Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States

Ziqiao Wang
University of Ottawa
zwang286@uottawa.ca

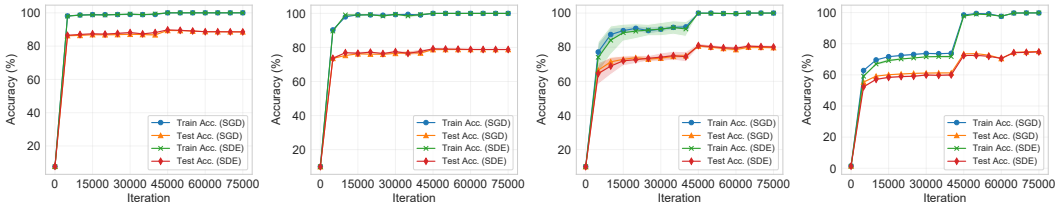
Yongyi Mao
University of Ottawa
ymao@uottawa.ca

Abstract

Stochastic differential equations (SDEs) have been shown recently to well characterize the dynamics of training machine learning models with SGD. This provides two opportunities for better understanding the generalization behaviour of SGD through its SDE approximation. Firstly, viewing SGD as full-batch gradient descent with Gaussian gradient noise allows us to obtain trajectories-based generalization bound using the information-theoretic bound. Secondly, assuming mild conditions, we estimate the steady-state weight distribution of SDE and use the information-theoretic bound to establish terminal-state-based generalization bounds.

1 Introduction

Recently, information-theoretic generalization bounds have been developed to analyze the expected generalization error of a learning algorithm. The main advantage of such bounds is that they are not only distribution-dependent, but also algorithm-dependent, making them an ideal tool for studying the generalization behaviour of models trained with a specific algorithm, such as SGD. Mutual information (MI) based bounds are first proposed by [44, 45, 61]. They are then strengthened by additional techniques [4, 34, 9, 49, 15, 53]. Particularly, Negrea et al. [34] derive MI-based bounds by developing a PAC-Bayes-like bounding technique, which upper-bounds the generalization error in terms of the KL divergence between the posterior distribution of learned model parameter given by a learning algorithm with respect to any data-dependent prior distribution. It is remarkable that the application of these information-theoretic techniques usually requires the learning algorithm to be an iterative noisy algorithm, such as stochastic gradient Langevin dynamics (SGLD) [43, 41], so as to avoid the MI bounds becoming infinity, and can not be directly applied to SGD. In order to apply such techniques to SGD, Neu et al. [35] and Wang and Mao [55] develop generalization bounds for SGD via constructing an auxiliary iterative noisy process, so additional complexity must be dealt with in that analysis.



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10 (d) ResNet on CIFAR100

Figure 1: Performance of VGG-11 and ResNet-18 trained with SGD and SDE. Standard data augmentation techniques are only used in (d).

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

Recent research has suggested that the SGD dynamics can be well approximated by using stochastic differential equations (SDEs), where the gradient signal in SGD is regarded as the full-batch gradient perturbed with an additive Gaussian noise. Specifically, [30] and [22] model this gradient noise drawn from a Gaussian distribution with a fixed covariance matrix, thereby viewing SGD as performing variational inference. [63, 56, 58, 59] further model the gradient noise as dependent of the current weight parameter and the training data. Modelling SGD in this way provide explanations as to when SGD finds flat minima [63, 58] and sharp minima [64], and inspire some new training techniques [56, 59]. Moreover, Li et al. [24, 25] and Wu et al. [56] prove that when the learning rate is sufficiently small, the SDE trajectories are theoretically close to those of SGD (see Lemma A.1). More recently, [27] has demonstrated that the SDE approximation well characterizes the optimization and generalization behavior of SGD without requiring small learning rates.

In this work, we also empirically verify the consistency between the dynamics of SGD and its associated discrete SDE (i.e. Eq. (5)). As illustrated in Figure 1, the strong agreement in their performance suggests that, despite the potential presence of non-Gaussian components in the SGD gradient noise, analyzing its SDE through a Gaussian approximation suffices for exploring SGD's generalization behavior. Furthermore, under the SDE formalism of SGD, SGD becomes an iterative noisy algorithm, on which the aforementioned information-theoretic bounding techniques can directly apply. In particular, we summarize our contributions below: (1) We obtain a generalization bound (Theorem 3.1) in the form of a summation over training steps of a quantity that involves both the the population gradient covariance and also the covariance of the gradient noise, and the generalization performance of SGD depends on the alignment of these two matrices; (2) We also apply the information-theoretic bound to obtain generalization upper bounds in terms of the KL divergence between the steady-state weight distribution of SGD with respect to a distribution-dependent prior distribution. This gives us a bound based on the alignment between the weight covariance matrix for each individual local minimum and the weight covariance matrix for the average of local minima (Theorem 4.1). Under mild assumptions, we can estimate the steady-state weight distribution of SDE (Lemma 4.1), leading to a variant of Theorem 4.1 (Corollary 4.1) and a norm-based bound (Corollary 4.2).

Other Related Literature Information-theoretic generalization bounds are typically useful to noisy iterative algorithms. For example, Pensia et al. [41] first apply the information-theoretic bound given by Xu and Raginsky [61] to analyze the generalization property of SGLD. Since the noise used in SGLD is usually an isotropic Gaussian, by utilizing the closed form of KL divergence between two Gaussian distributions, the information-theoretic generalization bound for SGLD is shown to have a tractable form. Their result is then improved by stronger bounds in [9, 34, 15, 55].

Recently, [46, 37, 47, 31, 14] challenge the traditional assumption that gradient noise is a Gaussian and argue that the noise is heavy-tailed (e.g., Lévy noise). In contrast, Xie et al. [58] and Li et al. [27] claim that non-Gaussian noise is not essential to SGD performance, and SDE with Gaussian gradient noise can well characterize the behavior of SGD. They also argue that the empirical evidence shown in [46] relies on a hidden strong assumption that gradient noise is isotropic and each dimension has the same distribution. Other works on SGD and SDE, see [20, 60, 38, 56, 63, 26, 65].

2 Preliminaries

Unless otherwise noted, a random variable will be denoted by a capitalized letter, and its realization by the corresponding lower-case letter. The distribution of a random variable X is denoted by P_X (or Q_X), and the conditional distribution of X given Y is denoted by $P_{X|Y}$. When conditioning on a specific realization y, we use the shorthand $P_{X|Y=y}$ or simply $P_{X|y}$. Denote by $\mathbb{E}_X$ expectation over $X \sim P_X$, and by $\mathbb{E}_{X|Y=y}$ (or $\mathbb{E}_X^y$) expectation over $X \sim P_{X|Y=y}$. We may omit the subscript of the expectation when there is no ambiguity. The KL divergence of probability distribution Q with respect to P is denoted by $\mathrm{D}_{\mathrm{KL}}(Q||P)$. The mutual information (MI) between random variables X and Y is denoted by $I(X;Y)$, and the conditional mutual information between X and Y given Z is denoted by $I(X;Y|Z)$. In addition, for a matrix $A \in \mathbb{R}^{d\times d}$, we let $tr\{A\}$ denote the trace of A and we use $tr\{\log A\}$ to indicate $\sum_{k=1}^{d} \log A_{k,k}$

Expected Generalization Error Let $\mathcal{Z}$ be the instance space and let μ be an unknown distribution on $\mathcal{Z}$, specifying random variable Z. We let $\mathcal{W} \subseteq \mathbb{R}^d$ be the space of hypotheses. In the information-theoretic analysis framework, there is a training sample $S = \{Z_1, Z_2, \ldots, Z_n\}_{i=1}^n$ drawn i.i.d. from μ and a stochastic learning algorithm $\mathcal{A}$ takes the training sample S as its input and outputs a hypothesis $W \in \mathcal{W}$ according to some conditional distribution $Q_{W|S}$. Given a loss function

$\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$, where $\ell(w, z)$ measures the "unfitness" or "error" of any $z \in \mathcal{Z}$ with respect to a hypothesis $w \in \mathcal{W}$. The goal of learning is to find a hypothesis w that minimizes the population risk, and for any $w \in \mathcal{W}$, the population risk is defined as $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$. In practice, since μ is only partially accessible via the sample S, we instead turn to use the empirical risk, defined as $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$. Then, the expected generalization error of $\mathcal{A}$ is defined as $\mathcal{E}_\mu(\mathcal{A}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$, where the expectation is taken over $(S, W) \sim \mu^n \otimes Q_{W|S}$.

Throughout this paper, we assume that ℓ is differentiable almost everywhere with respect to w. In some cases we will assume that $\ell(w, Z)$ is R-subgaussian for any $w \in \mathcal{W}$. Note that a bounded loss is guaranteed to be subgaussian. We will denote $\ell(w, Z_i)$ by ℓ_i when there is no ambiguity.

SGD and SDE At each time step t, given the current state $W_{t-1} = w_{t-1}$, let B_t be a random subset that is drawn uniformly from $\{1, 2, \ldots, n\}$ and $|B_t| = b$ is the batch size. Let $\widetilde{G}_t \triangleq \frac{1}{b}\sum_{i\in B_t} \nabla\ell_i$ be the mini-batch gradient. The SGD updating rule with learning rate η is then

$$W_t = w_{t-1} - \eta\widetilde{G}_t. \tag{1}$$

The full batch gradient is $G_t \triangleq \frac{1}{n}\sum_{i=1}^n \nabla\ell_i$. It follows that

$$W_t = w_{t-1} - \eta G_t + \eta V_t, \tag{2}$$

where $V_t \triangleq G_t - \widetilde{G}_t$ is the mini-batch *gradient noise*. Since $\mathbb{E}_{B_t}[V_t] = 0$, $\widetilde{G}_t$ is an unbiased estimator of the full batch gradient G_t. Moreover, the single-draw (i.e. $b = 1$) SGD gradient noise covariance (GNC) and the mini-batch GNC are $\Sigma_t = \frac{1}{n}\sum_{i=1}^n \nabla\ell_i\nabla\ell_i^T - G_tG_t^T$ and $C_t = \frac{n-b}{b(n-1)}\Sigma_t$, respectively. If $n \gg b$, then $C_t = 1/b\Sigma_t$. Notice that Σ_t (or C_t) is state-dependent, i.e. it depends on w_{t-1}. If t is not specified, we use Σ_w (or C_w) to represent its dependence on w. In addition, the population GNC at time t is

$$\Sigma_t^\mu \triangleq \mathbb{E}_Z\left[\nabla\ell(w_{t-1}, Z)\nabla\ell(w_{t-1}, Z)^T\right] - \mathbb{E}_Z\left[\nabla\ell(w_{t-1}, Z)\right]\mathbb{E}_Z\left[\nabla\ell(w_{t-1}, Z)^T\right]. \tag{3}$$

We assume that the initial parameter W_0 is independent of all other random variables, and SGD stops after T updates, outputting W_T as the learned parameter.

We now approximate V_t up to its second moment, e.g., $V_t \sim \mathcal{N}(0, C_t)$, then we have the following continuous-time evolution, i.e. Itô SDE:

$$d\omega = -\nabla L_S(\omega)dt + [\eta C_\omega]^{\frac{1}{2}}d\theta_t, \tag{4}$$

where C_ω is the GNC at ω and θ_t is a Wiener process. Furthermore, the *Euler-Maruyama* discretization, as the simplest approximation scheme to Itô SDE in Eq. (4), is

$$W_t = w_{t-1} - \eta G_t + \eta C_t^{1/2}N_t, \tag{5}$$

where $N_t \sim \mathcal{N}(0, \mathrm{I}_d)$ is the standard Gaussian random variable.

Validation of SDE It is important to understand how accurate of SDE in Eq. (4) for approximating the SGD process in Eq. (1). Previous research, such as [24, 25], has provided theoretical evidence supporting the idea that SDE can approximate SGD in a "weak sense". That is, the SDE processes closely mimic the original SGD processes, not on an individual sample path basis, but rather in terms of their distributions (see Lemma A.1 for a formal result).

Additionally, concerning the validation of the discretization of SDE in Eq. (5), Wu et al. [56, Theorem 2] has proved that Eq. (5) is *an oder* 1 *strong approximation* to SDE in Eq. (4). Moreover, we direct interested readers to the comprehensive investigations carried out by [56, 27], where the authors empirically verify that SGD and Eq. (5) can achieve the similar testing performance, suggesting that non-Gaussian noise is not essential to SGD performance. In other words, studying Eq. (5) is arguably sufficient to understand generalization properties of SGD. In Figure 1, we also empirically verify the approximation of Eq. (5), and show that it can effectively capture the behavior of SGD.

SGD and SDE Training Dynamics We implement the SDE training by following the same algorithm given in [56, Algorithm 1]. Our experiments involved training a VGG-11 architecture without BatchNormalization on a subset of SVHN (containing 25k training images) and CIFAR10.

Additionally, we trained a ResNet-18 on both CIFAR10 and CIFAR100. Data augmentation is only used in the experiments related to CIFAR100. We ran each experiment for ten different random seed, maintaining a fixed initialization of the model parameters. Further details about the experimental setup can be found in the Appendix. The results are depicted in Figure 1. As mentioned earlier, SDE exhibits a performance dynamics akin to that of SGD, reinforcing the similarities in their training behaviors.

Information-Theoretic Bound The original version of mutual information based bound is a sample-based MI bound whose main component is the mutual information between the output W and the entire input sample S. This result is given as follows:

Lemma 2.1 (Xu and Raginsky [61, Theorem 1.]). *Assume the loss $\ell(w, Z)$ is R-subGaussian for any $w \in \mathcal{W}$, then $|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} I(W;S)}$.*

3 Generalization Bounds Via Full Trajectories

We now discuss the generalization of SGD under the approximation of Eq. (5). In particular, we let $\widehat{G}_t = -G_t + C_t^{1/2} N_t$. We first have the following lemma.

Lemma 3.1. $I(\widehat{G}_t; S|W_{t-1}) = \mathbb{E}_{W_{t-1}}\left[\inf_{P_{\widehat{G}_t|W_{t-1}}} \mathbb{E}_S^{W_{t-1}}\left[\mathrm{D}_{\mathrm{KL}}(Q_{\widehat{G}_t|S,W_{t-1}}||P_{\widehat{G}_t|W_{t-1}})\right]\right]$, *where the infimum is achieved when the prior distribution $P_{\widehat{G}_t|w_{t-1}} = Q_{\widehat{G}_t|w_{t-1}}$ for any t.*

Lemma 3.1 suggests that every choice of $P_{\widehat{G}_t|W_{t-1}}$ gives rise to an upper bound of the MI of interest via $I(\widehat{G}_t; S|W_{t-1}) \leq \mathbb{E}_{W_{t-1}}\left[\mathbb{E}_S^{W_{t-1}}\left[\mathrm{D}_{\mathrm{KL}}(Q_{\widehat{G}_t|S,W_{t-1}}||P_{\widehat{G}_t|W_{t-1}})\right]\right]$. The closer is $P_{\widehat{G}_t|W_{t-1}}$ to $Q_{\widehat{G}_t|W_{t-1}}$, the tighter is the bound.

While choosing the isotropic Gaussian prior is common in the GLD or SGLD setting, given that we already know C_t is an anisotropic covariance, one can select an anisotropic prior to better incorporate the geometric structure in the prior distribution. A natural choice of the covariance is a scaled population GNC, namely $\tilde{c}_t \Sigma_t^\mu$, where $\tilde{c}_t$ is some positive state dependent scaling factor. By optimizing over c_t, we have the bound below.

Theorem 3.1. *Under the conditions of Lemma 2.1 and assume C_t and Σ_t^μ are positive-definite matrices, then $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \mathbb{E}_{W_{t-1},S}\left[tr\left\{\log \frac{\Sigma_t^\mu C_t^{-1}}{b}\right\}\right]}$.*

Remark 3.1. *If we let the diagonal element of Σ_t^μ in dimension k be $\alpha_t(k)$ and let the corresponding diagonal element of Σ_t be $\beta_t(k)$, and assume $n \gg b$ (so $\Sigma_t = bC_t$), then $tr\left\{\log(\Sigma_t^\mu C_t^{-1}/b)\right\} = \sum_{k=1}^d \log \frac{\alpha_t(k)}{\beta_t(k)}$. Thus, Theorem 3.1 implies that a favorable alignment between the diagonal values of Σ_t and Σ_t^μ will positively impact generalization performance.*

Theorem 3.1 emphasizes the significance of gradient-related information along entire trajectories in comprehending the generalization dynamics of understanding the generalization of SGD. In Figure 2, we visually show that some key gradient-based measures during SDE training closely mirror the dynamics observed in SGD.

4 Generalization Bounds Via Terminal State

In this section, we directly bound the generalization error by the properties of the terminal state instead of using the full training trajectory information. Particularly, we will first use the stationary distribution of weights at the end of training as $Q_{W_T|S}$.

Let w_s^* be a local minimum for a given training sample $S = s$, then the classical result of Mandt et al. [30] shows that the posterior $Q_{W|s}$ around w_s^* is a Gaussian distribution $\mathcal{N}(w_s^*, \Lambda_{w_s^*})$, where $\Lambda_{w^*} \triangleq \mathbb{E}\left[(W - w^*)(W - w^*)^T\right]$ is the covariance of the stationary distribution.

We are ready to give the terminal state-dependent bounds.

Theorem 4.1. *Under the conditions in Lemma 2.1. Let $w_\mu^* = \mathbb{E}[W_S^*]$ and let $\Lambda_{w_\mu^*} = \mathbb{E}\left[\left(W_T - w_\mu^*\right)\left(W_T - w_\mu^*\right)^{\mathbf{T}}\right]$, then $\mathcal{E}_\mu(\mathcal{A}) \leq \frac{R}{\sqrt{2n}} \sqrt{\mathbb{E}_{S,W_S^*}\left[tr\left\{\log\left(\Lambda_{W_S^*}^{-1} \Lambda_{w_\mu^*}\right)\right\}\right]}$.*

Note that $\Lambda_{w_\mu^*} = \mathbb{E}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^{\mathbf{T}}\right] + \mathbb{E}\left[\Lambda_{W_S^*}\right]$. By Jensen's inequality, we can bring the expectation over W_s^* inside the logarithmic function. Additionally, if $\mathbb{E}_{W_s^*}\left[\Lambda_{W_s^*}^{-1}\mathbb{E}\left[\Lambda_{W_s^*}\right]\right]$ is close to the identity matrix—especially evident in scenarios where each s has only one minimum, as in convex learning—we obtain the upper bound $\mathcal{O}\left(\sqrt{\mathbb{E}\left[\mathrm{d}_{\mathrm{M}}^2\left(W_S^*, w_\mu^*; \Lambda_{W_S^*}\right)\right]/n}\right)$, where $\mathrm{d}_{\mathrm{M}}\left(x, y; \Sigma\right) \triangleq \sqrt{(x-y)^T\Sigma^{-1}(x-y)}$ is the Mahalanobis distance. Intuitively, this quantity measures the sensitivity of a local minimum to the combined randomness introduced by both the algorithm and the training sample, relative to its local geometry.

In practice, one can estimate $\Lambda_{w_\mu^*}$ and $\Lambda_{w_s^*}$ by repeatedly conducting training processes and storing numerous checkpoints at the end of each training run. As an alternative strategy, one may leverage the analytical expression available for $\Lambda_{w_s^*}$.

Lemma 4.1. *Let H_{w^*} be the Hessian matrix of s at w^*. If $L_s(w) \approx L_s(w^*) + \frac{1}{2}(w-w^*)^T H_{w^*}(w-w^*)$ holds when w is close to any local minimal w^*, then in the long term limit, we have*

$$\Lambda_{w^*}H_{w^*} + H_{w^*}\Lambda_{w^*} - \eta H_{w^*}\Lambda_{w^*}H_{w^*} = \eta C_T.$$

Moreover, consider the conditions: (i) H_{w^*} and Λ_{w^*} commute; (ii) $H_{w^*}^{-1}\Sigma_T \approx \mathrm{I}_d$; (iii) $\frac{2}{\eta} \gg \lambda_1$ where λ_1 is the top-1 eigenvalue of H_{w^*}. Under (i), we have $\Lambda_{w^*} = \left[H_{w^*}\left(\frac{2}{\eta}\mathrm{I}_d - H_{w^*}\right)\right]^{-1} C_T$; under (i-ii), we have $\Lambda_{w^*} = (\frac{2}{\eta}\mathrm{I}_d - H_{w^*})^{-1}$; under (i-iii), we have $\Lambda_{w^*} = \frac{\eta}{2b}\mathrm{I}_d$.

Notably, all the conditions in Lemma 4.1 are only discussed in the context of the terminal state of SGD training. Regarding the condition (ii), as being widely used in the literature [22, 63, 26, 58, 59, 28], Hessian is proportional to the GNC near local minima when the loss is the negative log likelihood, i.e. cross-entropy loss. For condition (iii), the initial learning rate is typically set at a high value, and this condition may not be satisfied until the learning rate undergoes decay in the later stages of SGD training. This observation is evident in Figure 4a-4b, where the condition becomes easily met at the terminal state following the learning rate decay. Moreover, the interplay between $\frac{2}{\eta}$ and λ_1 is extensively explored in the context of the *edge of stability* [57, 10, 3], which suggests that during the training of GD, λ_1 approaches $\frac{2}{\eta}$ and hovers just above it in the "edge of stability" regime. In this case, as indicated by Lemma 4.1, the diagonal elements of $\Lambda_{w_s^*}$ tend to be close to zero before reaching the "edge of stability". Consequently, the bound presented in Theorem 4.1 diverges to infinity. This aligns with the fact that $I(W;S)$ may approach infinity for deterministic algorithms, e.g., GD with a fixed initialization.

The following results can be obtained by combining Theorem 4.1 and Lemma 4.1.

Corollary 4.1. *Under (i,iii) in Lemma 4.1, then $\mathcal{E}_\mu(\mathcal{A}) \leq \frac{R}{\sqrt{n\eta}}\sqrt{\mathbb{E}\left[tr\left\{\log\left(\left[H_{w^*}C_T^{-1}\right]\Lambda_{w_\mu^*}\right)\right\}\right]}$.*

Corollary 4.2. *Under (i-iii) in Lemma 4.1, then $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n}\log\left(\frac{2b}{\eta d}\mathbb{E}||W_S^* - w_\mu^*||^2 + 1\right)}$.*

By $\log(x+1) \leq x$, the bound in Corollary 4.2 is dimension-independent if the weight norm does not grow with d. Furthermore, the information-theoretic bound becomes a norm-based bound in Corollary 4.2, which is widely studied in the generalization literature [5, 36]. In fact, w_μ^* can be replaced by any data-independent vector, for example, the initialization, w_0 (see Corollary D.1). In this case, the corresponding bound suggests that generalization performance can be characterized by the "distance from initialization". Nagarajan and Kolter [32] also derived a "distance from initialization" based generalization bound by using Rademacher complexity, and Hu et al. [21] use "distance from initialization" as a regularizer to improve the generalization performance on noisy data.

5 Concluding Remarks

In this paper, we invoke the SDE approximation of SGD so that information-theoretic generalization bounds are directly applicable to SGD with two opportunities. First, dynamics characterized by SDE enable us to obtain trajectories-based bounds by the step-wise analysis of mutual information. In addition, with some mild assumptions, we also obtain some new bounds based on the terminal state of SGD. More theoretical and empirical results can be found in Appendix.

References

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

[2] Pierre Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

[3] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[4] Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 2017.

[6] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*. PMLR, 2018.

[7] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xHKVVHGDOEk.

[9] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591. IEEE, 2019.

[10] 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*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

[11] R Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

[12] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[13] Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In *Conference on Learning Theory*. PMLR, 2021.

[14] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*. PMLR, 2021.

[15] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

[16] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[17] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[18] Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Estimating informativeness of samples with smooth unique information. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kEnBH98BGs5.

[19] Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3): 824–839, 2020.

[20] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[21] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hke3gyHYwH.

[22] 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.

[23] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[24] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*. PMLR, 2017.

[25] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

[26] Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 2020.

[27] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 2021.

[28] Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pages 7045–7056. PMLR, 2021.

[29] Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

[30] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.

[31] Qi Meng, Shiqi Gong, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Dynamic of stochastic gradient descent with state-dependent noise. *arXiv preprint arXiv:2006.13719*, 2020.

[32] Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

[33] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 2019.

[35] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*. PMLR, 2021.

[36] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[37] Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *Advances in neural information processing systems*, 2019.

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

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[40] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[41] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018.

[42] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale), 2019.*, 2019.

[43] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[44] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*. PMLR, 2016.

[45] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

[46] 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*. PMLR, 2019.

[47] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

[48] Sidak Pal Singh, Aurelien Lucchi, Thomas Hofmann, and Bernhard Schölkopf. Phenomenology of double descent in finite-width neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lTqGXfn9Tv.

[49] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*. PMLR, 2020.

[50] Kei Takeuchi. The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153:12–18, 1976.

[51] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.

[52] Bohan Wang, Huishuai Zhang, Jieyu Zhang, Qi Meng, Wei Chen, and Tie-Yan Liu. Optimizing information-theoretical generalization bound via anisotropic noise of sgld. *Advances in Neural Information Processing Systems*, 2021.

[53] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.

[54] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. PAC-bayes information bottleneck. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iLHOIDsPv1P.

[55] Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022.

[56] 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*. PMLR, 2020.

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

[58] 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*, 2021.

[59] Zeke Xie, Li Yuan, Zhanxing Zhu, and Masashi Sugiyama. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *International Conference on Machine Learning*. PMLR, 2021.

[60] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

[61] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.

[62] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 581–590. IEEE, 2020.

[63] 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*. PMLR, 2019.

[64] Liu Ziyin, Botao Li, James B Simon, and Masahito Ueda. SGD can converge to local maxima. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9XhPLAjjRB.

[65] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uorVGbWV5sw.

A Additional Background

Lemma A.1 (Li et al. [24, Theorem 1]). *Let $\eta \in (0,1)$, $T > 0$ and $N = \lfloor T/\eta \rfloor$. Let $\mathcal{F}$ be the set of functions of polynomial growth, i.e. $f \in \mathcal{F}$, if there exists constants $K, \kappa > 0$ s.t. $|f(x)| < K(1+|x|^\kappa)$. Assume $\nabla \ell$ is Lipschitz continuous, has at most linear asymptotic growth and has sufficiently high derivatives belonging to $\mathcal{F}$, then SDE in Eq. (4) is an order 1 weak approximation of the SGD in Eq. (1). Or equivalently, for every $f \in \mathcal{F}$, there exists $C > 0$, independent of η, s.t. for all $k = 0, 1, \ldots, N$, $|\mathbb{E}[f(\omega_{k\eta})] - \mathbb{E}[f(W_k)]| < C\eta$.*

Lemma 2.1 is further improved by a data-dependent prior based bound. Following the setup in Negrea et al. [34], let J be a random subset uniformly drawn from $\{1, \ldots, n\}$ and $|J| = m > b$. Let $S_J = \{Z_i\}_{i \in J}$. Typically, we choose $m = n - 1$, then the following result is known.

Lemma A.2 (Negrea et al. [34, Theorem 2.5]). *Assume the loss $\ell(w, Z)$ is bounded in $[0, M]$, then for any $P_{W|S_J}$, $\mathcal{E}_\mu(\mathcal{A}) \leq \frac{M}{\sqrt{2}} \mathbb{E}_{S,J} \sqrt{\mathrm{D}_{\mathrm{KL}}(Q_{W|S} || P_{W|S_J})}$.*

Note that J is drawn before the training starts and is independent of $\{W_t\}_{t=0}^T$. We use the subset S_J to conduct a parallel SGD training process based to obtain a data-dependent prior ($P_{W|S_J}$). When $m = n - 1$, we call this prior process the leave-one-out (LOO) prior.

B Some Useful Lemmas

We present the variational representation of mutual information below.

Lemma B.1 (Polyanskiy and Wu [42, Corollary 3.1.]). *For two random variables X and Y, we have*

$$I(X;Y) = \inf_P \mathbb{E}_X \left[\mathrm{D}_{\mathrm{KL}}(Q_{Y|X} || P) \right],$$

where the infimum is achieved at $P = Q_Y$.

The following lemma is inspired by the classic Log sum inequality in Cover and Thomas [12, Theorem 2.7.1].

Lemma B.2. *For non-negative numbers $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$,*

$$\sum_{i=1}^n b_i \log \frac{a_i}{b_i} \leq \left(\sum_{i=1}^n b_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i},$$

with equality if and only if $\frac{a_i}{b_i} = const.$

Proof. Since $\log$ is a concave function, according to Jensen's inequality, we have

$$\sum_{i=1}^n \alpha_i \log(x_i) \leq \log(\sum_{i=1}^n \alpha_i x_i),$$

where $\sum_{i=1}^n \alpha_i = 1$.

Let $\alpha_i = \frac{b_i}{\sum_{i=1}^n b_i}$ and $x_i = \frac{a_i}{b_i}$, and plugging them into the inequality above, we have

$$\sum_{i=1}^n \frac{b_i}{\sum_{i=1}^n b_i} \log(\frac{a_i}{b_i}) \leq \log \left(\sum_{i=1}^n \frac{b_i}{\sum_{i=1}^n b_i} \frac{a_i}{b_i} \right) = \log \left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right),$$

which implies

$$\sum_{i=1}^n b_i \log(\frac{a_i}{b_i}) \leq \left(\sum_{i=1}^n b_i \right) \log \left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right).$$

This completes the proof. □

Below is the KL divergence between two Gaussian distributions $p = \mathcal{N}(\mu_p, \Sigma_p)$ and $q = \mathcal{N}(\mu_q, \Sigma_q)$, where $\mu_p, \mu_q \in \mathbb{R}^d$ and $\Sigma_p, \Sigma_q \in \mathbb{R}^{d \times d}$.

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \frac{1}{2} \left[\log \frac{\det(\Sigma_q)}{\det(\Sigma_p)} - d + (\mu_p - \mu_q)^T \Sigma_q^{-1} (\mu_p - \mu_q) + tr\left\{ \Sigma_q^{-1} \Sigma_p \right\} \right]. \tag{6}$$

C Omitted Proofs and Additional Results in Section 3

C.1 Lemma C.1: Unrolling Mutual Information

We first unroll the terminal parameters' mutual information $I(W_T; S)$ to the full trajectories' mutual information via the lemma below.

Lemma C.1. $I(W_T; S) \leq \sum_{t=1}^{T} I(-G_t + C_t^{1/2} N_t; S|W_{t-1})$.

This lemma can be proved by recurrently applying the data processing inequality (DPI) and chain rule of the mutual information [42].

Proof. Recall the SDE approximation of SGD, i.e., Eq (5), we then have,

$$\begin{aligned}
I(W_T; S) &= I(W_{T-1} - \eta G_T + \eta C_T^{1/2} N_T; S) \\
&\leq I(W_{T-1}, -\eta G_T + \eta C_T^{1/2} N_T; S) \qquad (7) \\
&= I(W_{T-1}; S) + I(-\eta G_T + \eta C_T^{1/2} N_T; S|W_{T-1}) \qquad (8) \\
&\vdots \\
&\leq \sum_{t=1}^{T} I(-\eta G_t + \eta C_t^{1/2} N_t; S|W_{t-1}) \\
&= \sum_{t=1}^{T} I(-G_t + C_t^{1/2} N_t; S|W_{t-1}).
\end{aligned}$$

where Eq. (7) is by the data processing inequality (e.g., $Z - (X, Y) - (X + Y)$ form a markov chain then $I(X + Y, Z) \leq I(X, Y; Z)$), Eq. (8) is by the chain rule of the mutual information, and learning rate η is dropped since mutual information is scale-invariant. □

C.2 Proof of Lemma 3.1

Proof. For any $t \in [T]$, similar to the proof of Lemma B.1 in [42]:

$$\begin{aligned}
& I(-G_t + C_t^{1/2} N_t; S|W_{t-1} = w_{t-1}) \\
= \; & \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || Q_{\widehat{G}_t|w_{t-1}}) \right] \\
= \; & \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || P_{\widehat{G}_t|w_{t-1}}) - \mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1}} || P_{\widehat{G}_t|w_{t-1}}) \right] \\
\leq \; & \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || P_{\widehat{G}_t|w_{t-1}}) \right], \qquad (9)
\end{aligned}$$

where Eq. (9) is due to the fact that KL divergence is non-negative, and the equality holds when $P_{\widehat{G}_t|w_{t-1}} = Q_{\widehat{G}_t|w_{t-1}}$ for $W_{t-1} = w_{t-1}$.

Thus, we conclude that

$$I(\widehat{G}_t; S|W_{t-1} = w_{t-1}) = \inf_{P_{\widehat{G}_t|w_{t-1}}} \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || P_{\widehat{G}_t|w_{t-1}}) \right].$$

Taking expectation over W_{t-1} for both side above, we have

$$I(\widehat{G}_t; S|W_{t-1}) = \mathbb{E}_{W_{t-1}} \left[\inf_{P_{\widehat{G}_t|W_{t-1}}} \mathbb{E}_S^{W_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|W_{t-1},S} || P_{\widehat{G}_t|W_{t-1}}) \right] \right].$$

This completes the proof. □

C.3 Proof of Theorem 3.1

Proof. Recall Lemma 3.1, we have

$$
\begin{aligned}
&I(-G_t + C_t^{1/2} N_t; S | W_{t-1} = w_{t-1}) \\
\leq &\inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D}_{\mathrm{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || P_{\widehat{G}_t|w_{t-1}}) \right] \\
= &\inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[\frac{1}{2} \left[\log \frac{\det(\tilde{c}_t \Sigma_t^\mu)}{\det(C_t)} - d + \frac{1}{\tilde{c}_t} ((G_t - \tilde{g}_t)^T (\Sigma_t^\mu)^{-1} (G_t - \tilde{g}_t)) + \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} C_t \right\} \right] \right] \\
= &\frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} \left[(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^T \right] \right\} \\
&\quad + \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} [C_t] \right\} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \\
= &\frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t n} tr \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + \frac{n-b}{\tilde{c}_t b n} tr \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \qquad (10) \\
= &\frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{\tilde{c}_t n} + \frac{(n-b)d}{\tilde{c}_t b n} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \\
= &\frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{b \tilde{c}_t} + d \log \tilde{c}_t + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} - d \\
= &\frac{d}{2} \log \frac{1}{b} + \frac{1}{2} tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\},
\end{aligned}
$$

where the last equality hold when $\tilde{c}_t^* = 1/b$ and Eq. (10) is by

$$
\mathbb{E}_S^{w_{t-1}} \left[(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^T \right] = \frac{1}{n} \Sigma_t^\mu, \quad \text{and}
$$

$$
\mathbb{E}_S^{w_{t-1}} [C_t] = \frac{n-b}{b(n-1)} \mathbb{E}_S^{w_{t-1}} [\Sigma_t] = \frac{n-b}{b(n-1)} \frac{n-1}{n} \Sigma_t^\mu = \frac{n-b}{bn} \Sigma_t^\mu.
$$

This completes the proof. □

D Omitted Proofs, Additional Results and Discussions in Section 4

In fact, this section provides a PAC-Bayes type analysis. The connection between information-theoretic bounds and PAC-Bays bounds have already been discussed in many previous works [6, 19, 2]. Roughly speaking, the most significant component of a PAC-Bayes bound is the KL divergence between the posterior distribution of a randomized algorithm output and a prior distribution, i.e. $\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S} || P_N)$ for some prior P_N. In essence, information-theoretic bounds can be view as having the same spirit. For concreteness, in Lemma 2.1, $I(W_T; S) = \mathbb{E}_S[\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S} || P_{W_T})]$, in which case the marginal P_{W_T} is used as a prior of the algorithm output. Furthermore, by using Lemma B.1, we have $I(W_T; S) \leq \inf_{P_N} \mathbb{E}_S[\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S} || P_N)]$. Hence, Lemma 2.1 can be regarded as a PAC-Bayes bound with the optimal prior. In addition, the PAC-Bayes framework is usually used to provide a high-probability bound, while information-theoretic analysis is applied to bounding the expected generalization error. In this sense, information-theoretic framework is closer to another concept called MAC-Bayes [13].

D.1 Proof of Lemma 4.1

Proof. When w is close to any local minimum w^*, we can use a second-order Taylor expansion to approximate the value of the loss at w,

$$
L_s(w) \approx L_s(w^*) + \frac{1}{2} (w - w^*)^{\mathbf{T}} H_{w^*} (w - w^*). \qquad (11)
$$

Then, when $w_t \to w^*$, we have $G_t = \nabla L_s(w_t) = H_{w^*} (w_t - w^*)$. Recall Eq. (2), then

$$
\begin{aligned}
w_t = &w_{t-1} - \eta G_t + \eta V_t \\
= &w_{t-1} - \eta H_{w^*} (w_{t-1} - w^*) + \eta V_t.
\end{aligned}
$$

Let $W'_t \triangleq W_t - w^*$. Thus, as $T \to \infty$,

$$
\begin{aligned}
&\mathbb{E}_{W'_T}\left[W'_T {W'_T}^{\mathbf{T}}\right] \\
=&\mathbb{E}_{W'_{T-1},V_T}\left[\left(W'_{T-1} - \eta H_{w^*} W'_{T-1} + \eta V_t\right)\left(W'_{T-1} - \eta H_{w^*} W'_{T-1} + \eta V_t\right)^{\mathbf{T}}\right] \\
=&\mathbb{E}_{W'_{T-1}}\left[W'_{T-1}{W'}^{\mathbf{T}}_{T-1} - \eta H_{w^*} W'_{T-1}{W'}^{\mathbf{T}}_{T-1} - \eta W'_{T-1}{W'}^{\mathbf{T}}_{T-1} H_{w^*} + \eta^2 H_{w^*} W'_{T-1}{W'}^{\mathbf{T}}_{T-1} H_{w^*}\right] \\
&\qquad + \eta^2 \mathbb{E}_{V_T}\left[V_T {V_T}^{\mathbf{T}}\right],
\end{aligned}
$$

where the last equation is by $\mathbb{E}_{V_T}^{w_{T-1}}[V_T] = 0$.

Recall that $\mathbb{E}_{V_T}\left[V_T {V_T}^{\mathbf{T}}\right] = C_T$ and notice that $\mathbb{E}_{W'_T}\left[W'_T {W'_T}^{\mathbf{T}}\right] = \mathbb{E}_{W'_{T-1}}\left[W'_{T-1} {W'_{T-1}}^{\mathbf{T}}\right] = \Lambda_{w^*}$ when $T \to \infty$ (i.e. ergodicity), we have

$$
\Lambda_{w^*} H_{w^*} + H_{w^*} \Lambda_{w^*} - \eta H_{w^*} \Lambda_{w^*} H_{w^*} = \eta C_T.
$$

Furthermore, if H_{w^*} and Λ_{w^*} commute, namely $\Lambda_{w^*} H_{w^*} = H_{w^*} \Lambda_{w^*}$, we have

$$
\left[H_{w^*}\left(2\mathrm{I}_d - \eta H_{w^*}\right)\right] \Lambda_{w^*} = \eta C_T,
$$

which will give use $\Lambda_{w^*} = \eta \left[H_{w^*}\left(2\mathrm{I}_d - \eta H_{w^*}\right)\right]^{-1} C_T$.

This completes the proof. □

D.2 Theorem D.1: A General Bound

The following bound can be easily proved by using Eq. 6.

Theorem D.1. *Under the same conditions in Lemma 2.1 and Lemma 4.1, then for any* $P_{W_T} = \mathcal{N}\left(\tilde{w}, \widetilde{\Lambda}\right)$, *where* $\tilde{w}$ *and* $\widetilde{\Lambda}$ *are independent of* S, *we have*

$$
\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{2n} \inf_{\tilde{w}, \widetilde{\Lambda}} \mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\widetilde{\Lambda}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\widetilde{\Lambda}^{-1}\Lambda_{W_S^*} - \mathrm{I}_d\right\} + \mathrm{d}_{\mathrm{M}}^2\left(W_S^*, \tilde{w}; \widetilde{\Lambda}\right)\right]},
$$

where $\mathrm{d}_{\mathrm{M}}(x, y; \Sigma) \triangleq \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$ *is the Mahalanobis distance.*

D.3 Proof of Theorem 4.1

Proof. Let $P_{W_T} = \mathcal{N}\left(w_\mu^*, \Lambda_{w_\mu^*}\right)$, then

$$
\begin{aligned}
&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\Lambda_{w_\mu^*}^{-1}\Lambda_{W_S^*} - \mathrm{I}_d\right\} + \left(W_S^* - w_\mu^*\right)^T \Lambda_{w_\mu^*}^{-1}\left(W_S^* - w_\mu^*\right)\right] \\
=&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\Lambda_{w_\mu^*}^{-1}\Lambda_{W_S^*} - \mathrm{I}_d\right\} + tr\left\{\Lambda_{w_\mu^*}^{-1}\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^T\right\}\right] \\
=&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)}\right] + tr\left\{\Lambda_{w_\mu^*}^{-1}\mathbb{E}_{S, W_S^*}\left[\Lambda_{W_S^*}\right] - \mathrm{I}_d + \Lambda_{w_\mu^*}^{-1}\mathbb{E}_{W_S^*}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^T\right]\right\}.
\end{aligned} \tag{12}
$$

Denote $\widetilde{\Sigma}_\mu \triangleq \mathbb{E}_{S, W_S^*}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^T\right] = \mathbb{E}_{W_S^*}\left[W_S^* {W_S^*}^T\right] - w_\mu^* {w_\mu^*}^T$.

Notice that

$$\begin{aligned}\mathbb{E}_{S,W_S^*}\left[\Lambda_{W_S^*}\right] =&\mathbb{E}_{S,W_S^*,W_T}\left[\left(W_T - W_S^*\right)\left(W_T - W_S^*\right)^T\right]\\ =&\mathbb{E}_{W_T}\left[W_T {W_T}^T\right] - \mathbb{E}_{W_S^*}\left[W_S^* {W_S^*}^T\right]\\ =&\mathbb{E}_{W_T}\left[W_T {W_T}^T\right] - w_\mu^* {w_\mu^*}^T - \left(\mathbb{E}_{W_S^*}\left[W_S^* {W_S^*}^T\right] - w_\mu^* {w_\mu^*}^T\right)\\ =&\Lambda_{w_\mu^*} - \widetilde{\Sigma}_\mu.\end{aligned}$$

Therefore,

$$\begin{aligned}&tr\left\{\Lambda_{w_\mu^*}^{-1}\mathbb{E}_{S,W_S^*}\left[\Lambda_{W_S^*}\right] - \mathrm{I}_d + \Lambda_{w_\mu^*}^{-1}\mathbb{E}_{W_S^*}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^T\right]\right\}\\ =&tr\left\{\Lambda_{w_\mu^*}^{-1}\mathbb{E}_{S,W_S^*}\left[\Lambda_{W_S^*}\right] - \Lambda_{w_\mu^*}^{-1}\Lambda_{w_\mu^*} + \Lambda_{w_\mu^*}^{-1}\mathbb{E}_{W_S^*}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^T\right]\right\}\\ =&tr\left\{\Lambda_{w_\mu^*}^{-1}\left(\mathbb{E}_{S,W_S^*}\left[\Lambda_{W_S^*}\right] - \Lambda_{w_\mu^*} + \widetilde{\Sigma}_\mu\right)\right\}\\ =&0.\end{aligned}$$

Plugging this into Eq. (12), we have

$$\begin{aligned}&\mathbb{E}_{S,W_S^*}\left[\log\frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\Lambda_{w_\mu^*}^{-1}\Lambda_{W_S^*} - \mathrm{I}_d\right\} + \left(W_S^* - w_\mu^*\right)^T\Lambda_{w_\mu^*}^{-1}\left(W_S^* - w_\mu^*\right)\right]\\ =&\mathbb{E}_{S,W_S^*}\left[\log\frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)}\right] = \mathbb{E}_{S,W_S^*}\left[tr\left\{\log\left(\Lambda_{W_S^*}^{-1}\Lambda_{w_\mu^*}\right)\right\}\right].\end{aligned}$$

Finally, applying Theorem D.1 will conclude the proof. □

D.4 Proof of Corollary 4.1

Proof. The proof is straightforward by plugging $\Lambda_{w^*} = \left[H_{w^*}\left(\frac{2}{\eta}\mathrm{I}_d\right)\right]^{-1} C_T$ in Theorem 4.1. □

D.5 Proof of Corollary 4.2

Proof. By Lemma B.2, it's easy to obtain the following bound according to Theorem 4.1.

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2 d}{2n}\log\left(\frac{\mathbb{E}\left[\mathrm{d}_{\mathrm{M}}^2\left(W_S^*, w_\mu^*; \mathbb{E}\left[\Lambda_{W_S^*}\right]\right)\right]}{d} + 1\right) + \mathbb{E}\left[tr\left\{\log\left(\Lambda_{W_S^*}^{-1}\mathbb{E}\left[\Lambda_{W_S^*}\right]\right)\right\}\right]}.$$

Then, plugging $\Lambda_{W_S^*} = \frac{\eta}{2b}\mathrm{I}_d$ will conclude the proof. □

D.6 Corollary D.1: Distance to Initialization

Corollary D.1. *Under (i-iii) in Lemma 4.1, then* $\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n}\log\left(\frac{2b}{\eta d}\mathbb{E}||W_S^* - W_0||^2 + 1\right)}$.

In this case, the corresponding bound suggests that generalization performance can be characterized by the "distance from initialization". Nagarajan and Kolter [32] also derived a "distance from initialization" based generalization bound by using Rademacher complexity, and Hu et al. [21] use "distance from initialization" as a regularizer to improve the generalization performance on noisy data.

Proof. Notice that $I(W_T;S) \leq \mathbb{E}_S \mathrm{D}_{\mathrm{KL}}(Q_{W_T|S}||P_{W_T})$ holds for any $\sigma > 0$, then for a given $\tilde{w}$, we have

$$\begin{aligned}
I(W_T;S) &= \inf_{P_{W_T}} \mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S}||P_{W_T})\right] \\
&\leq \inf_{\sigma} \mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}}(P_{W_S^*+\sqrt{\frac{\eta}{2b}}N,W_S^*|S}||P_{\tilde{w}+\sigma N})\right] \qquad (13) \\
&= \inf_{\sigma} \mathbb{E}_{S,W_S^*} \left[\mathrm{D}_{\mathrm{KL}}(P_{W_S^*+\sqrt{\frac{\eta}{2b}}N,|S,W_S^*}||P_{\tilde{w}+\sigma N})\right] \\
&= \inf_{\sigma} \frac{1}{2}\mathbb{E}_{S,W_S^*} \left[\frac{1}{\sigma^2}(W_S^*-\tilde{w})^T(W_S^*-\tilde{w}) + \log\frac{\sigma^{2d}}{(\eta/2b)^d} + tr\{\frac{\eta}{2b\sigma^2}\mathrm{I}_d\} - d\right] \\
&= \frac{1}{2}\inf_{\sigma} \frac{1}{\sigma^2}\mathbb{E}_{S,W_S^*} \left[||W_S^*-\tilde{w}||^2 + \frac{\eta d}{2b}\right] + d\log\sigma^2 + d\log\frac{2b}{\eta} - d \\
&= \frac{1}{2} d\log\left(\frac{2b}{\eta d}\mathbb{E}_{S,W_S^*}\left[||W_S^*-\tilde{w}||^2\right] + 1\right), \qquad (14)
\end{aligned}$$

where Eq. (13) is by the chain rule of KL divergence, and the optimal $\sigma^* = \sqrt{\mathbb{E}_{S,W_S^*}\left[||W_S^*-\tilde{w}||^2/d + \frac{\eta}{2b}\right]}$. Let $\tilde{w} = W_0$ will conclude the proof. □

Additionally, Corollary D.1 can be used to recover a trajectory-based bound.

Corollary D.2. *Let $W_T = W_s^*$, $\tilde{w} = 0$ and W.L.O.G, assume $W_0 = 0$, then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{dR^2}{n}\log\left(\frac{4bT\eta}{d}\sum_{t=1}^{T}\mathbb{E}\left[||G_t||^2 + tr\{C_t\}\right] + 1\right)},$$

Proof. When $W_0 = 0$, we notice that

$$W_T = \sum_{t=1}^{T} -\eta G_t + \eta N_{C_t},$$

where $N_{C_t} = C_t^{1/2}N_t$.

Thus,

$$||W_T||^2 = ||\sum_{t=1}^{T} -\eta G_t + \eta N_{C_t}||^2 \leq 2T\eta^2 \sum_{t=1}^{T} ||G_t||^2 + ||N_{C_t}||^2$$

Let $\tilde{w} = 0$, recall the bound in Corollary D.1 and plugging the inequality above, we have

$$\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) &\leq \sqrt{\frac{R^2}{n} d\log\left(\frac{2b}{\eta d}\mathbb{E}_{S,W_T}\left[||W_T-\tilde{w}||^2\right] + 1\right)} \\
&\leq \sqrt{\frac{dR^2}{n}\log\left(4bT\eta/d\mathbb{E}_{S,W_{0:T-1},N_{C_{0:t-1}}}\left[\sum_{t=1}^{T}||G_t||^2 + ||N_{C_t}||^2\right] + 1\right)} \\
&= \sqrt{\frac{dR^2}{n}\log\left(\frac{4bT\eta}{d}\sum_{t=1}^{T}\mathbb{E}_{S,W_{t-1}}\left[||G_t||^2 + tr\{C_t\}\right] + 1\right)}
\end{aligned}$$

This concludes the proof. □

D.7 Proof of Theorem E.1

Proof. Let $P_{W_T|S_J=s_j} = \mathcal{N}(W^*_{s_j}, \frac{\eta}{2b}\mathrm{I}_d)$, then

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S=s}||P_{W_T|S_J=s_j}) &= \mathrm{D}_{\mathrm{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N|S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N|S_J=s_j}) \\
&\leq \mathrm{D}_{\mathrm{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N,W^*_s|S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N,W^*_{s_j}|S_J=s_j}) \qquad (15) \\
&= \mathbb{E}_{W^*_s,W^*_{s_j}}\left[\mathrm{D}_{\mathrm{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N|W^*_s,S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N|W^*_{s_j},S_J=s_j})\right] \\
&= \mathbb{E}_{W^*_s,W^*_{s_j}}\left[\frac{b}{\eta}||W^*_s - W^*_{s_j}||^2\right], \qquad (16)
\end{aligned}$$

where Eq. (15) is by the chain rule of KL divergence. Plugging the Eq. (16) into Lemma A.2 will obtain the final result. □

E Additional Result: Data-Dependent Prior

In the sequel, we use the data-dependent prior bound, namely, Lemma A.2, to derive new results.

Let $P_{W_T|S_J=s_j} = \mathcal{N}(W^*_{s_j}, \Lambda(W^*_{s_j}))$ where $W^*_{s_j}$ is the local minimum found by the LOO training.

Theorem E.1. *Under the same conditions in Lemma A.2 and (i-iii) in Lemma 4.1, assuming $\Lambda(W^*_{s_j})$ is close to $\Lambda(W^*_s)$ for a given s, then $\mathcal{E}_\mu(\mathcal{A}) \leq \mathbb{E}_{S,J}\sqrt{\frac{M^2b}{2\eta}\mathbb{E}^{S,J}_{W^*_S,W^*_{S_J}}||W^*_S - W^*_{S_J}||^2}$.*

This bound implies a strong connection between generalization and the algorithmic stability exhibited by SGD. Specifically, if the hypothesis output does not change much (in the squared L_2 distance sense) upon the removal of a single training instance, the algorithm is likely to generalize effectively. In fact, $\mathbb{E}^{S,J}_{W^*_S,W^*_{S_J}}||W^*_S - W^*_{S_J}||^2$ can be regarded as an average version of squared *argument stability* [29]. Moreover, stability-based bounds often demonstrate a fast decay rate in the convex learning cases [17, 7]. It is worth noting that if argument stability achieves the fast rate, e.g., $\sup_{s,j}||w^*_s - w^*_{s_j}|| \leq \mathcal{O}(1/n)$, then Theorem E.1 can also achieve the same rate. In addition, note that the stability-based bound usually contains a Lipshitz constant, while the bound in Theorem E.1 discards such undesired constant.

Ideally, to estimate the distance of $||w^*_s - w^*_{s_j}||^2$, one can utilize the influence function [16, 11, 23], namely $w^*_{s_j} - w^*_s \approx \frac{1}{n}H^{-1}_{W^*_s}\nabla\ell(w^*_s, z_i)$, where i is the instance index that is not selected in j. However, for deep neural network training, the approximation made by influence function is often erroneous [8]. While this presents a challenge, it motivates further exploration and refinement, seeking to enhance the practical application of Theorem E.1 in the context of deep learning.

F Additional Result: Inverse Population FIM as both Posterior and Prior Covariance

Inspired by some previous works of [1, 18, 54], we can also select the inverse population Fisher information matrix $F^\mu_{w^*} = \mathbb{E}_Z\left[\nabla\ell(w^*, Z)\nabla\ell(w^*, Z)^T\right]$ as the posterior covariance. Then, the following theorem is obtained.

Theorem F.1. *Under the same conditions in Theorem E.1, and assume the distribution $P_{W^*_{S_J}|S_J}$ is invariant of J, then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{M}{2n}\mathbb{E}_S\left[\sqrt{\mathbb{E}^S_{W^*_S}\left[tr\{H^{-1}_{W^*_S}F^\mu_{W^*_S}\}\right]}\right].$$

Remark F.1. *Notice that $F^\mu_{W^*_S} \approx H^\mu_{W^*_S} \approx \Sigma^\mu(W^*_S)$ near minima [40, Chapter 8], then $tr\{H^{-1}_{W^*_S}\Sigma^\mu(W^*_S)\}$ is very close to the Takeuchi Information Criterion [50]. In addition, our bound in Theorem F.1 is similar to Singh et al. [48, Theorem 3.] with the same convergence rate, although strictly speaking, their result is not a generalization bound. Moreover, as also pointed out in [48],*

here $H^{-1}_{W^*_S}$ is evaluated on the training sample, unlike other works that evaluates the inverse Hessian on the testing sample (e.g., Thomas et al. [51]).

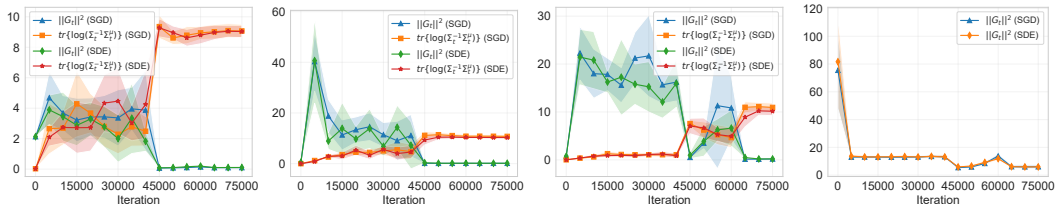
The invariance assumption is also used in Wang et al. [52]. In practice, n is usually very large, when $m = n - 1$, this assumption indicates that replacing one instance in s_j will not make $P_{W^*_{s_j}|s_j}$ be too different.

G Proof of Theorem F.1

Proof. We now use $(F^{\mu}_{W^*_S})^{-1}$ as both the posterior and prior covariance (again, we assume $F^{\mu}_{W^*_S} \approx F^{\mu}_{W^*_{S_j}}$ for any j), then

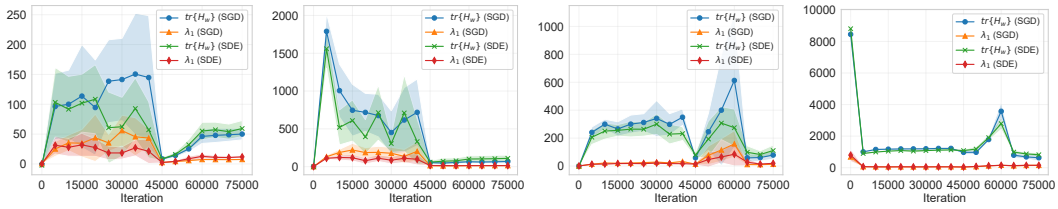
$$\begin{aligned}
\mathcal{E}_{\mu}(\mathcal{A}) \leq & \mathbb{E}_S \left[\sqrt{\frac{M^2}{4} \mathbb{E}^S_{J, W^*_S, W^*_{S_J}} \left[\left(W^*_S - W^*_{S_J} \right) F^{\mu}_{W^*_S} \left(W^*_S - W^*_{S_J} \right)^T \right]} \right] \\
= & \frac{M}{2n} \mathbb{E}_S \left[\sqrt{\mathbb{E}^S_{W^*_S, W^*_{S_j}} \left[tr \left\{ F^{\mu}_{W^*_S} H^{-1}_{W^*_S} H^{-1}_{W^*_S} \mathbb{E}_J \left[\nabla \ell(W^*_S, Z_i) \nabla \ell(W^*_S, Z_i)^T \right] \right\} \right]} \right] \\
= & \frac{M}{2n} \mathbb{E}_S \left[\sqrt{\mathbb{E}^S_{W^*_S} \left[tr \left\{ F^{\mu}_{W^*_S} H^{-1}_{W^*_S} \right\} \right]} \right],
\end{aligned}$$

which completes the proof. □



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10 (d) ResNet on CIFAR100

Figure 2: Gradient-related quantities of SGD or its discrete SDE approximation. In (d), since per-sample gradient is ill-defined when BatchNormalization is used, we do not track $tr\left\{\log\left(\Sigma_t^{-1}\Sigma_t^{\mu}\right)\right\}$.



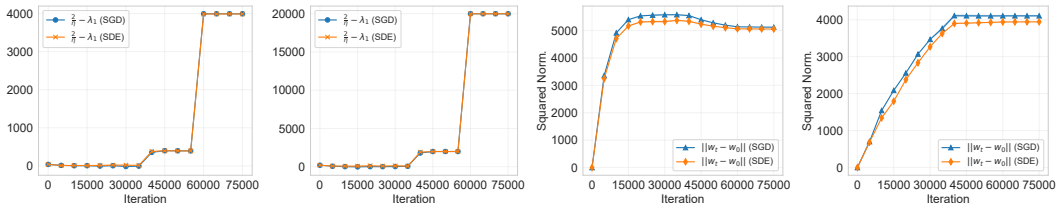
(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10 (d) ResNet on CIFAR100

Figure 3: Hessian-related quantities of SGD or its discrete SDE approximation.

H Empirical Study

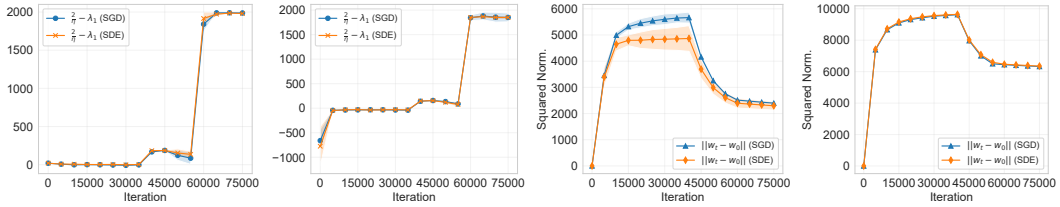
In this section, we present some empirical results including tracking training dynamics of SGD and SDE, along with the estimation of several obtained generalization bounds.

SGD and SDE Training Dynamics We implement the SDE training by following the same algorithm given in [56, Algorithm 1]. Our experiments involved training a VGG-11 architecture without BatchNormalization on a subset of SVHN (containing 25k training images) and CIFAR10.



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) VGG on (small) SVHN (d) VGG on CIFAR10

Figure 4: (a-b) The dynamics of $\eta/2 - \lambda_1$. Note that learning rate decays by 0.1 at the $40,000^{\text{th}}$ and the $60,000^{\text{th}}$ iteration. (c-d) The distance of current model parameters from its initialization.

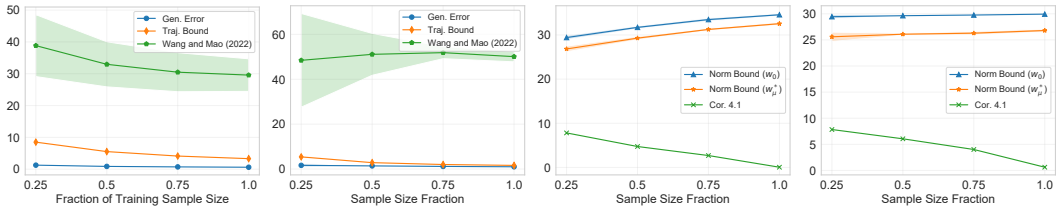


(a) ResNet on CIFAR10 (b) ResNet on CIFAR100 (c) ResNet on CIFAR10 (d) ResNet on CIFAR100

Figure 5: (a-b) The dynamics of $\eta/2 - \lambda_1$. Note that learning rate decays by 0.1 at the $40,000^{\text{th}}$ and the $60,000^{\text{th}}$ iteration. (c-d) The distance of current model parameters from its initialization.

Additionally, we trained a ResNet-18 on both CIFAR10 and CIFAR100. Data augmentation is only used in the experiments related to CIFAR100. We ran each experiment for ten different random seed, maintaining a fixed initialization of the model parameters. Further details about the experimental setup can be found in the Appendix. The results are depicted in Figure 1. As mentioned earlier, SDE exhibits a performance dynamics akin to that of SGD, reinforcing the similarities in their training behaviors.

Evolution of Key Quantities for SGD and SDE We show $||G_t||^2$ and $tr\left\{\log\left(\Sigma_t^{-1}\Sigma_\mu\right)\right\}$ in Figure 2. Recognizing the computational challenges associated with computing $tr\left\{\log\left(\Sigma_t^{-1}\Sigma_\mu\right)\right\}$, we opted to draw estimates based on 100 training and 100 testing samples. Notably, both SGD and SDE exhibit similar behaviors in these gradient-based metrics. It is noteworthy that despite the absence of the learning rate in the trajectories-based bounds, we observed that modifications to the learning rate at the $40,000^{\text{th}}$ and $60,000^{\text{th}}$ steps had discernible effects on these gradient-based quantities. Additionally, in Figure 3, we examine the trace of the Hessian and its largest eigenvalue during training, leveraging the PyHessian library [62]. Note that we still use only 100 training data to estimate the Hession for efficiency. Notice that the Hessian-related quantities of SGD and SDE are nearly perfectly matched in the terminal state of training. Furthermore, Figures 4c-4d illustrate the "distance to initialization," revealing a consistent trend shared by both SGD and SDE.



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) VGG on (small) SVHN (d) VGG on CIFAR10

Figure 6: Estimated trajectories-based bound and terminal-state based bound, with R excluded. Zoomed-in figures of generalization error are given in Figure 7 in Appendix.

Bound Comparison We vary the size of the training sample and empirically estimate several of our bounds in Figure 6, with the subgaussian variance proxy R excluded for simplicity. Thus,

the estimated values in Figure 6 don't accurately represent the true order of the bounds. Despite the general unbounded nature of cross-entropy loss, common training strategies, such as proper weight initialization, training techniques, and appropriate learning rate selection, ensure that the cross-entropy loss remains bounded in practice. Therefore, it is reasonable to assume subgaussian behavior of the cross-entropy loss under SGD training. In Figure 6a-6b, we compare our Theorem 3.1 with Wang and Mao [55, Theorem 2]. Since both bounds incorporate the same R, the results in Figures6a to 6b show that our Theorem 3.1 outperforms Wang and Mao [55, Theorem2]. This aligns with expectations, considering that the isotropic Gaussian used in the auxiliary weight process of Wang and Mao [55, Theorem2] is suboptimal. Moreover, Figures 6c to 6d hint that norm-based bounds Corollary 4.2 and Corollary D.1) exhibit growth with n, which are also observed in [33]. In contrast, Corollary 4.1 effectively captures the trend of generalization error, emphasizing the significance of the geometric properties of local minima. Additionally, while trajectories-based bounds may appear tighter, terminal-state-based bounds seem to have a faster decay rate.

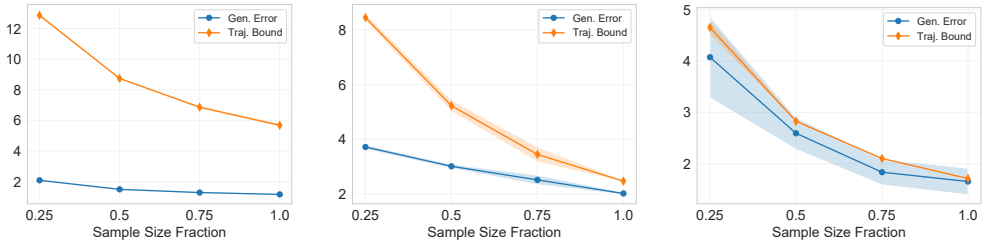
I Experiment Details and Additional Results

The implementation in this paper is on PyTorch [39], and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB). Most experiment settings follow [56], and the code is also based their implementation, which is available at: https://github.com/uuujf/MultiNoise.

I.1 Hyperparameters

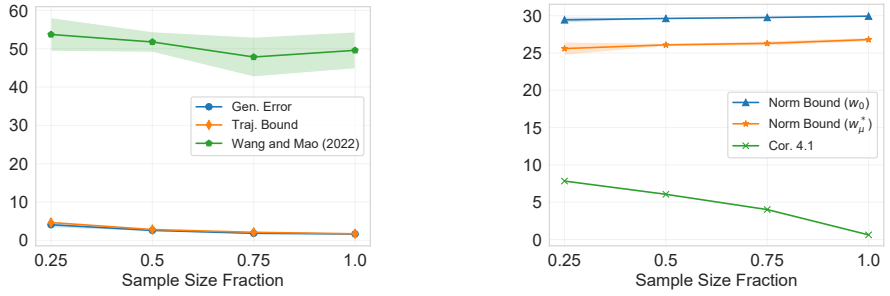
For CIFAR 10, the initial learning rates used for VGG-11 and ResNet-18 are 0.01 and 0.1, respectively. For SVHN, the initial learning rate is 0.05. For CIFAR100, the initial learning rate is 0.1. The learning rate is then decayed by 0.1 at iteration $40,000$ and $60,000$. If not stated otherwise, the batch size of SGD is 100.

I.2 Additional Results



(a) VGG on (small) SVHN (b) VGG on CIFAR10 (c) ResNet on CIFAR10

Figure 7: Zoomed-in of generalization error.



(a) ResNet (Traj. Bound) (b) ResNet (Term. Bound)

Figure 8: Estimated trajectories-based bound and terminal-state based bound, with R excluded. Models trained on CIFAR 10.