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

Ziqiao Wang[1] **Yongyi Mao**[1]

[1]School of Electrical Engineering and Computer Science,
University of Ottawa,
Ottawa, Ontario, Canada

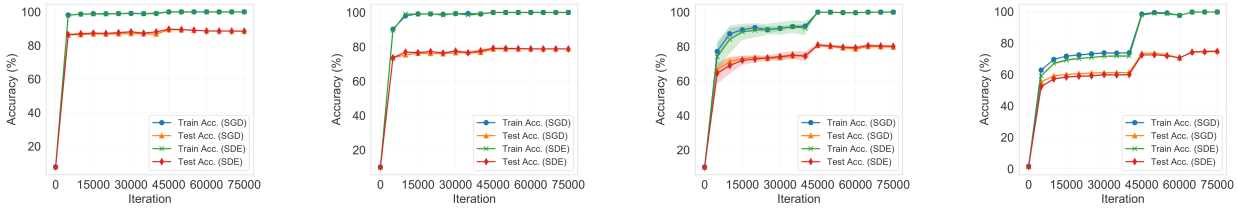
Abstract

Stochastic differential equations (SDEs) have been shown recently to characterize well the dynamics of training machine learning models with SGD. When the generalization error of the SDE approximation closely aligns with that of SGD in expectation, it provides two opportunities for understanding better 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 trajectory-based generalization bound using the information-theoretic bound from Xu and Raginsky [2017]. Secondly, assuming mild conditions, we estimate the steady-state weight distribution of SDE and use information-theoretic bounds from Xu and Raginsky [2017] and Negrea et al. [2019] to establish terminal-state-based generalization bounds. Our proposed bounds have some advantages, notably the trajectory-based bound outperforms results in Wang and Mao [2022], and the terminal-state-based bound exhibits a fast decay rate comparable to stability-based bounds.

1 INTRODUCTION

Modern deep neural networks trained with SGD and its variants have achieved surprising successes: the over-parametrized networks often contain more parameters than the size of training dataset, and yet are capable of generalizing well on the testing set; this contrasts the traditional wisdom in statistical learning that suggests such high-capacity models will overfit the training data and fail on the unseen data [Zhang et al., 2017]. Intense recent efforts have been spent to explain this peculiar phenomenon via investigating the properties of SGD [Arpit et al., 2017, Bartlett et al., 2017, Neyshabur et al., 2017, Arora et al., 2019], and the current understanding is still far from being complete. For example, neural tangent kernel (NTK)-based generalization bounds of SGD normally require the width of network to be sufficiently large (or even go to infinite) [Arora et al., 2019], and the stability-based bounds of SGD have a poorly dependence on an intractable Lipschitz constant [Hardt et al., 2016, Bassily et al., 2020].

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. The concept of mutual information (MI)-based bounds can be traced back to [Catoni, 2007, Section 1.3.1], where it is discussed how the prior distribution that minimizes the expected KL complexity term in a PAC-Bayesian (PAC-Bayes) bound can transform the KL term into the MI term. This idea has recently been popularized by [Russo and Zou, 2016, 2019, Xu and Raginsky, 2017], and further strengthened by additional techniques [Asadi et al., 2018, Negrea et al., 2019, Bu et al., 2019, Steinke and Zakynthinou, 2020, Haghifam et al., 2020, Wang et al., 2021b]. Particularly, Negrea et al. [2019] 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) [Raginsky et al., 2017, Pensia et al., 2018], so as to avoid the MI bounds becoming infinity, and can not be directly applied to SGD. While it is feasible to adopt similar techniques from the PAC-Bayes bounds in Lotfi et al. [2022] to analyze SGD, in order to study the generalization of SGD using methods akin to those for SGLD, Neu et al. [2021] and



(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.

Wang and Mao [2022] develop generalization bounds for SGD via constructing an auxiliary iterative noisy process. However, identifying an optimal auxiliary process is difficult, and arbitrary choices may not provide meaningful insights into the generalization of SGD, see Appendix A.1 for more discussions.

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, Mandt et al. [2017] and Jastrzębski et al. [2017] model this gradient noise drawn from a Gaussian distribution with a fixed covariance matrix, thereby viewing SGD as performing variational inference. Zhu et al. [2019], Wu et al. [2020], Xie et al. [2021a], and Xie et al. [2021b] further model the gradient noise as dependent of the current weight parameter and the training data. Moreover, Li et al. [2017, 2019] and Wu et al. [2020] prove that when the learning rate is sufficiently small, the SDE trajectories are theoretically close to those of SGD (cf. Lemma A.1). More recently, Li et al. [2021] 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 (cf. 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.

- We obtain a generalization bound (cf. Theorem 3.1) in the form of a summation over training steps of a quantity that involves both the sensitivity of the full-batch gradient to the variation of the training set and the covariance of the gradient noise (which makes the SGD gradient deviate from the full-batch gradient). We also give a tighter bound in Theorem 3.2, where the generalization performance of SGD depends on the alignment of the population gradient covariance and the batch gradient covariance. These bounds highlight the significance of (the trace of) the gradient noise covariance in the generalization ability of SGD.

- In addition to the time-dependent trajectory-based bounds, we also provide time-independent (or asymptotic) bounds by some mild assumptions. Specifically, based on previous information-theoretic bounds, we obtain generalization bounds in terms of the KL divergence between the steady-state weight distribution of SGD with respect to a distribution-dependent prior distribution (by Lemma 2.1) or data-dependent prior distribution (by Lemma 2.2). The former 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 (cf. Theorem 4.1). Under mild assumptions, we can estimate the steady-state weight distribution of SDE (cf. Lemma 4.1), leading to a variant of Theorem 4.1 (cf. Corollary 4.1) and a norm-based bound (cf. Corollary 4.2). Additionally, we obtain a stability-based like bound by Lemma 2.2 (cf. Theorem 4.2), with the notable omission of the Lipschitz constant in other stability-based bounds. Since stability-based bounds often achieve fast decay rates, e.g., $\mathcal{O}(1/n)$, Theorem 4.2 provides theoretical advantages compared with other information-theoretic bounds, as it can attain the same rate of decay as the stability-based bound. Comparing to the first family of bounds (i.e., trajectory-based bounds), the second family of bounds directly upper-bound the generalization error via the terminal state, which avoids summing over training steps; these bounds can be tighter when the steady-state estimates are accurate. On the other hand, not relying on the steady-state estimates and the approximating assumptions they base upon is arguably an advantage of the first family.

- We empirically analyze key components within the derived bounds for both algorithms. Our empirical findings reveal that these components for SGD and SDE align remarkably well, further validating the effectiveness of our bounds for assessing the generalization of SGD. Moreover, we provide numerical validation of the presented bounds and demonstrate that our trajectory-based bound

is tighter than the result in Wang and Mao [2022]. Additionally, compared with norm-based bounds, we show that the terminal-state-based bound that integrates the geometric properties of local minima can better characterize generalization.

2 PRELIMINARY

Notation 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 the 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 $S = \{Z_i\}_{i=1}^n$ 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 addition, 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 $\widetilde{G}_t \triangleq \frac{1}{b}\sum_{i\in B_t} \nabla\ell(w_{t-1},Z_i)$ 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 $\Sigma_t = \frac{1}{n}\sum_{i=1}^n \nabla\ell_i\nabla\ell_i^{\mathbf{T}} - G_tG_t^{\mathbf{T}}$ 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

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

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 is the SDE in Eq. (4) for approximating the SGD process in Eq. (1). Previous research, such as [Li et al., 2017, 2019], 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. [2020, Theorem 2] has proved that Eq. (5) is *an order* 1 *strong approximation* to SDE in Eq. (4). Moreover, we direct interested readers to the comprehensive investigations carried out by [Wu et al., 2020, Li et al., 2021], where the authors empirically verify that SGD and Eq. (5) can achieve the similar testing performance in the deep learning scenarios, suggesting that non-Gaussian noise may not be 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.

Two Information-Theoretic Bounds The original information-theoretic bound in Xu and Raginsky [2017] 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 [2017, Theorem 1.]). *Assume the loss $\ell(w, Z)$ is R-subgaussian*[1] *for any $w \in \mathcal{W}$, then*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} I(W;S)}.$$

This bound is further refined to a data-dependent prior based bound. Following the setup in Negrea et al. [2019], 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 2.2 (Negrea et al. [2019, Theorem 2.5]). *Let $Q_{W|S}$ be the posterior distribution of W given the training sample S, and let $P_{W|S_J}$ be the posterior distribution of W given the training sample S_J. Assume the loss $\ell(w, Z)$ is bounded in $[0, M]$, then for any $P_{W|S_J}$,*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{M}{\sqrt{2}} \mathbb{E}_{S,J} \sqrt{\mathrm{D_{KL}}(Q_{W|S} || P_{W|S_J})}.$$

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 on the same algorithm $\mathcal{A}$ (e.g. SGD) 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.

3 GENERALIZATION BOUNDS VIA FULL TRAJECTORIES

We now discuss the generalization of SGD under the approximation of Eq. (5). We first unroll the terminal parameters' mutual information $I(W_T; S)$ to the full trajectories' mutual information via the lemma below.

Lemma 3.1. *The MI term in Lemma 2.1 is upper bounded by $I(W_T; S) \leq \sum_{t=1}^T I(-G_t + C_t^{1/2} N_t; S | W_{t-1})$.*

This lemma can be proved by recurrently applying the data processing inequality (DPI) and chain rule of the mutual information [Polyanskiy and Wu, 2019].

Define $\widehat{G}_t = -G_t + C_t^{1/2} N_t$. Let $Q_{\hat{G}_t|s,w_{t-1}}$ and $Q_{\hat{G}_t|w_{t-1}}$ be the conditional and marginal distributions fully characterized by the algorithm, respectively. In addition, let $P_{\widehat{G}_t|w_{t-1}}$ be any prior distribution of $\widehat{G}_t$, satisfying $\mathrm{D_{KL}}(Q_{\widehat{G}_t|s,w_{t-1}} || P_{\widehat{G}_t|w_{t-1}}) < \infty$. we first have the following lemma.

Lemma 3.2. *For any time step t, we have $I(\widehat{G}_t; S | W_{t-1}) = \mathbb{E}_{W_{t-1}} \left[\inf_{P_{\widehat{G}_t|W_{t-1}}} \mathbb{E}_S^{W_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|S,W_{t-1}} || P_{\widehat{G}_t|W_{t-1}}) \right] \right]$, where the infimum is achieved when the prior distribution $P_{\widehat{G}_t|w_{t-1}} = Q_{\widehat{G}_t|w_{t-1}}$.*

Lemma 3.2 suggests that every choice of the prior $P_{\widehat{G}_t|W_{t-1}}$ gives rise to an upper bound of the MI of interest via $I(\widehat{G}_t; S | W_{t-1}) \leq \mathbb{E}\left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|S,W_{t-1}} || P_{\widehat{G}_t|W_{t-1}}) \right]$. The closer is $P_{\widehat{G}_t|W_{t-1}}$ to $Q_{\widehat{G}_t|W_{t-1}}$, the tighter is the bound. As the simplest choice, we will first choose an isotropic Gaussian prior, $P_{\widehat{G}_t|w_{t-1}} = \mathcal{N}(\tilde{g}_t, \sigma_t^2 \mathrm{I}_d)$ (where both $\tilde{g}_t$ and σ_t are only allowed to depend on W_{t-1}), and optimize the KL divergence in Lemma 3.2 over σ_t for a fixed $\tilde{g}_t$. The following result is obtained.

Theorem 3.1. *Under the conditions of Lemma 2.1 and assume C_t is a positive-definite matrix. For any $t \in [T]$, let $\tilde{g}_t$ be any constant vector for a given w_{t-1}, then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^T \mathbb{E}_{W_{t-1}} \left[d \log \frac{h_1(W_{t-1})}{d} - h_2(W_{t-1}) \right]}, \tag{6}$$

where $h_1(w) = \mathbb{E}_S^w \left[||G_t - \tilde{g}_t||^2 + tr\{C_t\} \right]$ and $h_2(w) = \mathbb{E}_S^w [tr\{\log C_t\}]$.

Furthermore, if $\tilde{g}_t = \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)]$, then $h_1(w) = \frac{1}{b} tr\{\Sigma_t^\mu\}$.

Notice that $\tilde{g}_t$ is any reference "gradient" independent of S, then the first term in $h_1(W_{t-1})$, $||G_t - \tilde{g}_t||^2$, characterizes the sensitivity of the full-batch gradient to some variation of the training set S, while the second term in $h_1(W_{t-1})$, i.e. $tr\{C_t\}$, reflects the gradient noise magnitude induced by the mini-batch based training. For example, if $\tilde{g}_t = \mathbb{E}_Z [\nabla \ell(w_{t-1}, Z)]$, then $\mathbb{E}_S^{w_{t-1}} \left[||G_t - \tilde{g}_t||^2 \right]$ is the variance of the gradient sample mean, and such $\tilde{g}_t$ will eventually convert $h_1(W_{t-1})$ to the population GNC, namely $h_1(W_{t-1}) = \frac{1}{b} tr\{\Sigma_t^\mu\}$.

Moreover, if we simply let $\tilde{g}_t = 0$, then Theorem 3.1 indicates that one can control the generalization performance via controlling the gradient norm along the entire training trajectories, e.g., if we further let $b = 1$, then $h_1(W_{t-1}) = \frac{1}{n} \sum_{i=1}^n ||\nabla \ell_i||^2$. This is consistent with the existing practice, for example, applying gradient clipping [Wang and Mao, 2022, Geiping et al., 2022] and gradient penalty [Jastrzebski et al., 2021, Barrett and Dherin, 2021, Smith et al., 2021, Geiping et al., 2022] as regularization techniques to improve generalization.

[1] A random variable X is R-subgaussian if for any $\rho \in \mathbb{R}$, $\log \mathbb{E} \exp(\rho(X - \mathbb{E}X)) \leq \rho^2 R^2 / 2$. Note that a bounded loss is guaranteed to be subgaussian.

As a by-product, we recover previous information-theoretic bounds for the Gradient Langevin dynamics (GLD) with noise distribution $\mathcal{N}(0, \eta^2 \mathrm{I}_d)$ below.

Corollary 3.1. *If $C_t = \mathrm{I}_d$, then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \sqrt{\frac{R^2 d}{n} \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}} \log \left(\mathbb{E}_S^{W_{t-1}} ||G_t - \tilde{g}_t||^2 / d + 1 \right)}.$$

Note that the bound in Corollary 3.1 can recover the bound in Neu et al. [2021, Proposition 3.] by using the inequality $\log(x+1) \leq x$. Furthermore, it can also recover the bound in Pensia et al. [2018] because we use a state-dependent quantity $\mathbb{E}_S^{w_{t-1}} \left[||G_t - \tilde{g}_t||^2 \right]$, which is smaller than the global Lipschitz constant used in Pensia et al. [2018].

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. Let $\tilde{g}_t = \mathbb{E}_Z \left[\nabla \ell(w_{t-1}, Z) \right]$ be the state-dependent mean. By optimizing over c_t, we have the bound below.

Theorem 3.2. *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 3.2 implies that a favorable alignment between the diagonal values of Σ_t and Σ_t^μ will positively impact generalization performance. In other words, the perfect alignment of these two matrices indicates that SGD is insensitive to the randomness of S. Recall the key quantity in Lemma 2.1, $I(W; S)$, which also measures the dependence of W with the randomness of S, the term $\Sigma_t^\mu \Sigma_t^{-1}$ conveys a similar intuition in this context.*

Compared with Theorem 3.1 under the same choice of $\tilde{g}_t$, we notice that the main difference is that the term $tr \left\{ \log(\Sigma_t^\mu / b) \right\}$, instead of $d \log(tr \left\{ \Sigma_t^\mu \right\} / bd)$, appears in the bound of Theorem 3.2. The following lemma demonstrates that Theorem 3.2 is tighter than the bound in Theorem 3.1.

Lemma 3.3. *For any t, we have $tr \left\{ \log \frac{\Sigma_t^\mu}{b} \right\} \leq d \log \frac{tr\{\Sigma_t^\mu\}}{bd}$, with the equality holds when all the diagonal elements in Σ_t^μ have the same value, i.e. $\alpha_t(1) = \alpha_t(2) = \cdots = \alpha_t(d)$.*

The trajectory-based bounds in Theorem 3.1 and Theorem 3.2 emphasize the significance of gradient-related information along entire trajectories, including metrics such as gradient norm and gradient covariance alignment, in comprehending the generalization dynamics of understanding the generalization of SGD. In Figure 2, we visually show that these key gradient-based measures during SDE training closely mirror the dynamics observed in SGD.

Notably, these trajectory-based information-theoretic bounds are time-dependent, indicating that these bounds may grow with the training iteration number T, unless the gradient norm becomes negligible at some point during training. While the stability-based bounds for GD/SGD are also time-dependent [Hardt et al., 2016, Bassily et al., 2020] (in the convex learning case), the learning rate in these bounds helps mitigate the growth of T. However, the learning rate does not appear in our trajectory-based information-theoretic bounds, making the dependency on T even worse.

Note that Wang et al. [2021b] uses the strong data processing inequality to reduce this deficiency, but the bound still increases with T. To tackle this weakness, we will invoke some asymptotic SDE results on the terminal parameters of the algorithm, which will give us a crisp way to characterize the expected generalization gap without decomposing the mutual information.

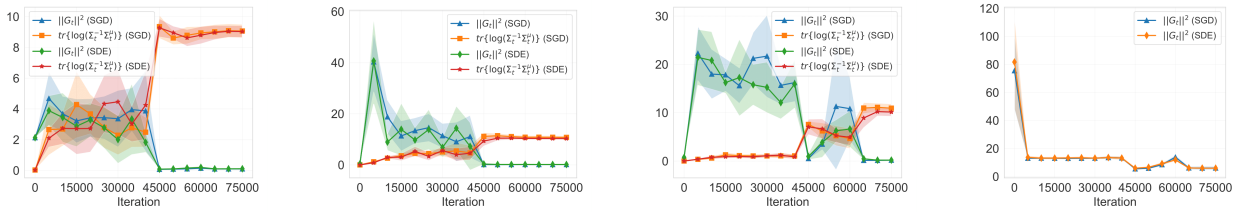
4 GENERALIZATION BOUNDS VIA TERMINAL STATE

In this section, we directly upper-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 $Q_{W_T|S}$. To overcome the explicit time-dependence present in the bounds discussed in Section 3, one has to introduce additional assumptions, with these assumptions being the inherent cost. For example, an important approximation used in this section is the quadratic approximation of the loss. Specifically, let w_s^* be a local minimum for a given training sample $S = s$, when w is close to w_s^*, we can use a second-order Taylor expansion to approximate the value of the loss at w,

$$L_s(w) = L_s(w_s^*) + \frac{1}{2}(w - w_s^*)^{\mathbf{T}} H_{w_s^*}(w - w_s^*). \tag{7}$$

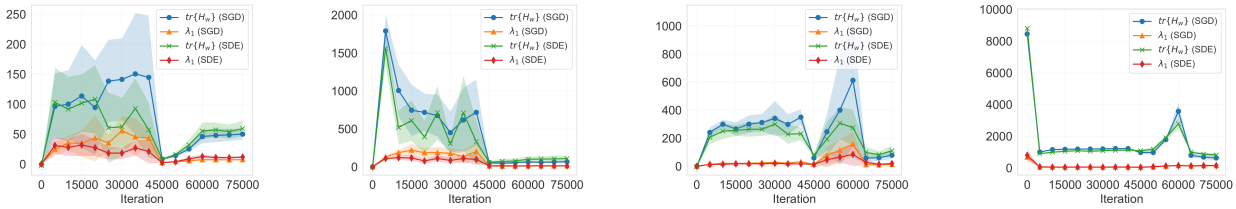
where $H_{w_s^*}$ is the Hessian matrix of s at w^*. Note that in this case, when $w_t \to w_s^*$, we have $G_t = \nabla L_s(w_t) = H_{w_s^*}(w_t - w_s^*)$. Our remaining analysis assumes the validity of Eq. (7).

In view of Eq. (7), a classical result by Mandt et al. [2017] shows that the posterior distribution $Q_{W|s}$ around w_s^* is a Gaussian distribution $\mathcal{N}(w_s^*, \Lambda_{w_s^*})$, where $\Lambda_{w^*} \triangleq \mathbb{E}\left[(W - w^*)(W - w^*)^{\mathbf{T}}\right]$ is the covariance of the station-



(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.

ary distribution (see Appendix A.3 for an elaboration). Furthermore, in the context of nonconvex learning, such as deep learning, where multiple local minima exist, we have multiple w_s^* for a give $S = s$. Therefore, it is necessary to treat the local minimum itself as a random variable for a fixed s, denoted as $W_s^* \sim Q_{W_s^*|s}$. In this case, we have $Q_{W|s,w_s^*} = \mathcal{N}(w_s^*, \Lambda_{w_s^*})$ and the posterior distribution $Q_{W|s} = \mathbb{E}_{W_s^*}\left[\mathcal{N}(W_s^*, \Lambda_{W_s^*})\right]$ should be a mixture of Gaussian distributions.

In addition, recall that $I(W_T; S) = \inf_{P_{W_T}} \mathbb{E}_S \mathrm{D}_{\mathrm{KL}}(Q_{W_T|S}||P_{W_T})$ where $P_{W_T} = Q_{W_T}$ achieves the infimum. Here, the oracle prior $Q_{W_T} = \mathbb{E}_{S,W_S^*}\left[\mathcal{N}(W_S^*, \Lambda_{W_S^*})\right]$ is also a mixture of Gaussian distributions. From a technical standpoint, given that the KL divergence between two mixtures of Gaussian distributions does not have a closed-form expression, we turn to analyze its upper bound, namely $\inf_{P_{W_T}} \mathbb{E}_{S,W_S^*} \mathrm{D}_{\mathrm{KL}}(Q_{W_T|S,W_S^*}||P_{W_T})$. When each s has only one local minimum, $I(W; S)$ reaches this upper bound.

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

Theorem 4.1. *Let $w_\mu^* = \mathbb{E}\left[W_S^*\right]$ be the expected ERM solution and let $\Lambda_{w_\mu^*} = \mathbb{E}\left[\left(W_T - w_\mu^*\right)\left(W_T - w_\mu^*\right)^{\mathbf{T}}\right]$ be its corresponding stationary covariance, then*

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

This result bears resemblance to Theorem 3.2 since both involve the alignment between a population covariance matrix and a sample (or batch) covariance matrix.

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 move 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 $\mathrm{d}_{\mathrm{M}}(x, y; \Sigma) \triangleq \sqrt{(x-y)^{\mathbf{T}}\Sigma^{-1}(x-y)}$ 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. This is still much easier than estimating $I(W; S)$ directly. As an alternative strategy, one may leverage the analytical expression available for $\Lambda_{w_s^*}$. Mandt et al. [2017] provides such analysis and give a equation to solve for $\Lambda_{w_s^*}$. However, the result in Mandt et al. [2017] relies on the unrealistic small learning rate, and the GNC in their analysis is regarded as a state-independent covariance matrix. To overcome these limitations, we give the following result under a quadratic approximation of the loss, which is refined from Liu et al. [2021, Theorem 1] by using the state-dependent GNC.

Lemma 4.1. *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 following conditions:*

(i) $H_{w^*}\Lambda_{w^*} = \Lambda_{w^*}H_{w^*}$;

(ii) $H_{w^*}^{-1}\Sigma_T = \mathrm{I}_d$;

(iii) $\frac{2}{\eta} \gg \lambda_1$, *where* λ_1 *is the top-*1 *eigenvalue of* H_{w^*}.

Then, given (i), we have $\Lambda_{w^*} = \left[H_{w^*}\left(\frac{2}{\eta}\mathrm{I}_d - H_{w^*}\right)\right]^{-1} C_T$; *given (i-ii), we have* $\Lambda_{w^*} = (\frac{2}{\eta}\mathrm{I}_d - H_{w^*})^{-1}$; *given (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 [Jastrzębski et al., 2017, Zhu et al., 2019, Li et al., 2020, Xie et al., 2021a,b, Liu et al., 2021], Hessian is proportional to the GNC near local minima when the loss is the negative log likelihood, i.e. cross-entropy loss. To see this, when $w_t \to w^*$, we have $\Sigma_{w^*} = \frac{1}{n}\sum_{i=1}^n \nabla\ell_i\nabla\ell_i^{\mathbf{T}} - G_t G_t^{\mathbf{T}} \approx \frac{1}{n}\sum_{i=1}^n \nabla\ell_i\nabla\ell_i^{\mathbf{T}} = F_{w^*}$, where F_{w^*} is the *Fisher information matrix* (FIM). This approximation is true because gradient noise dominates over gradient mean near local minima. Moreover, FIM is close to the Hessian near local minima with the log-loss [Pawitan, 2001, Chapter 8], namely, $F_{w^*} \approx H_{w^*}$. Let $n \gg b$, we have $H_{w^*} \approx \Sigma_{w^*} = bC_{w^*}$. Consequently, when Σ_T is sufficiently close to Σ_{w^*}, condition (ii) is satisfied. It's important to note that the debate surrounding $H_{w^*} \approx F_{w^*}$ arises when the loss function deviates from cross-entropy [Ziyin et al., 2022].

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* [Wu et al., 2018, Cohen et al., 2021, Arora et al., 2022], 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 the context of Theorem 4.1, 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" regime, the bound presented in Theorem 4.1 diverges to infinity. This, as a by-product, provides an alternative explanation to the failure mode of $I(W;S) \to \infty$ in the deterministic algorithm (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 [Bartlett et al., 2017, Neyshabur et al., 2018]. 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", namely, given that SGD achieves satisfactory performance on the training data, a shorter distance from the initialization tends to yield better generalization. Nagarajan and Kolter [2019a] also derived a "distance from initialization" based generalization bound by using Rademacher complexity, and Hu et al. [2020] use "distance from initialization" as a regularizer to improve the generalization performance on noisy data.

In the sequel, we use the data-dependent prior bound, namely, Lemma 2.2, to derive new results.

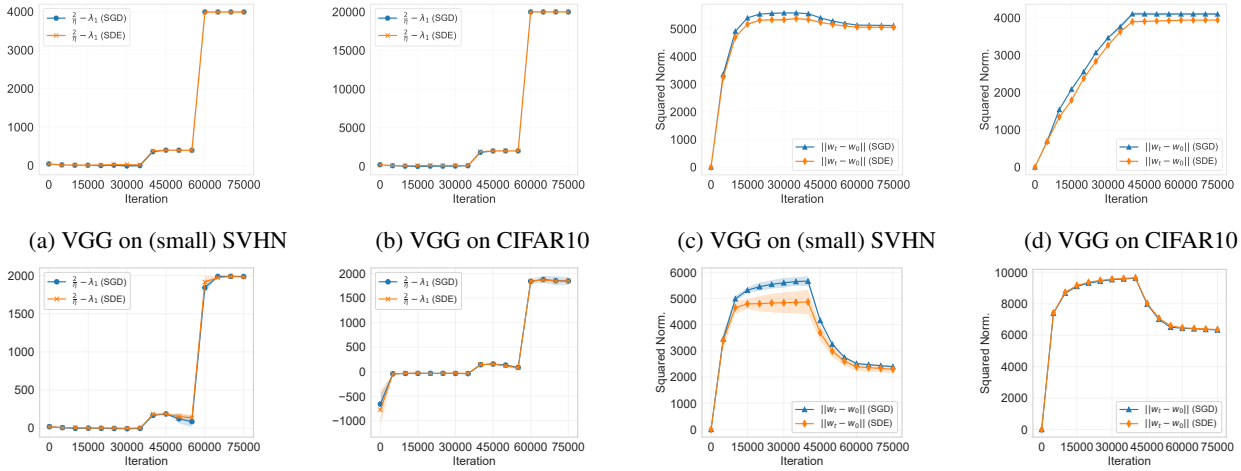
Theorem 4.2. *Let* $P_{W_T|S_J=s_j} = \mathcal{N}(W_{s_j}^*, \Lambda(W_{s_j}^*))$ *where* $W_{s_j}^*$ *is the local minimum found by the LOO training. Under the same conditions in Lemma 2.2 and (i-iii) in Lemma 4.1, assuming* $\Lambda(W_{s_j}^*) = \Lambda(W_s^*)$ *for a given* s, *then*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \mathbb{E}_{S,J}\sqrt{\frac{M^2 b}{2\eta}\mathbb{E}^{S,J}_{W_S^*,W_{S_J}^*}||W_S^* - W_{S_J}^*||^2}.$$

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* [Liu et al., 2017]. Moreover, stability-based bounds often demonstrate a fast decay rate in the convex learning cases [Hardt et al., 2016, Bassily et al., 2020]. 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 4.2 can also achieve the same rate. In addition, note that the stability-based bound usually contains a Lipshitz constant, while the bound in Theorem 4.2 discards such undesired constant.

Ideally, to estimate the distance of $||w_s^* - w_{s_j}^*||^2$, one can use the influence function [Hampel, 1974, Cook and Weisberg, 1982, Koh and Liang, 2017], namely $w_{s_j}^* - w_s^* \approx \frac{1}{n}H_{W_s^*}^{-1}\nabla\ell(w_s^*, z_i)$, 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 [Basu et al., 2021]. While this presents a challenge, it motivates further refinement, seeking to enhance the practical application of Theorem 4.2 in deep learning.

The main generalization bounds obtained in this paper are summarized in Table 1. In the remainder of this paper, we



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

(e) ResNet on CIFAR10 (f) ResNet on CIFAR100 (g) ResNet on CIFAR10 (h) ResNet on CIFAR100

Figure 4: (a-b,e-f) 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,g-h) The distance of current model parameters from its initialization.

Table 1: Comparison of the results in this work

	Bounds	Remarks
Trajectory-based Bounds. Pros: less assumptions, can track training dynamics; Cro: Time-Dependent		
Theorem 3.1	$\mathcal{O}\left(\sqrt{\frac{d}{n}\mathbb{E}\left[\log\frac{h_1}{d} - \frac{h_2}{d}\right]}\right)$	Isotropic covariance for Gaussian prior
Corollary 3.1	$\mathcal{O}\left(\sqrt{\frac{d}{n}\sum_{t=1}^{T}\mathbb{E}\log\left(\frac{\mathbb{E}\lvert\lvert G_t-\tilde{g}_t\rvert\rvert^2}{d}+1\right)}\right)$	Bound for langevin dynamic; tighter than Neu et al. [2021, Prop. 3.]
Theorem 3.2	$\mathcal{O}\left(\sqrt{\frac{1}{n}\sum_{t=1}^{T}\mathbb{E}\left[tr\left\{\log\frac{\Sigma_t^{\mu}C_t^{-1}}{b}\right\}\right]}\right)$	Population GNC for prior; tighter than Thm. 3.1
Terminal-State-based Bounds. Pro: time-indepedent; Cro: more assumptions, cannot track training dynamics		
Theorem 4.1	$\mathcal{O}\left(\sqrt{\frac{1}{n}\mathbb{E}\left[tr\left\{\log\left(\Lambda_{W_S^*}^{-1}\Lambda_{w_\mu^*}\right)\right\}\right]}\right)$	General result; hard to measure in practice
Corollary 4.1	$\mathcal{O}\left(\sqrt{\frac{1}{n\eta}\mathbb{E}\left[tr\left\{\log\left(\left[H_{w^*}C_T^{-1}\right]\Lambda_{w_\mu^*}\right)\right\}\right]}\right)$	Under conditions: $H_{w^*}\Lambda_{w^*} = \Lambda_{w^*}H_{w^*}$ and $H_{w^*}\Sigma_T = \mathrm{I}_d$
Corollary 4.2	$\mathcal{O}\left(\sqrt{\frac{d}{n}\log\left(\frac{b}{\eta d}\mathbb{E}\lvert\lvert W_S^*-\hat{w}\rvert\rvert^2+1\right)}\right)$	$\hat{w}$ is flexible; $\frac{2}{\eta} \gg \lambda_1$; other conditions same as Cor. 4.1
Theorem 4.2	$\mathcal{O}\left(\mathbb{E}\sqrt{\frac{M^2 b}{\eta}\mathbb{E}\lvert\lvert W_S^*-W_{S_J}^*\rvert\rvert^2}\right)$	Bounded loss; $\Lambda(W_{s_j}^*) = \Lambda(W_s^*)$; other conditions same as Cor. 4.2

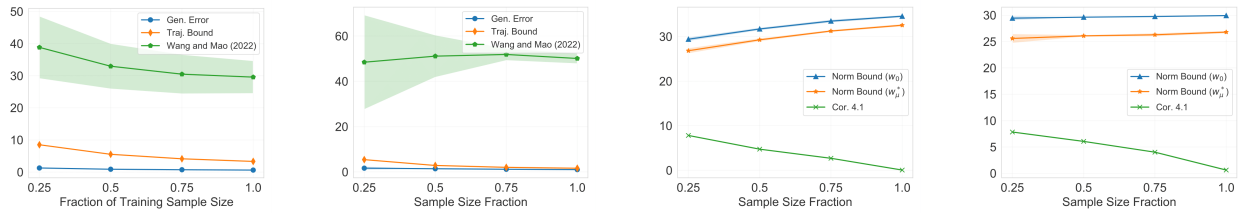
will empirically verify our theoretical results.

5 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.

Implementation and Hyperparameters The implementation in this paper is on PyTorch [Paszke et al., 2019], and all the experiments are carried out on NVIDIA Tesla V100 GPUs (32 GB). Most experiment settings follow Wu et al. [2020], and the code is also based their implementation, which is available at: https://github.com/uuujf/MultiNoise. 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.

SGD and SDE Training Dynamics We implement the SDE training by following the same algorithm given in [Wu et al., 2020, 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 Wu et al. [2020]. The results are depicted in Figure 1. As mentioned earlier,



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

Figure 5: Estimated trajectory-based bound and terminal-state based bound, with R excluded. Zoomed-in figures of generalization error are given in Figure 6 in Appendix.

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 trajectory-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 [Yao et al., 2020]. 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.

Bound Comparison We vary the size of the training sample and empirically estimate several of our bounds in Figure 5, with the subgaussian variance proxy R excluded for simplicity. Thus, the estimated values in Figure 5 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 5a-5b, we compare our Theorem 3.2 with Wang and Mao [2022, Theorem 2]. Since both bounds incorporate the same R, the results in Figures5a to 5b show that our Theorem 3.2 outperforms Wang and Mao [2022, Theorem2]. This aligns with expectations (see Appendix A.1 for additional discussions), considering that the isotropic Gaussian used in the auxiliary weight process of Wang and Mao [2022, Theorem 2] is suboptimal, as demonstrated in Lemma 3.3. Moreover, Figures 5c to 5d hint that norm-based bounds Corollary 4.2 (and Corollary D.1) exhibit growth with n, which are also observed in Nagarajan and Kolter [2019b]. In contrast, Corollary 4.1 effectively captures the trend of generalization error, emphasizing the significance of the geometric properties of local minima. Additionally, while trajectory-based bounds may appear tighter, terminal-state-based bounds seem to have a faster decay rate.

6 OTHER RELATED LITERATURE

Recently, Simsekli et al. [2019], Nguyen et al. [2019], Simsekli et al. [2020], Meng et al. [2020], Gurbuzbalaban et al. [2021], and Gurbuzbalaban et al. [2021] 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. [2021a] and Li et al. [2021] 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 Simsekli et al. [2019] relies on a hidden strong assumption that gradient noise is isotropic and each dimension has the same distribution. For other works on SGD and SDE, see [Hoffer et al., 2017, Xing et al., 2018, Panigrahi et al., 2019, Wu et al., 2020, Zhu et al., 2019, Li et al., 2020, Ziyin et al., 2022].

In addition, there are some generalization bounds using fractal dimensions [Simsekli et al., 2020, Camuto et al., 2021, Dupuis et al., 2023], which are also trajectory-based generalization bounds. Notably, Dupuis et al. [2023] improves previous works by removing the Lipschitz continuity assumption, yet direct comparison of our results with theirs remains challenging. Specifically, one notable difference is that in both Section 3 and Section 4 of our work, we provide in-expectation generalization bounds, while they present high-probability generalization guarantees, which require additional developments for comparison. Moreover, some key components in their bounds are not directly comparable to our gradient noise covariance or Hessian-based quantities, such as their upper and lower box-counting dimensions. On one hand, we believe our results have several advantages. For instance, while the boundedness of loss is essential in their work, we can relax it to a sub-Gaussian condition in

ours. Additionally, our bound is easier to estimate for more complex models. On the other hand, their utilization of intrinsic dimension in the analysis is inspiring and may be possible to incorporate into our analysis for obtaining better results.

7 LIMITATIONS AND FUTURE WORKS

While our current work exhibits certain limitations, such as the requirement of positive definiteness for C_t in our trajectory-based bounds, it's worth noting that recent studies [Frankle and Carbin, 2019, Li et al., 2018, Gur-Ari et al., 2018, Larsen et al., 2022] indicate that many parameters in deep neural networks might be dispensable without affecting generalization. This implies that GD/SGD could potentially occur in a subspace of $\mathbb{R}^d$ termed the "intrinsic dimension" d_{int}. Defining C_t within this invertible subspace, utilizing d_{int}, could potentially overcome our current limitations. Theoretical characterization of intrinsic dimension, however, remains an open problem, and further exploration in this direction is poised to significantly improve our work. In addition, there are also some other promising directions for further improving this work, for example, via using other posterior and prior covariance instead of steady-state covariance (e.g., we also give one in Theorem E.1 in Appendix), and via extending the similar analytical approach used in this work to other optimizer (e.g., Adam, Adagrad, etc.).

Acknowledgements

This work is supported partly by an NSERC Discovery grant. Ziqiao Wang is also supported in part by the NSERC CREATE program through the Interdisciplinary Math and Artificial Intelligence (INTER-MATH-AI) project. The authors would like to thank all the anonymous reviewers and the ACs for their careful reading and valuable suggestions.

References

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

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

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*. PMLR, 2019.

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.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*. PMLR, 2017.

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

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

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

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

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.

Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

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.

Alexander Camuto, George Deligiannidis, Murat A Erdogdu, Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Advances in neural information processing systems*, 34:18774–18788, 2021.

O Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *Vol. 56. Lecture Notes - Monograph Series. Institute of Mathematical Statistics*, 2007.

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.

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

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

Benjamin Dupuis, George Deligiannidis, and Umut Simsekli. Generalization bounds using data-dependent fractal dimensions. In *International Conference on Machine Learning*, pages 8922–8968. PMLR, 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer Berlin, Heidelberg, 2012.

Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022.

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.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

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

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.

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

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.

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.

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.

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.

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.

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.

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

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

Brett W Larsen, Stanislav Fort, Nic Becker, and Surya Ganguli. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. In *International Conference on Learning Representations*, 2022.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

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

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.

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.

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.

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.

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.

Sanae Lotfi, Marc Anton Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-bayes compression bounds so tight that they can explain generalization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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.

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.

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

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

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.

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.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 2017.

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.

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.

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

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.

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

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.

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

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.

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

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.

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.

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.

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.

Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

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

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

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.

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*, 2021a.

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, 2021b.

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.

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.

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.

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.

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*, 2021a.

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, 2021b.

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

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

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.

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

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.

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

Supplementary Material

Ziqiao Wang[1] **Yongyi Mao**[1]

[1]School of Electrical Engineering and Computer Science,

University of Ottawa,

Ottawa, Ontario, Canada

A ADDITIONAL BACKGROUND

A.1 INFORMATION-THEORETIC BOUNDS FOR SGD

Recently, [Neu et al., 2021, Wang and Mao, 2022] apply information-theoretic analysis to the generalization of models trained with SGD by invoking an auxiliary weight process (AWP). We now denote this auxiliary weight process by $\mathcal{A}_{AWP}$. Let $\mathcal{A}_{SGD}$ be the original algorithm of SGD, [Neu et al., 2021, Wang and Mao, 2022] obtain generalization bounds by the following construction,

$$\begin{aligned}\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) =& \mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) + \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) \\ \leq& \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{\text{AWP}};S)}{n}}\right)}_{\text{Lemma 2.1}} + \underbrace{\left|\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right)\right|}_{\text{residual term}},\end{aligned} \tag{8}$$

where W_{AWP} is the output hypothesis by $\mathcal{A}_{AWP}$.

Notably, it remains uncertain whether the residual term is sufficiently small for the information-theoretic bounds of $\mathcal{A}_{AWP}$ to yield meaningful insights into SGD. Although there exists an optimal $\mathcal{A}_{AWP}$ that tightens the bound in Eq. (8), finding such an optimal $\mathcal{A}_{AWP}$ beyond the isotropic Gaussian noise covariance case is challenging. It's worth noting that Wang and Mao [2022] provides an optimal bound for the time-invariant isotropic Gaussian noise case. Nevertheless, our empirical results, as illustrated in Figure 5, demonstrate that the bounds presented in this paper outperform the isotropic Gaussian noise case.

In this paper, we do not attempt to find an optimal $\mathcal{A}_{AWP}$, but instead, we invoke the SDE approximation (i.e. Eq. (5)), denoted as $\mathcal{A}_{SDE}$. Formally,

$$\begin{aligned}\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) =& \mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) + \mathcal{E}_\mu\left(\mathcal{A}_{SDE}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{SDE}\right) \\ \leq& \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{\text{SDE}};S)}{n}}\right)}_{\text{Lemma 2.1}} + \underbrace{\left|\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{SDE}\right)\right|}_{\text{residual term}},\end{aligned} \tag{9}$$

where W_{SDE} is the output hypothesis by $\mathcal{A}_{SDE}$.

Empirical evidence from [Wu et al., 2020, Li et al., 2021] and our Figure 1 suggests that the residual term in Eq. (9) is small. This observation motivates our investigation into the generalization of SGD using the information-theoretic bounds of SDE directly.

A.2 THEORETICAL VALIDATION OF SDE

To theoretically assess the validation of SDE in approximating SGD, two essential technical definitions are necessary.

Definition A.1 (Test Functions)**.** *Let $\mathcal{F}$ denote the set of continuous functions ($\mathbb{R}^d \to \mathbb{R}$) with polynomial growth, i.e. if $\forall$ $f \in \mathcal{F}$, there exists constants $K, \kappa > 0$ s.t. $|f(x)| < K(1 + |x|^\kappa)$ for all $x \in \mathbb{R}$.*

Definition A.2 (Order$-\alpha$ weak approximation)**.** *Let $\eta \in (0, 1)$, $T > 0$ and $N = \lfloor T/\eta \rfloor$. Let $\mathcal{F}$ be the set of test Functions. We say that the SDE in Eq. (4) is an order α weak approximation of the SGD in Eq. (1) if for every $f \in \mathcal{F}$, there exists $C > 0$, independent of η, s.t. for all $k = 0, 1, \ldots, N$,*

$$|\mathbb{E}[f(\omega_{k\eta})] - \mathbb{E}[f(W_k)]| \leq C\eta^\alpha.$$

Below is a classical result.

Lemma A.1 (Li et al. [2017, Theorem 1])**.** *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. $\max_{k=0,1,\ldots,N} |\mathbb{E}[f(\omega_{k\eta})] - \mathbb{E}[f(W_k)]| \leq C\eta$.*

This theorem suggests that SGD and SDE closely track each other when they result in similar distributions of outcomes, such as the returned hypothesis W. In addition, the closeness of distributions is formulated through expectations of suitable classes of test functions, as defined in Definition A.1. As mentioned in Li et al. [2021], of particular interest for machine learning are test functions like generalization error $\mathcal{E}_\mu$, which may not adhere to formal conditions such as differentiability assumed in classical theory but are still valuable for experimental use. Other typical choices of test functions includes weight norm, gradient norm, and the trace of noise covariance.

A.3 GAUSSIAN DISTRIBUTION AROUND LOCAL MINIMUM

A multi-dimensional Ornstein-Uhlenbeck process is defined as

$$dx_t = -\mathbf{H}x_t dt + \mathbf{B}d\theta_t, \tag{10}$$

where $x_t \in \mathbb{R}^d$, $\mathbf{H}$, $\mathbf{B}$ are $d \times d$ matrices and θ_t is an d-dimensional Wiener process.

Denote the density function of x_t as $P(x, t)$, then the corresponding Fokker-Planck equation describes the evolution of $P(x, t)$:

$$\frac{\partial P(x,t)}{\partial t} = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial}{\partial x_i}\left(P(x,t)\sum_{j=1}^{d}\mathbf{H}_{i,j}x_j\right) + \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbf{D}_{i,j}\frac{\partial^2 P(x,t)}{\partial x_i \partial x_j},$$

where $\mathbf{D} = \mathbf{B}\mathbf{B}^\mathbf{T}/2$.

Moreover, if $\mathbf{H}$ is positive define, then a stationary solution of P is given by [Freidlin and Wentzell, 2012]:

$$P(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}x^\mathbf{T}\Sigma^{-1}x\right), \tag{11}$$

where $\Sigma = \mathbb{E}\left[xx^\mathbf{T}\right]$ is the covariance matrix of x.

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,

$$L_s(w) \approx L_s(w^*) + \frac{1}{2}(w - w^*)^\mathbf{T} H_{w^*}(w - w^*). \tag{12}$$

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

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

Let $W_t' \triangleq W_t - w^*$ and recall Eq (10), we thus have the Ornstein-Uhlenbeck process for $x_t = W_t'$ as

$$dW_t' = -\eta H_{w^*} W_t' dt + \eta\sqrt{C_t} d\theta_t. \tag{13}$$

By Eq. (11), we have

$$P(W') \propto \exp\left(-\frac{1}{2} W'^{\mathbf{T}} \Lambda_{w^*}^{-1} W'\right).$$

Consequently, the stationary distribution of W for a given w^* is $\mathcal{N}(w^*, \Lambda_{w^*})$.

For discrete case, we have

$$\begin{aligned} w_t' &= (\mathrm{I}_d - \eta H_{w^*})\, w_{t-1}' + \eta V_t \\ &= (\mathrm{I}_d - \eta H_{w^*})^2\, w_{t-2}' + \eta\left((\mathrm{I}_d - \eta H_{w^*})\, V_{t-1} + V_t\right) \\ &\vdots \\ &= \bar{H}^t w_0' + \eta \sum_{i=0}^{t} \bar{H}^i V_{t-i}, \end{aligned}$$

where $\bar{H} = \mathrm{I}_d - \eta H_{w^*}$. Notably, when t is sufficiently large, then the first term is negligible, especially with a small learning rate, we have $w_t' = w_t - w^* = \eta \sum_{i=0}^{t} \bar{H}^i V_{t-i}$. When C_t does not change in the long time limit, then W_t' is the weighted sum of independent Gaussian random variables, which follows a Gaussian distribution, namely $w_t \sim \mathcal{N}(w^*, \Lambda_{w^*})$. We refer readers to [Liu et al., 2021, Theorem 1-2.] for a relaxed analysis in the discrete case.

B SOME USEFUL FACTS

We present the variational representation of mutual information below.

Lemma B.1 (Polyanskiy and Wu [2019, Corollary 3.1.]). *For two random variables X and Y, we have*

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

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

The following lemma is inspired by the classic Log-sum inequality in Cover and Thomas [2012, Theorem 2.7.1].

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

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

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

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

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

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

which implies

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

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 $\Sigma_p, \Sigma_q \in \mathbb{R}^{d\times d}$.

$$\mathrm{D_{KL}}(p||q) = \frac{1}{2}\left[\log\frac{\det(\Sigma_q)}{\det(\Sigma_p)} - d + (\mu_p - \mu_q)^{\mathbf{T}}\Sigma_q^{-1}(\mu_p - \mu_q) + tr\left\{\Sigma_q^{-1}\Sigma_p\right\}\right]. \tag{14}$$

C OMITTED PROOFS AND ADDITIONAL RESULTS IN SECTION 3

C.1 PROOF OF LEMMA 3.1

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

$$\begin{aligned}
I(W_T; S) &= I(W_{T-1} - \eta G_T + \eta C_T^{1/2} N_T; S) \\
&\leq I(W_{T-1}, -\eta G_T + \eta C_T^{1/2} N_T; S) && (15)\\
&= I(W_{T-1}; S) + I(-\eta G_T + \eta C_T^{1/2} N_T; S|W_{T-1}) && (16)\\
&\vdots \\
&\leq \sum_{t=1}^{T} I(-\eta G_t + \eta C_t^{1/2} N_t; S|W_{t-1}) \\
&= \sum_{t=1}^{T} I(-G_t + C_t^{1/2} N_t; S|W_{t-1}).
\end{aligned}$$

where Eq. (15) 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. (16) 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 3.2

Proof. For any $t \in [T]$, similar to the proof of Lemma B.1 in Polyanskiy and Wu [2019]:

$$\begin{aligned}
& I(-G_t + C_t^{1/2} N_t; S|W_{t-1} = w_{t-1}) \\
= \;& \mathbb{E}_S^{w_{t-1}}\left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S}||Q_{\widehat{G}_t|w_{t-1}})\right] \\
= \;& \mathbb{E}_S^{w_{t-1}}\left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S}||P_{\widehat{G}_t|w_{t-1}}) - \mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1}}||P_{\widehat{G}_t|w_{t-1}})\right] \\
\leq \;& \mathbb{E}_S^{w_{t-1}}\left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S}||P_{\widehat{G}_t|w_{t-1}})\right], && (17)
\end{aligned}$$

where Eq. (17) 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. We first prove Eq. (6). Recall Lemma 3.2 and assume C_t is a positive-definite matrix, for any $t \in [T]$, we have

$$
\begin{aligned}
& I(-G_t + C_t^{1/2} N_t; S | W_{t-1} = w_{t-1}) \\
\leq & \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{-G_t + C_t^{1/2} N_t | w_{t-1}, S} || P_{-\tilde{g}_t + \sigma_t N_t | w_{t-1}}) \right] \\
= & \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[\frac{1}{2} \left[\log \frac{\det(\sigma_t^2 \mathrm{I}_d)}{\det(C_t)} - d + \frac{1}{\sigma_t^2} ((G_t - \tilde{g}_t)^{\mathbf{T}} \mathrm{I}_d^{-1} (G_t - \tilde{g}_t)) + \frac{1}{\sigma_t^2} tr \left\{ \mathrm{I}_d^{-1} C_t \right\} \right] \right] && (18) \\
= & \frac{1}{2} \inf_{\tilde{g}_t, \sigma_t} \mathbb{E}_S^{w_{t-1}} \left[\frac{1}{\sigma_t^2} \left(||G_t - \tilde{g}_t||^2 + tr \left\{ C_t \right\} \right) + d \log \sigma_t^2 - d - tr \left\{ \log C_t \right\} \right], && (19)
\end{aligned}
$$

where Eq. (18) is by Eq. (14), Eq. (19) is due to the fact that $\log \det(C_t) = tr\{\log C_t\}$ when C_t is positive definite.

Recall that $h_1(w) = \mathbb{E}_S^w \left[||G_t - \tilde{g}_t||^2 + tr \left\{ C_t \right\} \right]$ and $h_2(w) = \mathbb{E}_S^w \left[tr \left\{ \log C_t \right\} \right]$, then we have

$$
\begin{aligned}
& \frac{1}{2} \inf_{\tilde{g}_t, \sigma_t} \frac{1}{\sigma_t^2} \mathbb{E}_S^{w_{t-1}} \left[||G_t - \tilde{g}_t||^2 + tr \left\{ C_t \right\} \right] + d \log \sigma_t^2 - d - \mathbb{E}_S^{w_{t-1}} \left[tr \left\{ \log C_t \right\} \right] \\
\leq \; & \frac{1}{2} \inf_{\sigma_t > 0} \frac{1}{\sigma_t^2} h_1(w_{t-1}) + d \log \sigma_t^2 - d - h_2(w_{t-1}) \\
= \; & \frac{1}{2} d \log \frac{h_1(w_{t-1})}{d} - \frac{1}{2} h_2(w_{t-1}),
\end{aligned}
$$

where we fix an arbitrary $\tilde{g}_t$ and use the optimal $\sigma^* = \sqrt{\frac{h_1(w_{t-1})}{d}}$.

Plugging everything into Lemma 3.1 and Lemma 2.1 will obtain Eq. (6).

We then prove the second part. Let $\tilde{g}_t = \mathbb{E}_Z \left[\nabla \ell(w_{t-1}, Z) \right]$, then

$$
\begin{aligned}
h_1(W_{t-1}) = & \mathbb{E}_S^{W_{t-1}} \left[||G_t - \tilde{g}_t||^2 + tr \left\{ C_t \right\} \right] \\
= & \mathbb{E}_S^{W_{t-1}} \left[tr \left\{ (G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^{\mathbf{T}} \right\} \right] + tr \left\{ \mathbb{E}_S^{W_{t-1}} \left[C_t \right] \right\} \\
= & \frac{1}{n} tr \left\{ \Sigma_t^\mu \right\} + \frac{n-b}{b(n-1)} tr \left\{ \mathbb{E}_S^{W_{t-1}} \left[\Sigma_t \right] \right\} && (20) \\
= & \frac{1}{n} tr \left\{ \Sigma_t^\mu \right\} + \frac{n-b}{bn} tr \left\{ \Sigma_t^\mu \right\} && (21) \\
= & \frac{1}{b} tr \left\{ \Sigma_t^\mu \right\},
\end{aligned}
$$

where Eq. (20) is by $\mathbb{E}_S \left[(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^{\mathbf{T}} \right] = \frac{1}{n} \Sigma_t^\mu$ for a given $W_{t-1} = w_{t-1}$ and $C_t = \frac{n-b}{b(n-1)} \Sigma_t$, and Eq. (21) is by $\mathbb{E}_S \left[\Sigma_t \right] = \frac{n-1}{n} \Sigma_t^\mu$. This completes the proof. □

C.4 PROOF OF COROLLARY 3.1

Proof. Let $C_t = \mathrm{I}_d$, by Theorem 3.1,

$$
\begin{aligned}
\mathcal{E}_\mu(\mathcal{A}) \leq & \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} d \mathbb{E}_{W_{t-1}} \left[\log \frac{\mathbb{E}_S^{W_{t-1}} \left[||G_t - \tilde{g}_t||^2 + tr \left\{ C_t \right\} \right]}{d} \right] - \mathbb{E}_{W_{t-1}, S} \left[tr \left\{ \log C_t \right\} \right]} \\
= & \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} d \mathbb{E}_{W_{t-1}} \left[\log \frac{\mathbb{E}_S^{W_{t-1}} \left[||G_t - \tilde{g}_t||^2 \right]}{d} + 1 \right]}.
\end{aligned}
$$

This completes the proof. □

C.5 PROOF OF THEOREM 3.2

Proof. Recall Lemma 3.2, we have

$$
\begin{aligned}
& I(-G_t + C_t^{1/2} N_t; S | W_{t-1} = w_{t-1}) \\
\leq & \inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[\mathrm{D_{KL}}(Q_{\widehat{G}_t|w_{t-1},S} || P_{\widehat{G}_t|w_{t-1}}) \right] \\
= & \inf_{\tilde{c}_t} \mathbb{E}_S^{w_{t-1}} \left[\frac{1}{2} \left[\log \frac{\det(\tilde{c}_t \Sigma_t^\mu)}{\det(C_t)} - d + \frac{1}{\tilde{c}_t} ((G_t - \tilde{g}_t)^{\mathbf{T}} (\Sigma_t^\mu)^{-1} (G_t - \tilde{g}_t)) + \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} C_t \right\} \right] \right] \\
= & \frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} \left[(G_t - \tilde{g}_t)((G_t - \tilde{g}_t)^{\mathbf{T}} \right] \right\} \\
& \quad + \frac{1}{\tilde{c}_t} tr \left\{ (\Sigma_t^\mu)^{-1} \mathbb{E}_S^{w_{t-1}} [C_t] \right\} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \\
= & \frac{1}{2} \inf_{\tilde{c}_t} \frac{1}{\tilde{c}_t n} tr \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + \frac{n-b}{\tilde{c}_t bn} tr \left\{ (\Sigma_t^\mu)^{-1} \Sigma_t^\mu \right\} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \qquad (22) \\
= & \frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{\tilde{c}_t n} + \frac{(n-b)d}{\tilde{c}_t bn} + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} + d \log \tilde{c}_t - d \\
= & \frac{1}{2} \inf_{\tilde{c}_t} \frac{d}{b \tilde{c}_t} + d \log \tilde{c}_t + tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\} - d \\
= & \frac{d}{2} \log \frac{1}{b} + \frac{1}{2} tr \left\{ \log \Sigma_t^\mu - \mathbb{E}_S^{w_{t-1}} [\log C_t] \right\},
\end{aligned}
$$

where the last equality hold when $\tilde{c}_t^* = 1/b$ and Eq. (22) is by

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

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

This completes the proof. □

C.6 PROOF OF LEMMA 3.3

Proof. Let the diagonal element of Σ_t^μ / b in dimension k be a_k, then

$$
\sum_{k=1}^{d} \log a_k \leq (\sum_{k=1}^{d} 1) \cdot \log (\sum_{k=1}^{d} a_k) / (\sum_{k=1}^{d} 1) = d \log (tr \{ \Sigma_t^\mu \} / bd),
$$

where we invoke Lemma B.2.

This completes the proof. □

C.7 ADDITIONAL RESULT VIA DATA-DEPENDENT PRIOR

With the same spirit of Lemma 3.1, to apply Lemma 2.2 to iterative algorithms, we also need the lemma below, which using the full training trajectories KL divergence to upper bound the final output KL divergence.

Lemma C.1 (Negrea et al. [2019, Proposition 2.6.]). *Assume that* $P_{W_0} = Q_{W_0}$, *then* $\mathrm{D_{KL}}(P_{W_T} || Q_{W_T}) \leq \sum_{t=1}^{T} \mathbb{E}_{W_{0:t-1}} \left[\mathrm{D_{KL}}(P_{W_t|W_{0:t-1}} || Q_{W_t|W_{0:t-1}}) \right]$.

Let $G_{Jt} \triangleq \nabla L_{S_J}(W_{t-1})$, the SDE approximation of this prior updating is defined as:

$$
W_t = W_{t-1} - \eta G_{Jt} + \eta C_{Jt}^{\frac{1}{2}} N_t,
$$

where $C_{Jt} = \frac{1}{b}\left(\frac{1}{m}\sum_{i\in J}\nabla\ell_i\nabla\ell_i^{\mathbf{T}} - G_{Jt}G_{Jt}^{\mathbf{T}}\right)$ is the gradient noise covariance of the prior process. In this case, the prior distribution $P_{\mathcal{G}_{Jt}|W_{0:t-1}}$ will be an anisotropic Gaussian distribution. We also assume $n \gg b$, then $C_t = \frac{1}{b}\Sigma_t$.

We denote the difference between G_t and G_{Jt} by

$$\xi_t \triangleq G_{Jt} - G_t.$$

To see the relationship between ξ_t, C_{Jt} and C_t, we present a useful lemma below.

Lemma C.2. *If $m = n-1$, then the following two equations hold,*

$$\mathbb{E}\left[\xi_t\xi_t^{\mathbf{T}}\right] = \frac{b}{(n-1)^2}C_t, \quad \mathbb{E}\left[C_{Jt}\right] = \frac{n(n-2)}{(n-1)^2}C_t,$$

where the expectation is taken over J.

Instead of using Lemma 2.2, we invoke the following result which is a simple extension of [Negrea et al., 2019, Theorem 2.5].

Lemma C.3 ([Wang et al., 2021a, Theorem 1.])**.** *Assume the loss $\ell(w, Z)$ is bounded in $[0, M]$, the expected generalization gap is bounded by*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \frac{M}{\sqrt{2}}\mathbb{E}_{S,J}\sqrt{\mathrm{D_{KL}}(P_{W|S_J}||Q_{W|S})}$$

Comparison with the work of Wang et al. [2021a] Wang et al. [2021a] studies the algorithm of SGD with anisotropic noise, while our SDE analysis focuses on GD with anisotropic noise. This means that the discrete gradient noise arising from mini-batch sampling still exists in their analyzed algorithm, whereas the gradient noise is fully modeled as Gaussian in our Section A.1. Moreover, Wang et al. [2021a] uses matrix analysis tools to optimize the prior distribution. A significant distinction lies in their optimization analysis, which relies on the assumption that the trace of gradient noise covariance remains unchanged during training (see **Constriant 1** in their paper). Additionally, their final optimal posterior covariance is derived based on the assumption that the posterior distribution of W is invariant to the data index, see Assumption 1 in their paper. In contrast, our Section A.1 avoids making these assumptions and demonstrates the superiority of population gradient noise covariance (GNC) in Lemma 3.3, by invoking a variant of the log-sum inequality. In summary, our proof is simpler and more straightforward, while Wang et al. [2021a] makes a stronger claim about the optimality of population GNC based on their additional assumptions.

As introduced in Wang et al. [2021a], the subsequent analysis based on the data-dependent prior bound will rely on an additional assumption.

Assumption 1. *When $m = n-1$, given dataset $S = s$, the distribution $P_{W_t|J,S_J}$ is invariant of J.*

In Wang et al. [2021a], authors mention that in practice, n is usually very large, so this assumption hints that changing one instance in S_J will not make $P_{W_t|J,S_J}$ be too different.

We are now in a position to state the following theorem.

Theorem C.1. *Assume the loss $\ell(w, Z)$ is bounded in $[0, M]$ and Assumption 1 hold,the expected generalization gap of SGD is bounded by*

$$\mathcal{E}_\mu(\mathcal{A}) \leq \mathbb{E}_S\left[\sqrt{M^2\sum_{t=1}^{T}\mathbb{E}_{W_{t-1}}\left[\left(\frac{(b-1)d}{(n-1)^2} + tr\left\{\mathbb{E}_J\left[\log C_t C_{Jt}^{-1}\right]\right\}\right)\right]}\right].$$

Proof. By Lemma C.3 and Lemma C.1, we have

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) \leq & \mathbb{E}_{S,J}\left[\sqrt{\frac{R'^2}{2}\sum_{t=1}^{T}\mathbb{E}_{W_{0:t-1}|S,J}\left[\mathrm{D_{KL}}(P_{W_t|W_{0:t-1},S_J}||Q_{W_t|W_{0:t-1},S})\right]}\right] \\ \leq & \mathbb{E}_S\left[\sqrt{\frac{R'^2}{2}\sum_{t=1}^{T}\mathbb{E}_{W_{0:t-1}|S,J}\left[\mathbb{E}_J\left[\mathrm{D_{KL}}(P_{W_t|W_{0:t-1},S_J}||Q_{W_t|W_{0:t-1},S})\right]\right]}\right],\end{aligned} \tag{23}$$

where Eq. (23) is by Jensen's inequality and Assumption 1.

Recall $\Sigma_t = \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(W_{t-1}, Z_i)\nabla\ell(W_{t-1}, Z_i)^{\mathbf{T}} - \nabla L_S(W_{t-1})\nabla L_S(W_{t-1})^{\mathbf{T}}$ and $C_t = \frac{1}{b}\Sigma_t$.

By the KL divergence between two Gaussian distributions, for any $t \in [T]$, we have

$$
\begin{aligned}
& \mathbb{E}_J\left[\mathrm{D}_{\mathrm{KL}}(P_{W_t|W_{0:t-1},S_J}||Q_{W_t|W_{0:t-1},S})\right] \\
= \; & \mathbb{E}_J\left[\frac{1}{2}\left(\xi_t^{\mathbf{T}} C_t^{-1}\xi_t + \log\frac{\det(C_t)}{\det(C_{Jt})} + tr\{C_t^{-1}C_{Jt}\} - d\right)\right] && (24) \\
= \; & \frac{1}{2}\left(tr\{C_t^{-1}\mathbb{E}_J\left[\xi_t\xi_t^{\mathbf{T}}\right]\} + \mathbb{E}_J\left[\log\frac{\det(C_t)}{\det(C_{Jt})}\right] + \mathbb{E}_J\left[tr\{C_t^{-1}C_{Jt}\}\right] - d\right) \\
= \; & \frac{1}{2}\left(\frac{1}{(n-1)^2}tr\{C_t^{-1}\Sigma_t\} + \mathbb{E}_J\left[\log\frac{\det(C_t)}{\det(C_{Jt})}\right] + \mathbb{E}_J\left[tr\{C_t^{-1}C_{Jt}\}\right] - d\right) && (25) \\
= \; & \frac{1}{2}\left(\frac{b}{(n-1)^2}tr\{\Sigma_t^{-1}\Sigma_t\} + \mathbb{E}_J\left[\log\frac{\det(C_t)}{\det(C_{Jt})}\right] + tr\{C_t^{-1}\mathbb{E}_J\left[C_{Jt}\right]\} - d\right) \\
= \; & \frac{1}{2}\left(\frac{bd}{(n-1)^2} + \frac{n(n-2)d}{(n-1)^2} - d + tr\{\log C_t - \mathbb{E}_J\left[\log C_{Jt}\right]\}\right) && (26) \\
= \; & \frac{1}{2}\left(\frac{(b-1)d}{(n-1)^2} + tr\{\log C_t - \mathbb{E}_J\left[\log C_{Jt}\right]\}\right)
\end{aligned}
$$

where Eq. (25) and Eq. (26) are by Lemma C.2. This concludes the proof. □

Remark C.1. *If the bound in Negrea et al. [2019] is used, then the first term in Eq. (24) is $\xi_t^{\mathbf{T}} C_{Jt}^{-1}\xi_t$, where both C_{Jt} and ξ_t dependent on J, making the bound difficult to analyze.*

The effect of $tr\{\log C_t\}$ on the magnitude of the bound can be decreased by the $tr\{\mathbb{E}_J \log C_{Jt}\}$. If we further consider Taylor expansion of the function $\log C_{Jt}$ around $\mathbb{E}_J[C_{Jt}]$, we have a well-known approximation

$$\mathbb{E}\left[\log C_{Jt}\right] \approx \log \mathbb{E}\left[C_{Jt}\right] - \mathrm{Var}(C_{Jt})/(2\mathbb{E}^2[C_{Jt}]).$$

Thus, recall Lemma C.2, the difference between $tr\{\log C_t\}$ and $tr\{\mathbb{E}_J \log C_{Jt}\}$ would become:

$$\log\left(1 + 1/(n^2 - 2n)\right) + \mathrm{Var}(C_{Jt})/(2\mathbb{E}^2[C_{Jt}]).$$

Thus, the generalization gap should be characterized by the second term above.

When $n \to \infty$, the first term will converges to zero, and for the second term, $\mathbb{E}^2[C_{Jt}]$ will converge to a constant by Lemma C.2, and then the bound is $\mathrm{Var}(C_{Jt})$ will also converges to zero.

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 [Bassily et al., 2018, Hellström and Durisi, 2020, Alquier, 2021]. 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 [Grunwald et al., 2021].

D.1 PROOF OF LEMMA 4.1

Proof. Recall $G_t = \nabla L_s(w_t) = H_{w^*}(w_t - w^*)$. and 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 $T \to \infty$,

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

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

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. (14).

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 $\mathrm{d}_{\mathrm{M}}\left(x, y; \Sigma\right) \triangleq \sqrt{(x-y)^{\mathbf{T}} \Sigma^{-1} (x-y)}$ 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

$$\begin{aligned}
&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\Lambda_{w_\mu^*}^{-1} \Lambda_{W_S^*} - \mathrm{I}_d\right\} + \left(W_S^* - w_\mu^*\right)^{\mathbf{T}} \Lambda_{w_\mu^*}^{-1}\left(W_S^* - w_\mu^*\right)\right] \\
=&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)} + tr\left\{\Lambda_{w_\mu^*}^{-1} \Lambda_{W_S^*} - \mathrm{I}_d\right\} + tr\left\{\Lambda_{w_\mu^*}^{-1}\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^{\mathbf{T}}\right\}\right] \\
=&\mathbb{E}_{S, W_S^*}\left[\log \frac{\det\left(\Lambda_{w_\mu^*}\right)}{\det\left(\Lambda_{W_S^*}\right)}\right] + tr\left\{\Lambda_{w_\mu^*}^{-1} \mathbb{E}_{S, W_S^*}\left[\Lambda_{W_S^*}\right] - \mathrm{I}_d + \Lambda_{w_\mu^*}^{-1} \mathbb{E}_{W_S^*}\left[\left(W_S^* - w_\mu^*\right)\left(W_S^* - w_\mu^*\right)^{\mathbf{T}}\right]\right\}. \qquad (27)
\end{aligned}$$

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

Notice that

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

Therefore,

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

Plugging this into Eq. (27), we have

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

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.

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

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* $\mathcal{E}_\mu(\mathcal{A})\leq\sqrt{\frac{dR^2}{n}\log\left(\frac{2b}{\eta d}\mathbb{E}||W_S^*-W_0||^2+1\right)}$.

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

$$\begin{aligned}
I(W_T;S) &= \inf_{P_{W_T}} \mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}}(Q_{W_T|S}||P_{W_T})\right] \\
&\leq \inf_{\sigma} \mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}}(P_{W_S^*+\sqrt{\frac{\eta}{2b}}N, W_S^*|S}||P_{\tilde{w}+\sigma N})\right] \qquad (28) \\
&= \inf_{\sigma} \mathbb{E}_{S,W_S^*} \left[\mathrm{D}_{\mathrm{KL}}(P_{W_S^*+\sqrt{\frac{\eta}{2b}}N, |S, W_S^*}||P_{\tilde{w}+\sigma N})\right] \\
&= \inf_{\sigma} \frac{1}{2}\mathbb{E}_{S,W_S^*} \left[\frac{1}{\sigma^2}(W_S^*-\tilde{w})^{\mathbf{T}}(W_S^*-\tilde{w}) + \log\frac{\sigma^{2d}}{(\eta/2b)^d} + tr\{\frac{\eta}{2b\sigma^2}\mathrm{I}_d\} - d\right] \\
&= \frac{1}{2}\inf_{\sigma}\frac{1}{\sigma^2}\mathbb{E}_{S,W_S^*}\left[||W_S^*-\tilde{w}||^2 + \frac{\eta d}{2b}\right] + d\log\sigma^2 + d\log\frac{2b}{\eta} - d \\
&= \frac{1}{2}d\log\left(\frac{2b}{\eta d}\mathbb{E}_{S,W_S^*}\left[||W_S^*-\tilde{w}||^2\right]+1\right), \qquad (29)
\end{aligned}$$

where Eq. (28) 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*

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

Remark D.1. *In Theorem 3.1, let $\tilde{g} = 0$ and by applying Jensen's inequality, we could also let the summation and factor T move inside the square root. Then the most different part in Corollary D.2 is that $A_2(t)$ is now removed from the bound.*

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. □

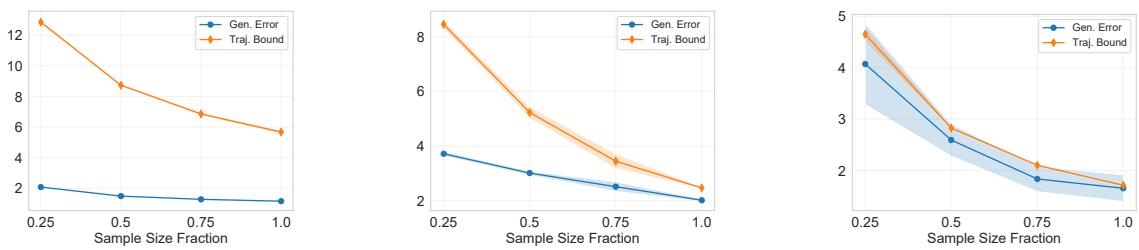
D.7 PROOF OF THEOREM 4.2

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

$$\begin{aligned}
\mathrm{D_{KL}}(Q_{W_T|S=s}||P_{W_T|S_J=s_j}) &= \mathrm{D_{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N|S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N|S_J=s_j}) \\
&\leq \mathrm{D_{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N,W^*_s|S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N,W^*_{s_j}|S_J=s_j}) \qquad (30) \\
&= \mathbb{E}_{W^*_s,W^*_{s_j}}\left[\mathrm{D_{KL}}(Q_{W^*_s+\sqrt{\frac{\eta}{2b}}N|W^*_s,S=s}||P_{W^*_{s_j}+\sqrt{\frac{\eta}{2b}}N|W^*_{s_j},S_J=s_j})\right] \\
&= \mathbb{E}_{W^*_s,W^*_{s_j}}\left[\frac{b}{\eta}||W^*_s - W^*_{s_j}||^2\right], \qquad (31)
\end{aligned}$$

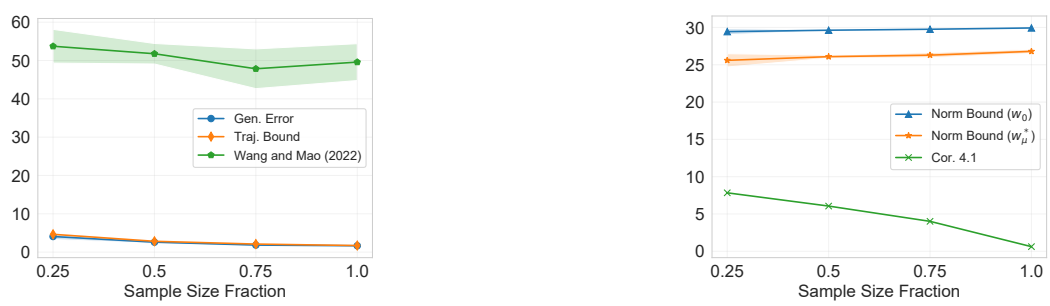
where Eq. (30) is by the chain rule of KL divergence. Plugging the Eq. (31) into Lemma 2.2 will obtain the final result. □

D.8 ADDITIONAL EMPIRICAL RESULTS



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

Figure 6: Zoomed-in of generalization error.



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

Figure 7: Estimated trajectory-based bound and terminal-state based bound, with R excluded. Models trained on CIFAR 10.

E ADDITIONAL RESULT: INVERSE POPULATION FIM AS BOTH POSTERIOR AND PRIOR COVARIANCE

Inspired by some previous works of [Achille et al., 2019, Harutyunyan et al., 2021, Wang et al., 2022], we can also select the inverse population Fisher information matrix $F^{\mu}_{w^*} = \mathbb{E}_Z\left[\nabla\ell(w^*,Z)\nabla\ell(w^*,Z)^{\mathbf{T}}\right]$ as the posterior covariance. Then, the following theorem is obtained.

Theorem E.1. *Under the same conditions in Theorem 4.2, 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 E.1. *Notice that $F^{\mu}_{W^*_S} \approx H^{\mu}_{W^*_S} \approx \Sigma^{\mu}(W^*_S)$ near minima [Pawitan, 2001, Chapter 8], then $tr\{H^{-1}_{W^*_S}\Sigma^{\mu}(W^*_S)\}$ is very close to the Takeuchi Information Criterion [Takeuchi, 1976]. In addition, our bound in Theorem E.1 is similar to Singh et al. [2022, Theorem 3.] with the same convergence rate, although strictly speaking, their result is not a generalization bound. Moreover, as also pointed out in Singh et al. [2022], 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. [2020]).*

The invariance assumption is also used in Wang et al. [2021a]. 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.

Proof of Theorem E.1. 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

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

which completes the proof. □