its strong anisotropy the SGD noise introduces an additional effective loss term that decreases with flatness and has an overall strength that increases with the learning rate and batch-to-batch variation. We find that the additional landscape-dependent SGD loss breaks the degeneracy and serves as an effective regularization for finding flat solutions. As a result, the flatness of the overall loss landscape increases during learning and reaches a higher value (flatter minimum) for a larger SGD noise strength before the noise strength reaches a critical value when the system fails to converge. These results, which are verified in realistic neural network models, elucidate the role of SGD for generalization, and they may also have important implications for hyperparameter selection for learning efficiently without divergence.

DOI: 10.1103/PhysRevLett.130.237101

Deep learning (DL) [1] has achieved tremendous success across various fields ranging from image recognition [2] to playing Go [3] and even solving complex scientific problems such as protein folding [4]. The parameters (weights) in a neural network model are trained by following gradient decent of a global loss function. Given the large number of parameters in DL, there are many solutions that have the same (or nearly the same) minimum loss. Of course, the "goodness" of a solution is measured by its generalizability, i.e., its performance in fitting previously unseen testing data, which differs from solution to solution. Indeed, generalization remains the most important problem in DL [5]. Previous empirical studies showed that generalization correlates with the local curvature of the loss landscape around the solution: Flatter solutions tend to generalize better than sharper ones [6–12]. Recent work based on an activity-weight duality showed the dependence of the generalization loss on the flatness of loss landscape explicitly [13].

In neural networks, the overall loss function is defined on the training set $\{\boldsymbol{s}_i\}_{i=1}^{M}$ with size M:

$$L(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} l(\boldsymbol{s}_i;\boldsymbol{\theta}), \tag{1}$$

where $l(\boldsymbol{s}_i;\boldsymbol{\theta})$ is the loss for single sample $\boldsymbol{s}_i$ and $\boldsymbol{\theta}$ denotes the parameter (weight) vector of the network. From previous studies [7,14–16], the loss landscapes are highly degenerate in the overparametrized regime, where the number of network parameters greatly exceeds the independent degrees of freedom in the training samples. This is evidenced from Hessian spectrum analysis [Fig. 1(a)]: Most eigenvalues of the Hessian matrix ($\mathbf{H} \equiv \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} L$) at the solution are nearly zero; only a few eigenvalues are significantly larger than zero, which means that most directions are approximately degenerate (flat) except for very few nondegenerate (sharp) directions. Furthermore, the structure of the Hessian spectrum is stabilized after only a few epochs [17] of training [15,18], and the subspaces composed of these degenerate directions are connected without barriers [19,20]. These studies indicate that, during most of the training processes, DL searches in a highly degenerate loss landscape instead of escaping from local minima in a rugged landscape [21].

A natural question thus arises: In a degenerate loss landscape with many solutions having almost the same low loss, how does DL pick a flatter one? The answer has to do with stochastic gradient descent (SGD), which was originally adopted due to computational limitation in calculating gradients of the overall loss function with all the training

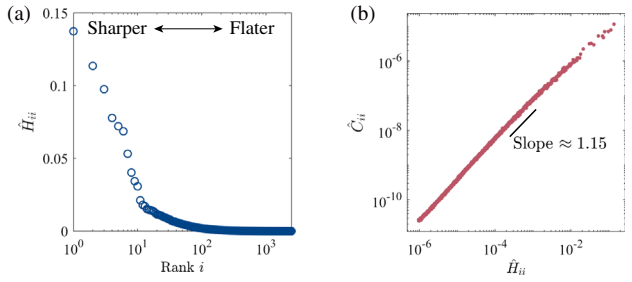

FIG. 1. (a) Rank-ordered eigenvalues of the Hessian matrix of a fully connected neural networks with two hidden layers (50 neurons each) in Modified National Institute of Standards and Technology (MNIST) classification. $\hat{\mathbf{H}}$ denotes the diagonalized Hessian matrix. (b) Top 2000 elements of the diagonalized Hessian matrix $\hat{\mathbf{H}}$ and the transformed SGD noise covariance matrix $\hat{\mathbf{C}}$. $\hat{\mathbf{C}}$ is obtained by transforming the original noise covariance matrix to the basis of $\mathbf{H}$. The network parameters are trained with SGD fixing learning rate $\eta = 0.05$ and batch size $B = 100$. The Hessian and covariance matrix are obtained after sufficient training at epoch 200. See Sec. IV A in Supplemental Material [22] for the details of the empirical experiments.

data [23,24]. Unlike the gradient descent (GD) training with all samples, SGD uses only a random minibatch of samples $\{\boldsymbol{s}_{\mu_i}\}_{i=1}^{B}$, $\mu_i \in \{1, \ldots, M\}$ with a small size $B(\ll M)$ to compute the gradients of a minibatch loss:

$$L^{\mu}(\boldsymbol{\theta}) = \frac{1}{B}\sum_{i=1}^{B} l(\boldsymbol{s}_{\mu_i}; \boldsymbol{\theta}) \tag{2}$$

and update parameters $\boldsymbol{\theta}$ at each iteration. The updating rule of SGD for iteration t is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} L^{\mu}(\boldsymbol{\theta}_t), \tag{3}$$

where η is the learning rate. The ensemble average of minibatch loss is equal to the overall loss [25], i.e., $\langle L^{\mu}(\boldsymbol{\theta})\rangle_{\mu} = L(\boldsymbol{\theta})$, and the SGD noise originates from the difference between $L(\boldsymbol{\theta})$ and $L^{\mu}(\boldsymbol{\theta})$. The SGD noise can be characterized by its covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ with its elements given as

$$C_{ij}(\boldsymbol{\theta}) = \left\langle \left(\frac{\partial L}{\partial \theta_i} - \frac{\partial L^{\mu}}{\partial \theta_i}\right)\left(\frac{\partial L}{\partial \theta_j} - \frac{\partial L^{\mu}}{\partial \theta_j}\right)\right\rangle_{\mu}. \tag{4}$$

Previous studies found that SGD noise is highly anisotropic [26]. In fact, the noise covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ was found to be correlated (aligned) with the anisotropic Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$ [6,27–29], which has also been verified by our numerical experiments shown in Fig. 1(b). Besides its anisotropic structure, the SGD noise has an overall strength that depends on the learning rate η and the minibatch size B [6,10], and increasing η or decreasing B drives the system to flatter solutions [6,7,12].

In this Letter, by using minimal models of SGD and the degenerate loss landscape, we aim to understand two general questions in DL: (i) how does the anisotropic SGD noise drive the system to flat minima in highly degenerate loss landscapes? (ii) how does the SGD noise strength affect the final solutions?

Model construction.—Based on the typical Hessian spectrum of overparametrized neural networks, we construct a degenerate loss function $L(\boldsymbol{\theta})$ where we separate all the parameters $\boldsymbol{\theta} = (\boldsymbol{x}, \boldsymbol{y})$ as nondegenerate (sharp) variables $\boldsymbol{y} \in \mathbb{R}^{N_S}$ and degenerate (flat) variables $\boldsymbol{x} \in \mathbb{R}^{N_F}$, respectively [21]:

$$L(\boldsymbol{\theta}) = L(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\sum_{i=1}^{N_S} \lambda^{(i)}(\boldsymbol{x}) y_i^2, \tag{5}$$

where $\lambda^{(i)}(\boldsymbol{x})$ are the eigenvalue function of the ith nondegenerate variable. The landscape has a region of connected degenerate minima (solutions) at $\boldsymbol{y} = \mathbf{0}$. At a given solution $(\boldsymbol{x}, \mathbf{0})$, the average flatness along the nondegenerate subspace is defined as

$$F(\boldsymbol{x}) \equiv \prod_{i=1}^{N_S} F^{(i)}(\boldsymbol{x})^{1/N_S}, \tag{6}$$

where $F^{(i)}(\boldsymbol{x}) \equiv 1/\sqrt{\lambda^{(i)}(\boldsymbol{x})}$ denotes the flatness for the ith nondegenerate variable, which is inversely correlated with the Gaussian curvature in the nondegenerate subspace. A larger $F(\boldsymbol{x})$ corresponds to a flatter minimum, and the flattest solution $(\boldsymbol{x}_F, \mathbf{0})$ is defined as the solution with the maximum $F(\boldsymbol{x})$; see Sec. I A 2 in Supplemental Material [22] for more details.

To mimic realistic SGD without explicitly introducing data, we construct an ensemble of minibatch losses by randomly shifting $L(\boldsymbol{\theta})$ with the random shifts representing the random data sampling (minibatch) in SGD. Intuition for this approach can be derived from a linear regression model, which is discussed further in Sec. I B 1 in Supplemental Material [22]. Specifically, we approximate minibatch loss $L^{\mu}(\boldsymbol{\theta})$ by taking small random shifts $\boldsymbol{\mu}$ in the overall loss $L(\boldsymbol{\theta})$: $L^{\mu}(\boldsymbol{\theta}) = L(\boldsymbol{\theta} - \boldsymbol{\mu})$. By keeping the first-order terms in $\boldsymbol{\mu}$, we have

$$L^{\mu}(\boldsymbol{\theta}) = L(\boldsymbol{\theta} - \boldsymbol{\mu}) \approx L(\boldsymbol{\theta}) - \boldsymbol{\mu} \cdot \nabla L(\boldsymbol{\theta}), \tag{7}$$

where we assume the shifts are independent identically distributed white noise with zero mean and noise strength 2σ. Empirically, noise strength σ depends inversely on the batch size, and $\sigma = 0$ corresponds to GD. Under this construction, we can show that the noise covariance matrix $\mathbf{C}$ is directly related to the Hessian matrix: $\mathbf{C} = 2\sigma\mathbf{H}^2$. Note that, in addition to the random shifts, the minibatch loss can vary in its shape and the value of its minimum [25]. These

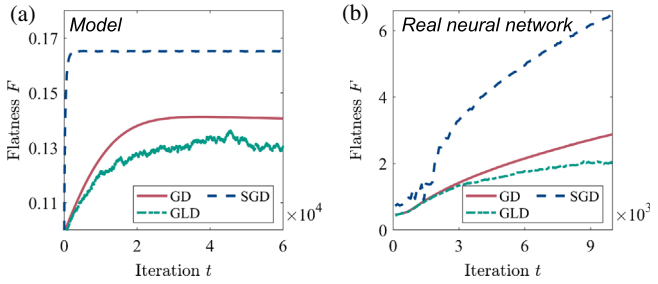

FIG. 2. Flatness dynamics of the high-dimensional model ($N_S = 5, N_F = 100$) (a) and real training in MNIST classification (b) using algorithms with different noise structures: gradient descent (red solid line), stochastic gradient descent (blue dashed line), and gradient Langevin dynamics with isotropic noise (green dash-dotted line). See Secs. II B and IV B in Supplemental Material [22] for more details.

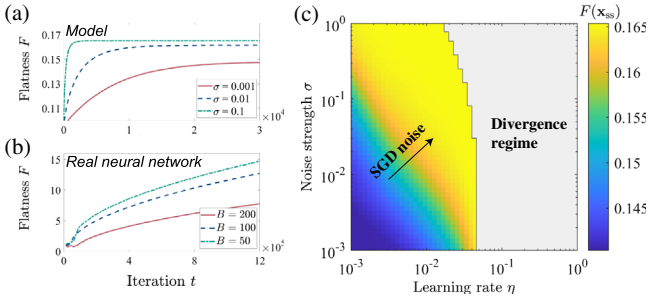

FIG. 3. (a) Flatness dynamics of the high-dimensional model ($N_S = 5$, $N_F = 100$) with fixed learning rate $\eta = 0.01$ and different noise strengths: $\sigma = 0.001$ (red solid line), $\sigma = 0.01$ (blue dashed line), and $\sigma = 0.1$ (green dash-dotted line). (b) Flatness dynamics of real training in MNIST classification with fixed learning rate $\eta = 0.05$ and different minibatch sizes: $B = 200$ (red solid line), $B = 100$ (blue dashed line), and $B = 50$ (green dash-dotted line). The flatness F in real scenarios is defined similarly as in our model where we choose the top ten eigendirections as the nondegenerate subspaces, i.e., $F \equiv \prod_{i=1}^{10} \hat{H}_{ii}{}^{-1/20}$. (c) (σ, η) phase diagram of the average flatness of the steady-state solution $F(\boldsymbol{x}_{ss})$ in the high-dimensional model. See Secs. II B and IV B in Supplemental Material [22] for more details.

variations may lead to different forms of the **C**-**H** dependence, as discussed further in Sec. I B2 in Supplemental Material [22]. However, our general conclusions hold true, regardless of the exact form of their dependence, as long as they are positively correlated (aligned), which results in an anisotropic noise that is stronger in the sharper direction, as observed in realistic SGD [Fig. 1(b)].

Evolution of flatness during learning.—By using the random-shift minibatch model [Eq. (7)] with L given in Eq. (5), we simulate the learning dynamics according to SGD updating rules [Eq. (3)]. To account for a more realistic loss landscape that is not perfectly degenerate, we use the expansion form of $\lambda^{(i)}(\boldsymbol{x})$ and include a term $(\varepsilon/2)\sum_{j=1}^{N_F}\overline{\lambda_0}x_j^2$ to slightly break the degeneracy, where $\overline{\lambda_0}$ is the average zero-order expansion coefficient and ε is a small constant ($0 < \varepsilon \ll 1$) that controls the degree of degeneracy breaking. Note that with $\varepsilon > 0$ the global minimum $(\mathbf{0}, \mathbf{0})$ is distinct from the flattest point $(\boldsymbol{x}_F, \mathbf{0})$ of the loss landscape. To demonstrate the effects of different noise structures, we also simulate the learning dynamics using GD without noise and gradient Langevin dynamics (GLD) with isotropic noise. As shown in Fig. 2(a), compared to these algorithms, the average flatness during SGD training increases faster with time and drives the system to a flatter solution. These findings from our model confirm the effects of anisotropic SGD noise in finding flatter solutions, which is consistent with the flatness dynamics observed in realistic neural networks for MNIST classification [Fig. 2(b)]. See Secs. II B and IV B in Supplemental Material [22] for details of the simulations.

Quantitatively, the flatness dynamics depends on the SGD hyperparameters. Figure 3(a) shows that a larger SGD noise strength σ drives the system to a flatter solution, which is also consistent with the flatness dynamics in real neural networks shown in Fig. 3(b). Specifically, we varied the SGD noise strength σ by changing the batch size B and found that the flatness at a given time is larger for smaller B (or, equivalently, larger σ). Increasing learning rate η has similar effects, which is also consistent with the model behavior. The case when fixing B and changing η is shown in Figs. S9 and S13 in Supplemental Material [22].

Note that the continued increase of flatness observed in Figs. 2(b) and 3(b) is due to a global scaling of weights in neural network models [25]. This scaling leads to a decrease in the cross-entropy loss function and an increase in the flatness function but does not affect test accuracy. In our high-dimensional models, there is no such scaling property, and the flatness eventually saturates after a certain training time, as shown in Figs. 2(a) and 3(a). Detailed discussions can be found in Sec. IV C in Supplemental Material [22].

The overall effects of hyperparameters are presented in a phase diagram shown in Fig. 3(c), where the average flatness of the steady-state solution $F(\boldsymbol{x}_{ss})$ obtained after the dynamics converge is shown for different choices of (η, σ). The phase diagram shows that increasing η and σ tends to increase the flatness of the solution, but excessive values of η and σ cause the system to diverge, as indicated by the gray area in Fig. 3(c).

Anisotropic noise breaks degeneracy of solutions.—To gain an analytical understanding of the results obtained empirically so far, we study a two-dimensional (2D) model, which is the minimal model that captures the key features of the degenerate loss landscape. The loss function is given by

$$L(x, y) = \frac{1}{2}\varepsilon\lambda_0 x^2 + \frac{1}{2}\lambda(x)y^2, \tag{8}$$

where we keep only one degenerate variable x and one nondegenerate variable y ($N_S = N_F = 1$) and we expand $\lambda(x) = \sum_{n=0}^{\infty} \lambda_n x^{2n}$ around the global minimum ($x^* = 0, y^* = 0$) with λ_n the $2n'$th-order expansion coefficient. As in the high-dimensional case, a small ε is introduced to break the degeneracy. Given $\varepsilon \ll 1$, points along the valley ($y = 0$) near $x = 0$ can be considered as near-degenerate solutions with low loss [$\sim\mathcal{O}(\varepsilon)$] but different flatness $F(x) \equiv 1/\sqrt{\lambda(x)}$.

Simulations of SGD dynamics in the 2D model exhibit qualitative agreement with the high-dimensional model. Beyond a critical line, larger η and σ would drive the system to the flatter solutions; see the Appendix for details. The anisotropy of SGD noise plays an important role in finding the flatter solutions, which can be demonstrated from comparison with the gradient Langevin dynamics with isotropic noise (see Fig. S3 in Supplemental Material [22]).

To demonstrate effects of the anisotropic noise, we first consider the stochastic dynamics of this system driven by gradients of fully degenerate loss function ($\varepsilon = 0$) and an anisotropic noise with constant noise strength. The dynamics of the system can be described by the following Langevin equations, which can be obtained as the continuous-time limit of SGD with a small learning rate [6,26,27,30] (see Sec. III A in Supplemental Material [22] for details):

$$\dot{x} = -\partial_x L(x,y) + \xi_1(t); \qquad \dot{y} = -\partial_y L(x,y) + \xi_2(t), \tag{9}$$

where $\xi_1(t)$ and $\xi_2(t)$ are the noise for x and y, respectively. The corresponding dynamics of the probability density $P(x,y,t)$ is governed by the following Fokker-Planck equation [31]:

$$\begin{aligned}\partial_t P = {} & \partial_x(\partial_x L + \partial_x D_{11} + \partial_y D_{12})P \\ & + \partial_y(\partial_y L + \partial_x D_{21} + \partial_y D_{22})P,\end{aligned} \tag{10}$$

where $\mathbf{D}$ is the covariance matrix of the noise. For simplicity, we first consider the case where $D_{11} = \kappa^{-1}\Delta$ and $D_{22} = \Delta$ with Δ denoting a constant noise strength in the y direction and $\kappa > 1$ characterizing the noise anisotropy with no ξ_1 ξ_2 correlation ($D_{12} = D_{21} = 0$).

By taking advantage of the separation of timescales in the degenerate and nondegenerate directions, we can integrate out the fast variable y and solve for the steady-state distribution (see Sec. III B 1 in Supplemental Material [22] for the detailed derivation) [32]:

$$P_{\rm ss}(x,y) = \frac{1}{Z_{\rm ss}} \exp\left[-\frac{\lambda(x)y^2}{2\Delta}\right] \lambda(x)^{-(\kappa-1)/2}, \tag{11}$$

where $Z_{\rm ss}$ is the normalization factor. The effective loss function $L_{\rm eff}(x,y) \equiv -\Delta \ln[Z_{\rm ss} P_{\rm ss}(x,y)]$ is given by

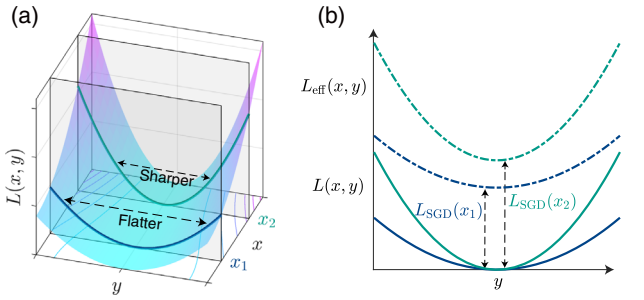

FIG. 4. (a) Illustration of two-dimensional degenerate loss function. Each point on $y = 0$ is a solution (minimum) with different flatness; the solution at x_1 (blue line) is flatter than that at x_2 (green line). (b) Loss landscape before (solid line) and after (dash-dotted line) reshaping by anisotropic noise; the landscape of flatter minimum x_1 is lower than sharper minimum x_2 after reshaping.

$$L_{\rm eff}(x,y) = \underbrace{\frac{1}{2}\lambda(x)y^2}_{L(x,y)} + \underbrace{\Delta(1-\kappa)\ln F(x)}_{L_{\rm SGD}(x)}, \tag{12}$$

which shows explicitly that the anisotropic noise reshapes the original loss landscape by introducing an additional flatness-dependent loss $L_{\rm SGD}(x) \equiv \Delta(1-\kappa)\ln F(x)$. This additional loss breaks the degeneracy in the original loss function as illustrated in Fig. 4. Since the noise covariance matrix is aligned with the Hessian [6,27–29], the SGD noise is stronger in the nondegenerate (sharp) direction, i.e., $\kappa > 1$, $L_{\rm SGD}$ is lower when $F(x)$ is larger, and, therefore, an anisotropic noise with $\kappa > 1$ is more likely to converge to the flatter solutions.

An effective flatness-dependent regularization due to SGD noise.—To better understand the general effects of the landscape-dependent SGD noise, we next consider a more realistic case where the loss landscape is not perfectly degenerate ($\varepsilon \neq 0$) and the noise covariance matrix $\mathbf{D}$ is aligned with the Hessian matrix. In our random-shift-minibatch model, $\mathbf{D}$ is found to depend on the Hessian matrix quadratically: $\mathbf{D} = \Delta_S \mathbf{H}^2$, where we define the product of two hyperparameters $\Delta_S \equiv \eta\sigma$ as the effective SGD noise level (see Sec. III A in Supplemental Material [22] for details).

Plugging the explicit expression of $\mathbf{D}$ in the Fokker-Plank equation [Eq. (10)], the steady-state distribution $P_{\rm ss}(x,y)$ can be solved approximately by integrating out the fast variable y. To characterize the SGD noise, we define an effective noise strength $\Delta(x) \equiv D_{22} \approx \Delta_S \lambda(x)^2$ and an effective anisotropy $\kappa(x) \equiv \Delta(x)/\langle D_{11}\rangle_y$, where we use $\langle\cdot\rangle_y$ to denote integration over y for a fixed x. Substituting the definition of effective loss function, i.e., $L_{\rm eff}(x,y) \equiv -\Delta(x)\ln P_{\rm ss}(x,y)$, we finally obtain the expression for the additional SGD loss $L_{\rm SGD}(x) \equiv L_{\rm eff}(x,y) - L(x,y)$ in terms of $\Delta(x)$ and $\kappa(x)$ (see Sec. III B 2 in Supplemental Material [22] for details):

$$L_{\mathrm{SGD}}(x) = \Delta(x) \int \left[\frac{\kappa(x)\varepsilon\lambda_0 x}{\Delta(x)} + \frac{\kappa(x)\lambda'(x)}{2\lambda(x)}\right] dx - \frac{\Delta(x)}{2}\ln\frac{\lambda(x)}{\Delta(x)} - \frac{1}{2}\varepsilon\lambda_0 x^2. \quad (13)$$

In the vicinity of global minimum where the noise level Δ_S is small enough to ensure the average loss $\langle L\rangle \ll 1$, $\kappa(x)$ and $\Delta(x)$ can be approximated as constants, i.e., $\Delta \approx \Delta_S\lambda_0^2 \ll 1$ and $\kappa \approx (\varepsilon^2 + 2\varepsilon\Delta_S\lambda_1 + 3\Delta_S^2\lambda_1^2)^{-1} \gg 1$, and $L_{\mathrm{SGD}}(x)$ can be expressed as

$$L_{\mathrm{SGD}}(x) \approx \Delta(1-\kappa)\ln F(x) + (\kappa-1)L(x, y=0), \quad (14)$$

which contains the flatness-dependent term $L_{\mathrm{SGD},F} \equiv \Delta(1-\kappa)\ln F(x)$ that decreases with $F(x)$ and another loss-dependent term $L_{\mathrm{SGD},L} \equiv (\kappa-1)L(x, y=0)$, which is proportional to the original loss function $L(x, y=0)$ at the nearly degenerate solutions [33] (see Sec. III B 1 in Supplemental Material [22] for details). Equation (14) shows clearly how $L_{\mathrm{SGD},F}$ compete with $L_{\mathrm{SGD},L}$: After integrating out the fast variable in the sharp direction, both $L_{\mathrm{SGD},F}$ and $L_{\mathrm{SGD},L}$ are affected by the anisotropy κ; however, the flatness-dependent SGD loss $L_{\mathrm{SGD},F}$ has an overall strength proportional to the SGD noise level Δ_S. Therefore, increasing Δ_S would increase the contribution of the flatness-dependent SGD loss $L_{\mathrm{SGD},F}$, which serves as an effective regularization that favors the flatter solutions.

For the high-dimensional case [Eq. (5)] with $N_S > 1$ and $N_F > 1$, the generalized effective loss $L_{\mathrm{SGD}}(\boldsymbol{x})$ in Eq. (14) can also be derived approximately under certain assumptions. The resulting $L_{\mathrm{SGD}}(\boldsymbol{x})$, albeit much more complex than that in 2D, also contains a flatness-dependent term $L_{\mathrm{SGD},F}$:

$$L_{\mathrm{SGD},F}(\boldsymbol{x}) \approx \sum_{i=1}^{N_S}(\bar{\Delta} - \kappa\Delta^{(i)})\ln F^{(i)}(\boldsymbol{x}), \quad (15)$$

where $\Delta^{(i)} \approx \Delta_S\lambda^{(i)}(\mathbf{0})^2$ denotes the effective noise strength for the ith nondegenerate variable, $\bar{\Delta} \equiv (1/N_S)\sum_{i=1}^{N_S}\Delta^{(i)}$ is the average noise strength over all nondegenerate variables, and κ is the overall anisotropy in the high-dimensional case (see Sec. III C 1 in Supplemental Material [22] for detailed derivations). We can see from Eq. (15) that $L_{\mathrm{SGD},F}$ also scales with the effective SGD noise level Δ_S and favors the flatter solutions.

Furthermore, we can define another average flatness:

$$\hat{F}(\boldsymbol{x}) \equiv \prod_{i=1}^{N_S} F^{(i)}(\boldsymbol{x})^{\gamma_i}, \quad (16)$$

as the geometric average flatness weighted by the relative noise strength $\gamma_i = \Delta^{(i)}/\sum_{i=1}^{N_S}\Delta^{(i)}$, which is positively correlated with the average flatness of equal weights $F(\boldsymbol{x})$. We can show that under certain assumptions $\hat{F}$ increases with time with a rate that is proportional to the SGD noise level Δ_S, which explains the flatness dynamics in Fig. 3 (see Sec. III C 2 in Supplemental Material [22] for details). Taken together, the main conclusions regarding effects of SGD from the 2D model hold true for the high-dimensional models.

Discussions.—Generalization is a fundamental problem in DL. Increasing empirical and theoretical evidence shows that a flatter solution has better generalization performance. In this Letter, we studied how SGD-based algorithms drive the learning system to flatter solution in a highly degenerate loss landscape that is typical in overparametrized neural network models. Our findings indicate that the anisotropic SGD noise introduces an additional flatness-dependent loss, which serves as an implicit regularization that favors the flatter solutions. The hyperparameters in SGD, learning rate η and batch noise level σ, together determine an overall SGD noise strength. A higher SGD noise strength within the convergence bound not only allows the system to find flatter solutions, but also speeds up the search process (see the Appendix for details). Although our current study is focused on the "solution valley," the effect of SGD in introducing an effective flatness -dependent loss function should also exist during early stage of training (learning), which may play an important role in guiding the system toward the solution valley or the flatter part of the valley. Overall, our work elucidates the effects of SGD in searching for (learning) flatter solutions, which may also shed light on the selection of SGD hyperparameters as well as design of more efficient learning algorithms.

This work of N. Y. and C. T. was supported by the National Natural Science Foundation of China (Grants No. 12090053 and No. 32088101). N. Y. acknowledges helpful discussions with Feng Yu, Jingxiang Shen, and Qiwei Yu.

Appendix: Simulation on the 2D landscape.—We simulate the learning dynamics of x and y according to SGD updating rules [Eq. (3)] using the random-shift minibath model [Eq. (7)] with the 2D loss function given in Eq. (8). To ensure that the global minimum x^* does not coincide with the flattest solution x_F, we set the expansion coefficients $\lambda_1 < 0$ and $\lambda_2 > 0$, with $\lambda_n = 0$ for $n > 2$; for the details of simulation, see Sec. II A 1 in Supplemental Material [22].

Similar to the high-dimensional case, the final steady-state solution x_{ss} (obtained by time averaging of x_t after convergence) depends on the noise strength σ and learning rate η. Figure 5(a) shows typical trajectories with fixed learning rate η and different noise strengths σ. In the parameter space spanned by (σ, η), there are three distinct phases characterized by the flatness of the solution, as shown in Fig. 5(b): For small σ and η, the steady-state solution is always the (sharper) global minimum x^*, i.e.,

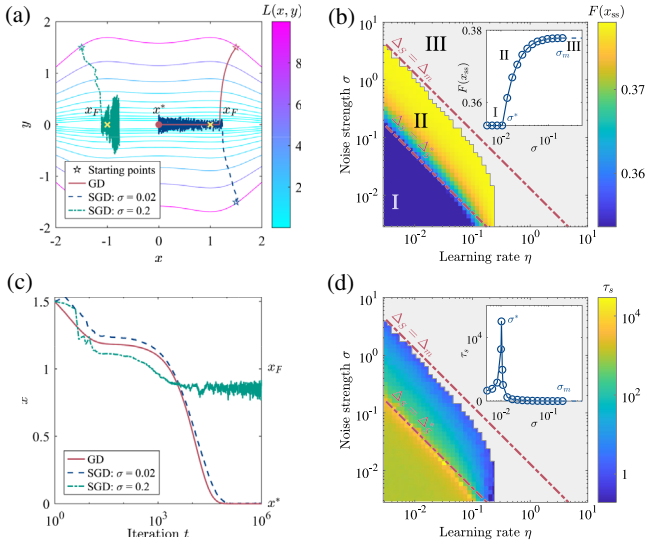

FIG. 5. (a) Trajectories of SGD on a highly degenerate loss with fixed learning rate $\eta = 0.01$ and different noise strengths: $\sigma = 0$ (red solid line), $\sigma = 0.02$ (blue dashed line), $\sigma = 0.2$ (green dash-dotted line). The starting point is (1.5,1.5), marked on symmetrical positions for visualization. Global minimum $x^* = 0$ and flattest solution $x_F = \pm 1$ are marked as a red point and yellow cross, respectively. The loss landscape parameters used in the simulation: $\varepsilon = 0.001$, $\lambda_0 = 8$, $\lambda_1 = -2$, and $\lambda_2 = 1$. (b) (σ, η) phase diagram of the flatness of steady-state solution $F(x_{ss})$. (c) Projections of the trajectories onto the degenerate direction x versus iteration t. (d) (σ, η) phase diagram for search time τ_s. In (b) and (d), the two transition boundaries between the three phases (I, II, and III) are marked by red dotted lines; the insets show $F(x_{ss})$ and τ_s versus σ for a fixed $\eta = 0.04$ with the critical σ and maximum σ denoted by σ^* and σ_m, respectively.

$F(x_{ss}) = F(x^*)$ (phase I); for intermediate σ and η, flatter solutions are found with $F(x_{ss})$ an increasing function of σ and η (phase II); for very large σ and/or η, the system diverges (phase III). The inset in Fig. 5(b) shows the three phases as we vary σ for a fixed $\eta = 0.04$, where σ^* and σ_m represent the transition boundary between the three phases.

The rate of convergence during training also depends on the SGD noise strength as shown in Fig. 5(c). To quantify the convergence rate, we define the search time $\tau_s \equiv \eta\langle t_s \rangle$ as the average first passage time t_s of reaching the solution $x = x_{ss}$ multiplied by the learning rate η (since η serves as the time interval). Analogous to critical slowing down [34], the search time is much longer near the transition point σ^* as shown in the inset in Fig. 5(d). Beyond the transition point ($\sigma > \sigma^*$), increasing the noise strength σ drastically speeds up the training processes until σ approaches its upper limit σ_m when the system diverges.

The behavior of the system in the convergence regime (phases I and II) is roughly determined by the product of these two hyperparameters $\eta\sigma$; therefore, we can define the same effective SGD noise level $\Delta_S \equiv \eta\sigma$ as in the analytical part. The transition boundary $\Delta_S = \Delta_S^*$ can be approximated by $\Delta_S^* = -\varepsilon/\lambda_1$. For large Δ_S beyond the transition boundary, a higher Δ_S would drive the system to flatter solutions with much faster speed. However, excessive SGD noise causes the system to diverge [7]. Explicitly, the convergence condition of the system satisfies $\eta\langle\lambda(x)\rangle_{ss} < 2$ with $\langle\lambda(x)\rangle_{ss} \approx \lambda(x_{ss}) + \frac{1}{2}\lambda''(x_{ss})(\langle x^2\rangle_{ss} - x_{ss}^2)$, which increases with Δ_S ($\langle\cdot\rangle_{ss}$ denotes average over the steady-state distribution; see Sec. II A 4 in Supplemental Material [22] for details). Therefore, there is a maximum noise level Δ_m, beyond which the system fails to converge [35]. The additional factor η in the convergence condition makes the dependence on σ and η asymmetric; i.e., for the same level of SGD noise (Δ_S), larger learning rate η makes the system more likely to diverge, which is consistent with numerical results shown in Figs. 5(b) and 5(d). Note that there is no sharp transition boundary between the global minimum and flatter solutions in the high-dimensional phase diagram [Fig. 3(c)], since the total contribution of all degenerate variables with different landscape parameters would blur the transition boundary.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, New York, 2016), pp. 770–778.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Mastering the game of go with deep neural networks and tree search, Nature (London) **529**, 484 (2016).

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).

[5] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma, Rethinking bias-variance trade-off for generalization of neural networks, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 10767–10777.

[6] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, Three factors influencing minima in SGD, arXiv:1711.04623.

[7] L. Wu, C. Ma *et al.*, How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective, Adv. Neural Inf. Process. Syst. **31** (2018).

[8] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning, Mol. Phys. **116**, 3214 (2018).

[9] 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, arXiv:1609.04836.

[10] S. L. Smith and Q. V. Le, A Bayesian perspective on generalization and stochastic gradient descent, in *Proceedings of the International Conference on Learning Representations* (ICLR, 2018).

[11] S. Hochreiter and J. Schmidhuber, Flat minima, Neural Comput. **9**, 1 (1997).
[12] E. Hoffer, I. Hubara, and D. Soudry, Train longer, generalize better: Closing the generalization gap in large batch training of neural networks, Adv. Neural Inf. Process. Syst. **30** (2017).
[13] Y. Feng and Y. Tu, The activity-weight duality in feed forward neural networks: The geometric determinants of generalization, arXiv:2203.10736.
[14] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, Empirical analysis of the Hessian of over-parametrized neural networks, arXiv:1706.04454.
[15] B. Ghorbani, S. Krishnan, and Y. Xiao, An investigation into neural net optimization via Hessian eigenvalue density, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2019), pp. 2232–2241.
[16] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, Entropy-SGD: Biasing gradient descent into wide valleys, J. Stat. Mech. (2019) 124018.
[17] An epoch is the training time when every sample in the training set has been used once.
[18] Y. Feng and Y. Tu, Phases of learning dynamics in artificial neural networks in the absence or presence of mislabeled data, Mach. Learn. **2**, 043001 (2021).
[19] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson, Loss surfaces, mode connectivity, and fast ensembling of dnns, Adv. Neural Inf. Process. Syst. **31** (2018).
[20] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, Essentially no barriers in neural network energy landscape, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2018), pp. 1309–1318.
[21] M. Wei and D. J. Schwab, How noise affects the Hessian spectrum in overparameterized neural networks, arXiv:1910.00195.
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.237101 for details on model construction, simulation and analytical results, as well as empirical experiments.
[23] H. Robbins and S. Monro, A stochastic approximation method, Ann. Math. Stat. **22**, 400–407 (1951).
[24] L. Bottou, Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010* (Springer, New York, 2010), pp. 177–186.
[25] Y. Feng and Y. Tu, The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima, Proc. Natl. Acad. Sci. U.S.A. **118**, e2015617118 (2021).
[26] P. Chaudhari and S. Soatto, Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks, in *Proceedings of the 2018 Information Theory and Applications Workshop (ITA)* (IEEE, New York, 2018), pp. 1–10.
[27] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma, The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2019), pp. 7654–7663.
[28] Z. Xie, I. Sato, and M. Sugiyama, A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima, in *Proceedings of the International Conference on Learning Representations* (ICLR, 2020).
[29] X. Li, Q. Gu, Y. Zhou, T. Chen, and A. Banerjee, Hessian based analysis of SGD for deep nets: Dynamics and generalization, in *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)* (Society for Industrial and Applied Mathematics, Philadelphia, 2020), pp. 190–198.
[30] Q. Li, C. Tai, and E. Weinan, Stochastic modified equations and adaptive stochastic gradient algorithms, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2017), pp. 2101–2110.
[31] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992).
[32] K. Kaneko, Adiabatic elimination by the eigenfunction expansion method, Prog. Theor. Phys. **66**, 129 (1981).
[33] The second term is absent in the purely degenerate case when $\epsilon = 0$; see Eq. (12).
[34] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, Boca Raton, 2018).
[35] The analytical approximations of transition boundaries Δ_S^*, Δ_m, and corresponding critical exponents are shown in Sec. III B in Supplemental Material [22], which includes Ref. [36].
[36] C. Gardiner, *Stochastic Methods* (Springer, Berlin, 2009), Vol. 4.