Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 2, 251–266
DOI: 10.61822/amcs-2025-0018

THE GENERALIZATION ERROR BOUND FOR A STOCHASTIC GRADIENT DESCENT FAMILY VIA A GAUSSIAN APPROXIMATION METHOD

HAO CHEN [a], ZHANFENG MO [b], ZHOUWANG YANG [a,*]

[a] School of Artificial Intelligence and Data Science
University of Science and Technology of China
96 Jinzhai Road, Hefei, 230026, China
e-mail: yangzw@ustc.edu.cn, ch330822@mail.ustc.edu.cn

[b] School of Computer Science and Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798, Singapore
e-mail: ZHANFENG001@ntu.edu.sg

Recent works have developed model complexity based and algorithm based generalization error bounds to explain how stochastic gradient descent (SGD) methods help over-parameterized models generalize better. However, previous works are limited by their scope of analysis and fail to provide comprehensive explanations. In this paper, we propose a novel Gaussian approximation framework to establish generalization error bounds for the $\mathcal{U}$-SGD family, which is a class of SGD with asymptotically unbiased and uniformly bounded gradient noise. We study $\mathcal{U}$-SGD dynamics, and we show both theoretically and numerically that the limiting model parameter distribution tends to be Gaussian, even when the original gradient noise is non-Gaussian. For a $\mathcal{U}$-SGD family, we establish a desirable iteration number independent generalization error bound at the order of $\mathcal{O}((1+\sqrt{\log(p\sqrt{n})})/\sqrt{n})$, where n and p stand for the sample size and parameter dimension. Based on our analysis, we propose two general types of methods to help models generalize better, termed as the additive and multiplicative noise insertions. We show that these methods significantly reduce the dominant term of the generalization error bound.

Keywords: stochastic gradient descent, Gaussian approximation, KL-divergence, generalization bound.

1. Introduction

Stochastic gradient descent (SGD) methods (Sutskever *et al.*, 2013; Bottou, 1998) have been empirically proven to be the most effective and reliable algorithms for training high-dimensional over-parameterized models, such as deep neural networks. Although there are many new attempts and additions to the parameter training of models (Chen, Chang, Meng and Zhang, 2019; Qian *et al.*, 2022; Sulaiman *et al.*, 2024), SGD methods and their variants are still the mainstream method in the industry. Instead of computing the faithful gradient for all training samples, as per standard gradient descent (GD) methods, SGD inserts noise in the full gradient in each iteration. It is surprising that gradient noise dramatically improves a model's generalization performance, especially in many large-scale optimization scenarios (Hardt *et al.*, 2016; Li *et al.*, 2018; Zhang *et al.*, 2016). Empirically, SGD helps over-parameterized models overcome the curse of dimensionality (Weinan *et al.*, 2019).

A fundamental and long-standing problem in machine learning is determining the way that model capacity, data distribution and algorithm properties affect the generalization performance. The key to this problem is to analyze generalization error bounds. A tight generalization error bound guarantees that gaps between training errors and testing errors are small. Hence, generalization error bounds are informative indicators of how well a model generalizes to unseen data. They provide us with guidance on designing effective algorithms to learn robust models.

There have been two main schools of generalization error bounds so far. One is based on the model

*Corresponding author

complexity of the hypothesis, and the other is based on algorithmic properties. A series of works (Bartlett *et al*., 2017; Li, Lu, Wang, Haupt and Zhao, 2019) have ascribed the generalization performance to model structures by showing that a well-designed network enjoys reduced model complexity (based on the Rademacher complexity or VC-dimension), which directly leads to tighter generalization bounds. However, most of these bounds become vacuous when the model complexity (size of parameter space) significantly exceeds the training sample size (Zhang *et al*., 2016). Recently, some stronger model-dependent results have been proposed in the over-parameterized settings (Arora *et al*., 2019; Chen, Mo, Yang and Wang, 2019; Weinan *et al*., 2019). Since model-dependent bounds provide uniform control of generalization error over a whole hypothesis family, any overfitted model in this family significantly impairs the bounds. In deep-learning practices, where advanced algorithms are employed to avoid overfitting, such a uniform bound is too loose to provide a comprehensive explanation of generalization phenomena.

To better understand the roles of algorithms in generalization performance, researchers have resorted to algorithm-dependent generalization error bounds. Algorithm-dependent bounds emphasize an essential connection between the generalization error of an algorithm and its uniform stability (Bousquet and Elisseeff, 2002; Hardt *et al*., 2016; Shalev-Shwartz and Ben-David, 2014). Roughly speaking, stable algorithms, which change by a bounded amount with switching a single data point, generalize well. Various algorithm-dependent bounds have been established for randomized algorithms, and these results rely on uniform stability (Elisseeff *et al*., 2005; Bousquet and Elisseeff, 2002). However, uniform stability is a strong condition, and it does not hold in general settings like univariate logistic regression (Negrea *et al*., 2019). Moreover, most of these results depend on the number of iterations T, which yield vacuous generalization bounds when T tends to infinity. To address these limitations, the PAC-Bayesian stability framework has been proposed to provide sharper generalization guarantees for SGD, see (Li *et al*., 2020; London *et al*., 2014).

In general, the above works either result in a divergent generalization error bound (Hardt *et al*., 2016; Shalev-Shwartz and Ben-David, 2014) or only apply to some special SGD methods such as SGLD (Li *et al*., 2020). The ultimate goal of this work is to obtain an informative generalization error bound for general SGD methods and provide theoretical explanations for how these algorithms attain small generalization errors. An intuitive idea is to use PAC-Bayesian theory to establish a generalization error bound for general SGDs. As we shall see, under standard assumptions, we develop desirable generalization error bounds for various SGD analogues that share important properties in common.

Contributions. The key contribution of this work is a novel Gaussian approximation framework that establishes a tight generalization error bound for a general class of SGD, termed $\mathcal{U}$-SGD. We present our contributions in three aspects.

- By studying the dynamics of $\mathcal{U}$-SGD, in Proposition 1 and 3, we provide the Gaussian approximation error bound of the limiting parameter distribution of SGD. Our result shows that the parameter distribution of SGD tends to be Gaussian-like as the iteration number goes to infinity. Moreover, according to numerical evidence in Section 6, our statement holds for more general cases.
- We establish a simple Gaussian approximation framework to yield generalization error bounds for $\mathcal{U}$-SGD at an order of $\mathcal{O}((1+\sqrt{\log(p\sqrt{n})})/\sqrt{n})$ which does not depend on the number of iterations T, where n and p stand for the sample size and parameter dimension, respectively (Propositions 2 and 4).
- The proposed framework induces two theoretically guaranteed methods, termed the additive and multiplicative noise insertions (Definition 2) to improve the generalization performance of $\mathcal{U}$-SGD. We show that such noise insertions significantly reduce the dominant term of the generalization bound (Proposition 5).

Related work. A stochastic differential equation (SDE) based framework (Li, Tai and E, 2019; Ljung *et al*., 1992; He *et al*., 2019) has been widely used to analyze the dynamics and generalization bounds for SGD variants, such as the stochastic gradient Langevin dynamic algorithm (SGLD) (Welling and Teh, 2011). Li, Tai and E (2019) proved that SGD dynamics can be approximated by a class of SDEs driven by Brownian motion. In (Ljung *et al*., 1992; He *et al*., 2019), for quadratic loss functions, the gradient noise of SGD is assumed to be Gaussian, and the parameter dynamic is treated as its continuous analogue, the Ornstein-Uhlenbeck process (OU-process). Generalization bounds are then derived from classical SDE results for the stationary distribution of the OU-process. On top of these bounds, He *et al*. (2019) show a positive correlation between the ratios of the batch size to the learning rate and the generalization error. However, the Gaussian assumption is contested in (Panigrahi *et al*., 2019; Simsekli *et al*., 2019), which shows that gradient noise can be generally non-Gaussian. The dynamic and approximation analyses of non-Gaussian SGD are discussed in (Dieuleveut *et al*.,

2017; Feng *et al*., 2020), while none of this work provides explicit bounds with respect to the model dimension and the limiting SGD covariance, and thus cannot be applied to derive generalization bounds. In this paper, under non-Gaussian assumptions, we show that SGD with non-Gaussian gradient noise yields Gaussian-like limiting distributions, which enables us to establish a PAC-Bayesian generalization bound by a Gaussian approximation approach.

2. Preliminaries

For any $\mathbf{v} \in \mathbb{R}^p$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$, let $\mathbf{v}[i]$ be the i-th component of $\mathbf{v}$, and let $\mathbf{V}[i,j]$ be the i-by-j component of $\mathbf{V}$. Write $\|\mathbf{v}\|_q \triangleq (\sum_{i=1}^p |\mathbf{v}[i]|^q)^{1/q}$. In the context of supervised learning, there is a training dataset $\mathcal{S} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, where $\mathcal{D}$ is the underlying distribution on $\mathbb{R}^d \times \mathbb{R}$. The hypothesis class is parameterized as $\{f_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$, where the dimension of the parameter space $\boldsymbol{\Theta}$ is p. For a given smooth loss function $l(\hat{y}, y) : \mathbb{R} \times \mathbb{R} \mapsto [0, 1]$, the ultimate goal of supervised learning is to find an optimal parameter $\boldsymbol{\theta}^*$ to minimize the expectation error $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} l(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$. In practice, however, we train the model by minimizing the empirical surrogate, which is the empirical error $\hat{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}} l(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \triangleq \frac{1}{n}\sum_{j=1}^n l(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$. For simplicity, we reparameterize the expectation error and the empirical error as functions of $\boldsymbol{\theta}$, denoted as $L(\boldsymbol{\theta})$ and $L_{\mathcal{S}}(\boldsymbol{\theta})$, respectively. We denote by $\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}) \triangleq \nabla^2 L_{\mathcal{S}}(\boldsymbol{\theta})$ the Hessian of $L_{\mathcal{S}}$ at $\boldsymbol{\theta}$. For a fixed $\boldsymbol{\theta}$, the generalization error is defined as $|L(\boldsymbol{\theta}) - L_{\mathcal{S}}(\boldsymbol{\theta})|$, which measures how well the model $f_{\boldsymbol{\theta}}(\cdot)$ generalizes to unseen data from $\mathcal{D}$.

When we apply SGD to minimize $L_{\mathcal{S}}(\boldsymbol{\theta})$, at the t-th iteration, $\boldsymbol{\theta}$ is updated as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\hat{\boldsymbol{g}}_t$, where $\alpha \in \mathbb{R}^+$ is the learning rate, and $\hat{\boldsymbol{g}}_t$ is a random noisy gradient, which is an estimate of the real gradient $\boldsymbol{g}_t$. It is convenient for us to consider $\boldsymbol{\theta} \sim P$ as a p-dimensional random vector, where P is a probability distribution over $\boldsymbol{\Theta}$. In this context, we are interested in the expectations of $L(\boldsymbol{\theta})$ and $L_{\mathcal{S}}(\boldsymbol{\theta})$. By a slight abuse of notation, we denote $L(P) \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim P} L(\boldsymbol{\theta})$ and $L_{\mathcal{S}}(P) \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim P} L_{\mathcal{S}}(\boldsymbol{\theta})$.

If we suppose that the parameter is initialized as $\boldsymbol{\theta}_0 \sim P_0$, and the learning rate is set as α, at the t-th iteration, the parameter is updated to $\boldsymbol{\theta}_t \sim P_t(\alpha)$. Thus, the updating path of SGD can be viewed as a discrete stochastic process $\{\boldsymbol{\theta}_t\}_{t\geq 1}$. If there is a limit $\boldsymbol{\theta}_\infty \sim P(\alpha)$ of $\{\boldsymbol{\theta}_t\}_{t\geq 1}$, the generalization error bound is defined as an upper bound of $|L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))|$. A tight generalization error bound implies that SGD returns a robust model $f_{\boldsymbol{\theta}_\infty}$.

$\mathcal{U}$-SGD family. We are now confronted with two central problems. What is the limiting distribution $P(\alpha)$ or does it even exist? How do we bound $|L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))|$? Empirically, a gradient noise with a grossly large magnitude significantly impairs the convergence and effectiveness of SGD. As SGD fails to converge, the limiting distribution and generalization error bounds become meaningless. To avoid these ill cases, we focus our interest on SGDs with reasonable noise insertion. Formally, we call this reasonable SGD that satisfies the following assumptions the $\mathcal{U}$-SGD.

Assumption 1. Assume that the following criteria hold for SGD updating:

1. Asymptotic unbiasedness: $\exists \boldsymbol{\theta}^*_{\mathcal{S}} \in \boldsymbol{\Theta}_{\mathcal{S}}$ such that $\lim_{t\to+\infty} \mathbb{E}\boldsymbol{\theta}_t = \boldsymbol{\theta}^*_{\mathcal{S}}$.
2. Uniform boundedness: The stochastic gradient $\hat{\boldsymbol{g}}_t$ satisfies that $\sup_{t\geq 0} \mathbb{E}\|\hat{\boldsymbol{g}}_t\|_2^2 < +\infty$.

In general, $\mathcal{U}$-SGD extends our scope of analysis to SGDs with non-Gaussian gradient noise, which are ubiquitous in stochastic optimization scenarios. For instance, when we employ integer-arithmetic-only quantization schemes to train neural networks (Jacob *et al*., 2018), the gradient noise introduced by the parameter clamping process is bounded, but highly non-Gaussian. The conditions imposed in $\mathcal{U}$-SGD are natural; asymptotic unbiasedness implies that the expectation of the parameter path converges to an optimum. The second condition is mild because the magnitude of the stochastic noisy gradient is observed to be stable in the vicinity of local minima. For a compact $\boldsymbol{\Theta}$, this condition is equivalent to imposing uniform boundedness on loss function l, which is widely embraced in the theoretical literature. In fact, the uniform boundedness of $\hat{\boldsymbol{g}}_t$ guarantees the existence of the covariance matrix $\mathrm{Cov}(\mathbf{u}_t), t \geq 0$, where $\mathbf{u}_t \triangleq \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*_{\mathcal{S}}) - \hat{\boldsymbol{g}}_t$ is an important auxiliary variable for the successive analysis and $\mathbf{H}_{\mathcal{S}} \triangleq \mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}^*_{\mathcal{S}})$. This further implies that the third-moment of $\{\mathbf{u}_t\}_{t\geq 0}$ is uniformly bounded, that is, $\exists\, \Gamma > 0$ such that $\sup_{t\geq 0} \|\mathbf{u}_t\|_\infty^3 \leq \Gamma$. In the next section, we take advantage of the theoretical merits of $\mathcal{U}$-SGD to find the limiting distribution $P(\alpha)$ and obtain desirable generalization bounds.

3. Generalization bound of $\mathcal{U}$-SGD

In this section, we introduce a Gaussian approximation framework to analyze the $\mathcal{U}$-SGD dynamics and establish generalization error bounds for objectives with non-degenerate and degenerate Hessian $\mathbf{H}_{\mathcal{S}}$, also known as the locally strongly convex objectives and generic convex objectives. The detailed proof is contained in Section 5. Roughly speaking, there are two main steps to establish a generalization error bound.

1. We study the $\mathcal{U}$-SGD dynamic, examine the existence of limiting distribution P, and establish

a Gaussian approximation $\hat{P}$ with respect to the Wasserstein-1 metrics (Propositions 1 and 3). We bound the deviation term $|(L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))) - (L(\hat{P}(\alpha)) - L_{\mathcal{S}}(\hat{P}(\alpha)))|$ (Lemmas 2 and 4).

2. The generalization error bound is converted to the KL-divergence bound (Lemma 3) between $\hat{P}$ and a sample-independent distribution Q (Propositions 2 and 4).

3.1. Generalization bound for non degenerate Hessian. To further understand the dynamic and generalization property of $\mathcal{U}$-SGD in locally strong convex cases, we first introduce some standard assumptions.

Assumption 2. Suppose the learning rate $\alpha \in \mathbb{R}^+$ satisfies $\alpha < \lambda_{\max}(\mathbf{H}_{\mathcal{S}})^{-1}$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

Remark 1. The positive definiteness of $\mathbf{H}_{\mathcal{S}}$ can be fulfilled in the local strong convex case. The optima of the loss function have been observed and proven to be stable, even in the presence of gradient noise (Kramers, 1940). The second assumption is that a moderate learning rate is required to train a model well with $\mathcal{U}$-SGD since $\lambda_{\max}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}}) < 1$, $(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})$ can be viewed as a contraction operator. These assumptions are also embraced in (He *et al*., 2019; Mandt *et al*., 2017).

The update equation for SGD is known as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha(\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*_{\mathcal{S}}) - \mathbf{u}_t) = (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\boldsymbol{\theta}_t + \alpha(\mathbf{H}_{\mathcal{S}}\boldsymbol{\theta}^*_{\mathcal{S}} + \mathbf{u}_t)$. At the t-th iteration, we have $(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*_{\mathcal{S}}) = (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*_{\mathcal{S}}) + \alpha\sum_{i=0}^{t-1}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^{t-i-1}\mathbf{u}_i$. The SGD dynamic involves adding new information $\mathbf{u}_t$ to the parameter each time while discounting the historical information $\boldsymbol{\theta}_t$ by multiplying it by $(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})$. As $t \to +\infty$, we have

$$(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\boldsymbol{\theta}_0 \overset{L^2}{\to} \mathbf{0},$$

which implies that the effect of the initialization tends to vanish during the SGD process. This theoretical evidence also coincides with the fact that SGD with proper random initialization works well in large scale optimization cases. Thus, we focus on the asymptotic behaviour of the last term. Thanks to the standard martingale convergence theorem, the existence of the limiting distribution is guaranteed (see Section 5) and we are free to denote the limiting distribution of $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ as $\boldsymbol{\theta}_\infty \sim P(\alpha)$. A well-known result in (McAllester, 1999) states that the information from $P(\alpha)$ provides us with a PAC-Bayesian bound.

Lemma 1. (McAllester, 1999) *In the previous settings, with a probability of at least* $1 - \delta$ *over the choice of* $\mathcal{S}$*, we have*

$$|L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))| \leq \sqrt{\frac{D_{KL}(P(\alpha)\|Q) + \ln n/\delta}{2(n-1)}},$$

where $P(\alpha)$ *is the limiting distribution of* $\{\boldsymbol{\theta}_t\}_{t>0}$*, and* Q *is a distribution that is independent of* $\mathcal{S}$*.* $D_{KL}(P\|Q)$ *is known as the KL-divergence between* P *and* Q*.*

However, this result is far from intended, since we know nothing about $P(\alpha)$ other than its existence. To obtain an informative bound, we turn to an approximate distribution of $P(\alpha)$. Suppose $P(\alpha)$ can be approximated by a parameterized distribution, denoted by $\hat{P}(\alpha)$, then we have

$$\begin{aligned}|L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))| &\leq |(L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))) \\ &\quad - (L(\hat{P}(\alpha)) - L_{\mathcal{S}}(\hat{P}(\alpha)))| \\ &\quad + \sqrt{\frac{D_{\mathrm{KL}}(\hat{P}(\alpha)\|Q) + \ln n/\delta}{2(n-1)}}, \end{aligned} \tag{1}$$

where the first term is called the deviation term, and the second term is the KL-divergence term. Our problem then consists in finding a desirable Gaussian approximation $\hat{P}(\alpha)$ and bounding $D_{\mathrm{KL}}(\hat{P}(\alpha)\|Q)$. The following proposition shows that $P(\alpha)$ can be locally approximated by a Gaussian distribution $\hat{P}(\alpha)$ with respect to the Wasserstein-1 metric (Villani, 2008).

Definition 1. The Wasserstein-1 metric is defined as

$$\begin{aligned}&\mathcal{W}_V^{(1)}(\mu_1, \mu_2) \\ &\quad \triangleq \inf\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \sum_{i=1}^{p} v_i^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\mu(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2) : \\ &\qquad\qquad \mu \in \mathrm{coupl}(\mu_1, \mu_2)\},\end{aligned}$$

where μ_1, μ_2 are two probability measures on $\mathbb{R}^p$ and $V = \{v_1, \ldots, v_p\}$ is an orthogonal basis. $\mathrm{coupl}(\mu_1, \mu_2)$ denotes the collection of probability measures on $\mathbb{R}^p \times \mathbb{R}^p$ with marginals μ_1, μ_2 on the first and second factors, respectively.

Proposition 1. (Gaussian approximation with Wasserstein-1 metric) *Suppose* $\{\mathbf{u}_t\}_{t\geq 0} \in \mathcal{U}$*-SGD, and* $\boldsymbol{\theta}_\infty \sim P(\alpha)$ *is the limiting parameter distribution. Under the previous assumptions, the following statements hold:*

(i) $\mathrm{Cov}(\boldsymbol{\theta}_\infty)$ *exists and* $\mathrm{Cov}(\boldsymbol{\theta}_\infty) = \mathcal{O}(\alpha)$.

(ii) Let $\mathbf{\Sigma}_{\mathcal{S}} = \mathrm{Cov}(\boldsymbol{\theta}_\infty)$ *and* $\hat{P}(\alpha)$ *be a Gaussian distribution* $N(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{\Sigma}_{\mathcal{S}})$. *We have* $\alpha^{-1/2}(P(\alpha) - \hat{P}(\alpha)) \xrightarrow{\text{law}} \mathbf{0}$, *as* $\alpha \to 0$, *where* $\xrightarrow{\text{law}}$ *stands for the convergence in law.*

(iii) Let $\mathbf{\Theta}_K \triangleq \{\boldsymbol{\theta} : |\boldsymbol{\theta}[i] - \boldsymbol{\theta}^*_{\mathcal{S}}[i]| \le K\sqrt{\mathbf{\Sigma}_{\mathcal{S}}[i,i]}, i = 1, 2, \ldots, p\}$. *Denote* $P|_{\mathbf{\Theta}_K}(\alpha), \hat{P}|_{\mathbf{\Theta}_K}(\alpha)$ *the restriction by of* $P(\alpha), \hat{P}(\alpha)$ *on* $\mathbf{\Theta}_K$, *respectively. For all* $K > 0$ *and* $\alpha < (2\lambda_{\max}(\mathbf{H}_{\mathcal{S}}))^{-1}$, *we have*

$$\mathcal{W}^{(1)}(P|_{\mathbf{\Theta}_K}(\alpha), \hat{P}|_{\mathbf{\Theta}_K}(\alpha)) \le \alpha^2 \frac{2\tilde{C}K\Gamma}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})}\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{S}}^{-1}), \quad (2)$$

where $\tilde{C}$ *is a constant smaller than* 36.

Remark 2.

1. Notice that $\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{S}}^{-1}) = \mathcal{O}(\alpha^{-1})$, the local approximation error is bounded by $\mathcal{O}(\alpha)$, and it tends to vanish as $\alpha \to 0$. This theorem reveals an important property of $\mathcal{U}$-SGD: the limiting distribution of $\{\boldsymbol{\theta}_t\}_{t\ge 0}$ is Gaussian-like. In essence, this conclusion follows the Central Limit Theorem (CLT); in other words, when independent non-Gaussian random vectors $\{\alpha\tilde{\mathbf{u}}_t\}_{t\ge 0}$ are added, the sum tends toward a Gaussian distribution as $\alpha \to 0$, even if the original gradient noise is not Gaussian.
2. Intuitively, a smaller learning rate makes the $\mathcal{U}$-SGD dynamic more similar to its continuous analogue, the OU-process, which has a Gaussian limiting distribution. This statement coincides with some previous results (Mandt *et al*., 2017; He *et al*., 2019), where the gradient noise $\{\mathbf{u}_t\}_{t\ge 0}$ is imposed to be independent and identically Gaussian distributed.

Our understanding is numerically validated in Section 6. Numerical experiments show that our statement also holds for more general cases, where the gradient noise is unbounded and the loss function is nonconvex. In these cases, SGD converges to a flat local minimum, where the loss function can be considered quadratic.

Given that $P(\alpha)$ is locally approximated by a Gaussian distribution $\hat{P}(\alpha)$, the following lemma further shows that the deviation term $|(L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))) - (L(\hat{P}(\alpha)) - L_{\mathcal{S}}(\hat{P}(\alpha)))|$ has a tight upper bound.

Lemma 2. *Suppose* $\{\mathbf{u}_t\}_{t\ge 0} \in \mathcal{U}$*-SGD, and* $\boldsymbol{\theta}_\infty \sim P(\alpha)$ *is the limiting parameter distribution. Under the previous assumptions,* $\forall\delta \in (0,1)$, *with a probability of at least* $1-\delta$ *over the choice of* $\mathcal{S}$, *we have*

$$|(L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))) - (L(\hat{P}(\alpha)) - L_{\mathcal{S}}(\hat{P}(\alpha)))| \le \alpha^2 \frac{2C\Gamma\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{S}}^{-1})}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta}} \times \left(\left(2\log(\frac{\sqrt{2/\pi}3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta}p}{2\alpha C\Gamma\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{S}}^{-1})})\right)^{\frac{1}{2}} + \left(2\log(\frac{\sqrt{2/\pi}3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta}p}{2\alpha C\Gamma\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{S}}^{-1})})\right)^{-\frac{1}{2}} \right), \quad (3)$$

where C *is a constant independent of* δ *and* θ_∞.

The deviation term (3) is bounded by $\mathcal{O}(\alpha((\log(\sqrt{n}p/\alpha))^{\frac{1}{2}} + (\log(\sqrt{n}p/\alpha))^{-\frac{1}{2}})/\sqrt{n}) \approx \mathcal{O}((\alpha\sqrt{n})^{1-\epsilon}(\log\sqrt{n}p)^{\frac{1}{2}})$ for an arbitrarily small constant $\epsilon > 0$. Again, the bound tends to zero when $\alpha \to 0$, and it remains desirable even when p is large. The only thing left is to bound the KL-divergence term. Let $Q \triangleq N(\boldsymbol{\theta}^*, \mathbf{\Sigma})$, where $\mathbf{\Sigma} \triangleq \mathbb{E}_{\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}}$, and we have $\mathbf{\Sigma} \succ \mathbf{0}$. To better illustrate our result, we introduce another assumption.

Assumption 3. Suppose that $\exists\lambda_* > 0$ such that $\lambda_{\min}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}}) \ge \lambda_*$ holds for all $\mathcal{S}$; $\exists M > 0$ such that $\mathbb{E}\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F^2 \le M$, where $\|\cdot\|_F$ is the Frobenius norm.

The first assumption $\lambda_{\min}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}}) > \lambda_* > 0$ guarantees that $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}}$ is not singular. In the second assumption, $\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F^2$ evaluates how significantly the eigenvalues of $\mathbf{\Sigma}_{\mathcal{S}}$ deviate from the ones of $\mathbf{\Sigma}$. A small M value implies that $\mathbf{\Sigma}_{\mathcal{S}}$ is highly concentrated to $\mathbf{\Sigma}$, which means the training data $\mathcal{S}$ are stable and representative. Since $\mathbb{E}_{\mathcal{S}}[\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}}] = \mathbf{I}$, the Law of Large Numbers yields $\mathbb{E}\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F^2 \xrightarrow{\text{law}} 0$ as the sample size $\#\mathcal{S} = n$ tends to infinity. This implies $\mathbb{E}\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F^2$ is uniformly bounded with respect to n and the soundness of Assumption 2.

Lemma 3. (KL-divergence bound) *Suppose* $\{\mathbf{u}_t\}_{t\ge 0}$ *belongs to* $\mathcal{U}$*-SGD. Let* $P(\alpha)$ *be the limited distribution of* $\{\boldsymbol{\theta}_t\}_{t\ge 0}$, *and* $\hat{P}(\alpha)$ *be the approximate distribution* $N(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{\Sigma}_{\mathcal{S}})$. *Under the previous assumptions, for* $\forall\delta > 0$, *with a probability of at least* $1-\delta$ *over the choice of* $\mathcal{S}$, *it holds that*

$$D_{KL}(\hat{P}\|Q) \le \frac{(1+\delta^{-1})}{2} \times \max\left\{\frac{-\log\lambda_* + \lambda_* - 1}{1-\lambda_*}, 1\right\}M + \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_{\mathcal{S}})^\top\mathbf{\Sigma}^{-1}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_{\mathcal{S}}). \quad (4)$$

We develop a dimension-free upper bound for the KL-divergence term. The combination of the aforementioned results yields a probabilistic generalization error bound for $\mathcal{U}$-SGD.

Proposition 2. (Generalization error bound of $\mathcal{U}$-SGD) *Under the previous assumptions, given* $\{\mathbf{u}_t\}_{t\ge 0}$ *belonging to* $\mathcal{U}$*-SGD, for* $\forall\delta_1, \delta_2 > 0$, *with a probability of at least* $1-\delta_1-\delta_2-\delta_3$ *over training data* $\mathcal{S}$ *of size* n, *we*

have the following inequality for the limiting parameter distribution $P(\alpha)$:

$$
\begin{aligned}
&|L(P(\alpha)) - L_{\mathcal{S}}(P(\alpha))| \\
&\le \frac{1}{\sqrt{2n-2}}\Bigg(\frac{(1+\delta_1^{-1})}{2} \\
&\quad \times \max\left\{\frac{-\log\lambda_* + \lambda_* - 1}{1-\lambda_*}, 1\right\} M \\
&\quad + \log(\frac{n}{\delta_2}) + \frac{1}{2}(\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*)\Bigg)^{\frac{1}{2}} \\
&\quad + \alpha^2 \frac{2C\Gamma \mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{S}}^{-1})}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta_3}} \\
&\quad \times \Bigg(\big(2\log(\frac{\sqrt{2/\pi}3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta_3}p}{2\alpha^2 C\Gamma \mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{S}}^{-1})})\big)^{\frac{1}{2}} \\
&\quad + \big(2\log(\frac{\sqrt{2/\pi}3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n\delta_3}p}{2\alpha^2 C\Gamma \mathrm{tr}(\boldsymbol{\Sigma}_{\mathcal{S}}^{-1})})\big)^{-\frac{1}{2}}\Bigg), \qquad (5)
\end{aligned}
$$

where C is a constant independent of δ and $\boldsymbol{\theta}_\infty$.

Remark 3. (*Comparison with existing work*)

1. Our order bound $\mathcal{O}(\sqrt{\log(n)+\alpha^{-1}}/\sqrt{n} + \alpha\sqrt{\log(p\sqrt{n})}/\sqrt{n})$ sheds light on a trade-off between the KL-divergence term and the deviation term. When the learning rate α is small, the limiting parameter distribution is highly Gaussian-like, and the first term dominates the bound. In contrast to algorithmic stability based generalization bounds that scale as $\mathcal{O}(T/n)$ (Hardt *et al*., 2016; Shalev-Shwartz and Ben-David, 2014), our bound is independent of the total iteration number T, which is more favorable for realistic SGD training.

2. Our work explicitly analyzes the order of the bound with respect to both model dimension p and sample size n. On the contrary, the dependence of algorithmic stability based bounds on p is blurred by same artificially defined constants. From this perspective, our bound can better explain generalization under the over-parameterized setting. In fact, by further assuming $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F$ and $\|\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*\|_2$ are at an order of $\mathcal{O}(1/\sqrt{n})$, we can refine the KL-divergence term to the order of $\mathcal{O}(1/n)$. Since $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}} - \mathbf{I}\|_F$ and $\|\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*\|_2$ tend to 0 almost surely, this additional assumption is merely asking for a mild error rate control for these terms, which does not exceed the naive Monte Carlo error rate.

3.2. Generalization bound for degenerate Hessian. To further understand how SGD helps high dimensional machine learning models generalize well, we need to establish a generalization bound for locally convex objectives with a degenerate Hessian $\mathbf{H}_{\mathcal{S}}$. Unfortunately, in these cases, the distribution of $\mathcal{U}$-SGD path $\{\boldsymbol{\theta}_t\}_{t\ge 0}$ is no longer guaranteed to converge, and the previous Gaussian approximation framework becomes inapplicable.

Example 1. (*SGD fails to converge for degenerate $\mathbf{H}_{\mathcal{S}}$*) Suppose we optimize $\boldsymbol{\theta}$ with SGD by minimizing the following objective with a degenerate Hessian:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= L_1 + L_2 \\
&= \big(\boldsymbol{\theta}[1]^2 + \boldsymbol{\theta}[3]\big) + \big(\boldsymbol{\theta}[2]^2 - \boldsymbol{\theta}[3]\big). \qquad (6)
\end{aligned}
$$

At the t-th iteration, $\boldsymbol{\theta}$ does not converge as $\boldsymbol{\theta}_t[3] = \boldsymbol{\theta}_0 + \sum_{i=1}^t \mathrm{Ber}(-1,1)$, where $\mathrm{Ber}(-1,1)$ is the Bernoulli distribution. ♦

A straightforward approach to address this issue, is to recover the positive definiteness of $\mathbf{H}_{\mathcal{S}}$ via ℓ_2-regularization. Again, we can apply the previous arguments by substituting $\mathbf{H}_{\mathcal{S}}$ with ℓ_2 penalized Hessian $\mathbf{H}_{\mathcal{S}} + \lambda\mathbf{I}$. However, ℓ_2 regularization introduces an inevitable bias to the objective, and machine learning models generalize well even in the absence of ℓ_2 regularization. This urges us to find another solution for understanding generalization of unregularized models. Hence, we recast our Gaussian approximation framework by employing a dynamic learning rate α_t and establishing a generalization bound for the T-th iteration. Specifically, we first establish an order $\mathcal{O}(\max_{t<T}\alpha_t + \frac{\sum_{i=1}^T \alpha_t^3}{\sum_{i=1}^T \alpha_t^2})$ Gaussian approximation error bound. Then, we establish the revised KL-divergence bound, which finally leads us to a PAC-Bayesian generalization bound.

Proposition 3. (Gaussian approximation with Wasserstein-1 metric) *Suppose $\{\mathbf{u}_t\}_{t=0}^{T-1} \in \mathcal{U}$-SGD, and $\boldsymbol{\theta}_T \sim P(\boldsymbol{\alpha}, T)$ is the parameter distribution at time T. Under the previous assumptions, the following statements holds:*

(i) $\mathrm{Cov}(\boldsymbol{\theta}_T) = \mathcal{O}(\sum_{i=1}^T \alpha_t^2)$.

(ii) Let $\boldsymbol{\Sigma}_{\mathcal{S}} = \mathrm{Cov}(\boldsymbol{\theta}_T)$ and $\hat{P}(\boldsymbol{\alpha}, T)$ be a Gaussian distribution $N(\mathbb{E}[\boldsymbol{\theta}_T], \boldsymbol{\Sigma}_{\mathcal{S}})$, and let $\boldsymbol{\Theta}_K \triangleq \{\boldsymbol{\theta} : |\boldsymbol{\theta}[i] - \mathbb{E}[\boldsymbol{\theta}_T][i]| \le K\sqrt{\boldsymbol{\Sigma}_{\mathcal{S}}[i,i]}, i = 1,2,\ldots,p\}$. Denote by $P|_{\boldsymbol{\Theta}_K}(\boldsymbol{\alpha}, T), \hat{P}|_{\boldsymbol{\Theta}_K}(\boldsymbol{\alpha}, T)$ the restrictions of $P(\boldsymbol{\alpha}, T), \hat{P}(\boldsymbol{\alpha}, T)$ on $\boldsymbol{\Theta}_K$, respectively. For $\forall K > 0$, we have

$$
\begin{aligned}
&\mathcal{W}^{(1)}(P|_{\boldsymbol{\Theta}_K}(\boldsymbol{\alpha}, T), \hat{P}|_{\boldsymbol{\Theta}_K}(\boldsymbol{\alpha}, T)) \\
&\quad \le 2\tilde{C}K\left(\frac{\max_{t<T}\alpha_t \cdot \boldsymbol{\Gamma}}{3\tilde{\lambda}_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^T \alpha_t^2}\right), \qquad (7)
\end{aligned}
$$

where $\tilde{C}$ is a constant independent of $\boldsymbol{\theta}_T$, and $\tilde{\lambda}_{\min}$ denotes the smallest nonzero eigenvalue of $\mathbf{H}_{\mathcal{S}}$.

Remark 4. This proposition is an analog of Proposition 1 in singular cases. In contrast to Proposition 1, (7) is independent of $\lambda_{\min}(\mathbf{H}_{\mathcal{S}})$ which is essentially 0. Hence, it does not explode when $\mathbf{H}_{\mathcal{S}}$ is semi-positive definite.

By an argument similar to Lemma 3, we obtain the following deviation bound.

Lemma 4. *Suppose $\{\mathbf{u}_t\}_{t=0}^{T} \in \mathcal{U}$-SGD, and $\boldsymbol{\theta}_T \sim P(\boldsymbol{\alpha}, T)$ is the limiting parameter distribution. Under the previous assumptions, $\forall \delta \in (0,1)$, with a probability of at least $1-\delta$ over the choice of $\mathcal{S}$, we have*

$$\begin{aligned}
&|(L(P(\boldsymbol{\alpha},T)) - L_{\mathcal{S}}(P(\boldsymbol{\alpha},T)))\\
&-(L(\hat{P}(\boldsymbol{\alpha},T)) - L_{\mathcal{S}}(\hat{P}(\boldsymbol{\alpha},T)))|\\
&\le \frac{2C\big(\frac{\max_{t\le T}\alpha_t\cdot\boldsymbol{\Gamma}}{3\lambda_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\big)}{\sqrt{n\delta}}\\
&\quad\times\left(\left(2\log\left(\frac{\sqrt{2/\pi}\sqrt{n\delta}p}{2C\big(\frac{\max_{t\le T}\alpha_t\cdot\boldsymbol{\Gamma}}{3\lambda_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\big)}\right)\right)^{\frac{1}{2}}\right.\\
&\quad\left.+\left(2\log\left(\frac{\sqrt{2/\pi}\sqrt{n\delta}p}{2C\big(\frac{\max_{t\le T}\alpha_t\cdot\boldsymbol{\Gamma}}{3\lambda_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\big)}\right)\right)^{-\frac{1}{2}}\right),
\end{aligned} \tag{8}$$

where C is a constant independent of δ and $\boldsymbol{\theta}_T$.

The next step is to bound the KL-divergence term $D_{\mathrm{KL}}(\hat{P}(\boldsymbol{\alpha},T)\|Q(\boldsymbol{\alpha},T))$, where $Q(\boldsymbol{\alpha},T)$ is chosen to be a data-independent Gaussian distribution $N(\boldsymbol{\theta}^*,\boldsymbol{\Sigma})$. Plugging in the deviation bound and the KL-divergence bound yields our desirable generalization bound.

Proposition 4. (Generalization error bound of $\mathcal{U}$-SGD) *Under the previous assumptions, given $\{\mathbf{u}_t\}_{t=0}^{T}$ belonging to $\mathcal{U}$-SGD, $\forall\delta>0$, with a probability of at least $1-\delta_1-\delta_2-\delta_3$ over training data $\mathcal{S}$ of size n, we have the following inequality for the limiting parameter distribution $P(\alpha)$:*

$$\begin{aligned}
&|L(P(\boldsymbol{\alpha},T)) - L_{\mathcal{S}}(P(\boldsymbol{\alpha},T))|\\
&\le \frac{1}{\sqrt{2n-2}}\left(\frac{(1+\delta_1^{-1})}{2}\max\left\{\frac{-\log\lambda_* + \lambda_* - 1}{1-\lambda_*}, 1\right\}M\right.\\
&\quad\left.+\log(\frac{n}{\delta_2}) + \frac{1}{2\alpha}(\boldsymbol{\theta}_{\mathcal{S}}^* - \boldsymbol{\theta}^*)^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_{\mathcal{S}}^* - \boldsymbol{\theta}^*)\right)^{\frac{1}{2}}\\
&\quad+\frac{2C\Gamma}{\sqrt{n\delta_3}}\left(\left(2\log\left(\frac{\sqrt{2/\pi}\sqrt{n\delta_3}p}{2C\Gamma\big(\frac{\max_{t\le T}\alpha_t\cdot\boldsymbol{\Gamma}}{3\lambda_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\big)}\right)\right)^{\frac{1}{2}}\right.\\
&\quad\left.+\left(2\log\left(\frac{\sqrt{2/\pi}\sqrt{n\delta_3}p}{2C\Gamma\big(\frac{\max_{t\le T}\alpha_t\cdot\boldsymbol{\Gamma}}{3\lambda_{\min}} + \frac{\sum_{i=0}^{T-1}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\big)}\right)\right)^{-\frac{1}{2}}\right),
\end{aligned} \tag{9}$$

where C is a constant independent of δ and $\boldsymbol{\theta}_T$.

Remark 5. (*Comparison with existing work*) This result substantially enlarges the scope of Proposition 2, by liberating us from the strongly convexity restriction. If we further assume $\alpha_t = \mathcal{O}(1/t)$ as in (Hardt *et al.*, 2016), our result is of order $\mathcal{O}((1+\sqrt{\log(p\sqrt{n})})/\sqrt{n})$ which is still independent of T and stays informative for large T, while the result in (Hardt *et al.*, 2016) is of order $\mathcal{O}(T^{1-\epsilon}/n)$, where $\epsilon\in(0,1)$.

4. $\mathcal{U}$-SGD with additive and multiplicative noise insertion

So far, we have studied the dynamics and the approximate limiting distribution of $\mathcal{U}$-SGD. On top of this framework, we established a tight generalization error bound in (5). In this subsection, we propose two general types of methods to help the model $f_{\boldsymbol{\theta}}$ generalize better. We first introduce the definitions of additive and multiplicative noise insertions. Then, we show that this noise insertion scheme further reduces the dominant term of the generalization bound for $\mathcal{U}$-SGD. In practice, some widely used algorithms are shown to be specific instances of our theoretical framework.

Suppose the gradient noise of $\mathcal{U}$-SGD is independent and identically distributed, that is, $\{\mathbf{u}_t\}_{t\ge0} \overset{\text{i.i.d.}}{\sim} \mathbf{u}$, with $\mathrm{Cov}(\mathbf{u}) = \mathbf{C}$. According to the previous discussions, the dominant term in (4) can be as bad as $\alpha^{-1}\lambda_{\min}(\mathbf{C})^{-1}\lambda_{\max}(\mathbf{H}_{\mathcal{S}})\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\mathcal{S}}^*\|_2^2$. The generalization error bound fails when $\lambda_{\min}(\mathbf{C}) \approx 0$. In these cases, the component-wise collinearity of $\mathbf{u}$ is strong, and it impairs the generalization bound. To alleviate this collinearity, a straightforward idea is to insert some isotropic noise into $\mathbf{u}$. Generally, these schemes can be characterized as additive and multiplicative noise insertion.

Definition 2. Assume that $\{\mathbf{u}_t\}_{t\ge0} \in \mathcal{U}$-SGD, $\boldsymbol{\theta}_t$ is the model parameter at the t-th iteration, and $\alpha>0$ is the learning rate.

1. Additive noise insertion: The SGD updating equation with additive noise insertion follows $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha(\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\mathcal{S}}^*) - \mathbf{u}_t + \boldsymbol{\eta}_t)$, where $\{\boldsymbol{\eta}_t\}_{t\ge0}$ are identically and independently distributed (i.i.d.), $\forall t\ge0$, $\boldsymbol{\eta}_t \perp \{\mathbf{u}_s\}_{0\le s\le t}$. Moreover, $\{\boldsymbol{\eta}_t\}_{t\ge0}$ is centralized and isotropic with finite variance, that is, $\forall t, \mathbb{E}\boldsymbol{\eta}_t = \mathbf{0}$, $\{\boldsymbol{\eta}_t[i]\}_{i=1}^{p}$ are i.i.d. and $\mathrm{Var}(\eta_0[1]) < +\infty$.

2. Multiplicative Noise insertion: The SGD updating equation with multiplicative noise insertion follows $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}_t \odot (\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\mathcal{S}}^*) - \mathbf{u}_t)$, where $\odot$ denotes the Hadamard product, $\{\boldsymbol{\gamma}_t\}_{t\ge0}$ are i.i.d., $\forall t\ge0, \boldsymbol{\gamma}_t \perp \{\mathbf{u}_s\}_{0\le s\le t}$. Moreover, we assume $\forall t, \mathbb{E}\boldsymbol{\gamma}_t = \mathbf{1}$, $\{\boldsymbol{\gamma}_t[i]\}_{i=1}^{p}$ are i.i.d. with $1 < \mathbb{E}(\boldsymbol{\gamma}_0[1]^2) < +\infty$.

Some SGD variants can be characterized as additive and multiplicative noise insertions. The SGLD algorithm is known as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha(g_{B_t}(\boldsymbol{\theta}_t) + \sigma\boldsymbol{\varepsilon}_t)$, where $\sigma > 0$, $g_{B_t}(\boldsymbol{\theta}_t)$ is the mini-batch gradient, and $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I})$. Hence, SGLD can be viewed as additive noise insertion by taking $\mathbf{u}_t \triangleq (\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*_{\mathcal{S}}) - g_{B_t}(\boldsymbol{\theta}_t))$ and $\boldsymbol{\eta}_t \triangleq \sigma\boldsymbol{\varepsilon}_t$. SGD with edge-wise dropout (Rong *et al*., 2020) during backpropagation can be formulated as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}_t \odot (\mathbf{H}_{\mathcal{S}}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*_{\mathcal{S}}) - \mathbf{u}_t)$, with $\{\gamma_t[i]\}_{i=1}^p \overset{\text{i.i.d.}}{\sim} \beta^{-1}\mathrm{Ber}(\beta)$, where $\mathrm{Ber}(\beta), \beta \in (0,1)$ is the Bernoulli distribution with a success probability equal to β. Thus, the edge-wise dropout can be considered as a specific multiplicative noise insertion.

Notice that, Wu *et al*. (2020) proposed the so-called 'multiplicative-framework' to extend the mini-batch sampling SGD to SGD variants with general sampling noise. In contrast, our new framework characterizes such data-wise sampling procedures (e.g., mini-batch SGD) as instances of additive noise insertion. For algorithms that perform parameter-wise sampling over the full gradient (e.g., edge-wise dropout), we characterize them as multiplicative noise insertions. The following proposition studies the Gaussian approximation for these cases.

Proposition 5. *Suppose $\{\boldsymbol{\theta}_t^{'}\}_{t\geq 0}$ is generated by $\mathcal{U}$-SGD with additive or multiplicative noise insertions. Let $\boldsymbol{\theta}_{\infty}^{'} \sim P^{'}(\alpha)$ be the limiting parameter distribution.*

(i) Additive noise insertion: $\boldsymbol{\theta}_{\infty}^{'} = \alpha\sum_{t\geq 0}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}^{'}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t$ exists with $\mathbf{C}^{'} = (\mathbf{C} + \mathrm{Var}(\boldsymbol{\eta}_0[1])\mathbf{I})$ and $\mathrm{Cov}(\boldsymbol{\theta}_{\infty}^{'}) = \mathcal{O}(\alpha)$. Let $\boldsymbol{\Sigma}_{\mathcal{S}}^{'} \triangleq \mathrm{Cov}(\boldsymbol{\theta}_{\infty}^{'})$ and $\hat{P}^{'}(\alpha)$ be a Gaussian distribution $N(\boldsymbol{\theta}^*_{\mathcal{S}}, \boldsymbol{\Sigma}_{\mathcal{S}}^{'})$. Then, $\alpha^{-1/2}(P^{'}(\alpha) - \hat{P}^{'}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$ as $\alpha \to 0$.

(ii) Multiplicative noise insertion: $\boldsymbol{\theta}_{\infty}^{'} = \alpha\sum_{t\geq 0}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}^{'}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t$ exists with $\mathbf{C}^{'} = (\mathbf{C} + (\mathbb{E}\boldsymbol{\gamma}_0[1]^2 - 1)\mathrm{diag}(\mathbf{C}))$ and $\mathrm{Cov}(\boldsymbol{\theta}_{\infty}^{'}) = \mathcal{O}(\alpha)$. Let $\boldsymbol{\Sigma}_{\mathcal{S}}^{'} \triangleq \mathrm{Cov}(\boldsymbol{\theta}_{\infty}^{'})$ and $\hat{P}^{'}(\alpha)$ be a Gaussian distribution $N(\boldsymbol{\theta}^*_{\mathcal{S}}, \boldsymbol{\Sigma}_{\mathcal{S}}^{'})$. Then, $\alpha^{-1/2}(P^{'}(\alpha) - \hat{P}^{'}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$ as $\alpha \to 0$.

Remark 6. Although the noise insertion method can reduce the generalization error, it will also increase the variance of the parameter distribution or affect the convergence of SGD. To some extent, this reflects the contradiction between empirical error and generalization error. From the perspective of model selection, if we use a suitable model so that the loss function has a flat landscape in the main range of parameter distribution, the noise insertion method can reduce the generalization error while maintaining the loss value. However, this is beyond the scope because the focus of this paper is on the generalization error with SGD.

By simply substituting $\boldsymbol{\Sigma}_{\mathcal{S}}$ with $\boldsymbol{\Sigma}_{\mathcal{S}}^{'}$, the previous generalization error bound can be parallelly extended to $\mathcal{U}$-SGD with additive and multiplicative noise insertions. Both of these methods share important theoretical merits: the diagonal entries of $\mathbf{C}$ are further augmented, and the component-wise collinearity of $\mathbf{u}$ is alleviated. Hence, additive and multiplicative noise insertions significantly reduce $\frac{\alpha}{4}(\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*)^\top\mathbf{H}_{\mathcal{S}}^{1/2}\mathbf{C}^{-1}\mathbf{H}_{\mathcal{S}}^{1/2}(\boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{\theta}^*)$, the dominant term of the generalization error bound, especially when p is extremely large and α is small.

5. Proof of propositions

5.1. Proof of Proposition 1. (i): Since $\mathbf{u}_t$ is uniformly bounded, $\exists\mathbf{C} \in \mathbb{R}^{p\times p}, \mathbf{C} \succ 0$ such that $\mathrm{Cov}(\mathbf{u}_t) \prec \mathbf{C}$ holds for any t. Then we have

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\theta}_\infty) &= \alpha^2\sum_{t\geq 0}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathrm{Cov}(\mathbf{u}_t)(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\\ &\leq \alpha^2\sum_{t\geq 0}\lambda_{\max}^{2t}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\mathbf{C}\\ &= \frac{\alpha^2}{1 - \lambda_{\max}^2(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})}\mathbf{C}\\ &= \mathcal{O}(\alpha).\end{aligned}$$

(ii): Let $\phi_{\boldsymbol{\theta}_t}$ be the characteristic function of $\boldsymbol{\theta}_t$. Thus

$$\begin{aligned}\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) &= \prod_{t\geq 0}\phi_{u_t}(\alpha(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s})\\ &= \prod_{t\geq 0}(1 - \alpha^2\mathbf{s}^\top(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathrm{Cov}(\mathbf{u}_t)\\ &\quad\times(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s} + o(\alpha^2\|\mathbf{s}\|_2^2))\\ &= 1 - \mathbf{s}^\top\mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s} + o(\|\mathbf{s}\|_2^2\alpha^2),\end{aligned}$$

By the proof of (i), $\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) \to 1 - \mathbf{s}^\top\mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s}$ as $\alpha \to 0$, thus $\alpha^{-1/2}(P(\alpha) - \hat{P}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$.

(iii): Let event $A = \{\boldsymbol{\theta} \mid |\boldsymbol{\theta}[i] - \boldsymbol{\theta}^*_{\mathcal{S}}[i]| \leq K\sqrt{\boldsymbol{\Sigma}[i][i]}, i = 1, \ldots, p\}$. We have

$$\begin{aligned}&\mathcal{W}^{(1)}(P|_{\Theta_K}, \hat{P}|_{\Theta_K})\\ &= \inf_{F_{\boldsymbol{\theta}_1} = F_{P|_{\Theta_K}}, F_{\boldsymbol{\theta}_2} = F_{\hat{P}|_{\Theta_K}}}\mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1\\ &\leq \inf_{F_{\boldsymbol{\theta}_1} = F_P, F_{\boldsymbol{\theta}_2} = F_{\hat{P}}}\mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1\\ &\quad\cdot\chi_A(\boldsymbol{\theta}_1)\cdot\chi_A(\boldsymbol{\theta}_2)]\\ &\leq \inf_{F_{\theta_1} = F_P, F_{\theta_2} = F_{\hat{P}}}\sum_{i=1}^p\int_{\theta^*_{\mathcal{S}}[i] - K\sqrt{\boldsymbol{\Sigma}[i][i]}}^{\theta^*_{\mathcal{S}}[i] + K\sqrt{\boldsymbol{\Sigma}[i][i]}}\\ &\quad\cdot|F_{P_i}(x) - F_{\hat{P}_i}(x)|\mathrm{d}x\end{aligned}$$

$$\le 2K\sum_{i=1}^{p}\sqrt{\mathbf{\Sigma}[i][i]}\cdot C\mathbb{E}\big|\boldsymbol{\theta}[i]/\sqrt{\Sigma[i][i]}\big|^3$$
$$\le 2\tilde{C}K\sum_{i=1}^{p}(\mathbf{\Sigma}[i][i])^{-1}\cdot\Big(\sum_{t\ge 0}\alpha^3(1-\alpha\lambda_{\min}(\mathbf{H}_{\mathcal{S}}))^{3t}\Gamma\Big)[i]$$
$$\le \frac{2\alpha^2\tilde{C}K\mathbf{\Gamma}}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})}\mathrm{tr}(\mathbf{\Sigma}^{-1}),$$

where F_{P_i} is the cumulative function of $\theta[i]$; the third inequality is obtained by the Berry-Essen inequality.

5.2. Proof of Lemma 2. Let us start with a claim: Suppose the parameter space Θ is compact, $\forall\delta\in(0,1)$, with probability of at least $1-\delta$ over the choice of S, there exists a constant $C(\delta,\Theta)$ such that $\|L_S-L\|_{\mathrm{lip}}\le C(\delta,\Theta)/\sqrt{n}$.

By the CLT, as $n\to\infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta l(f_\theta(x_i),y)-\nabla_\theta L(\theta)\xrightarrow{d}\mathcal{N}(0,\mathrm{Cov}(\nabla_\theta l(f_\theta(x),y))).$$

Hence, by the standard Chebyshev inequality, $\forall\delta\in(0,1)$, with probability of at least $1-\delta$ over the choice of S, we have

$$\sup_{\theta\in\Theta}\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta l(f_\theta(x_i),y)-\nabla_\theta L(\theta)\|^2\le\sup_{\theta\in\Theta}\mathrm{tr}(\mathrm{Cov}(\nabla_\theta l(f_\theta(x),y)))/(\delta n),$$

where Θ is the compact parameter space. Then, the proof is completed by taking

$$C(\delta,\Theta)=2\sqrt{\sup_{\theta\in\Theta}\mathrm{tr}(\mathrm{Cov}(\nabla_\theta l(f_\theta(x),y)))/\delta}.$$

Now let us move on to the proof of Lemma 2:

$$\begin{aligned}&|(L(P)-L_{\mathcal{S}}(P))-(L(\hat{P})-L_{\mathcal{S}}(\hat{P}))|\\&=|\mathbb{E}_{\theta\sim P}(L(\boldsymbol{\theta})-L_{\mathcal{S}}(\boldsymbol{\theta}))-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}}(L(\boldsymbol{\theta})-L_{\mathcal{S}}(\boldsymbol{\theta}))|\\&\le|\mathbb{E}_{\boldsymbol{\theta}\sim P|_{\mathbf{\Theta}_K}}L(\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}|_{\mathbf{\Theta}_K}}L(\boldsymbol{\theta})|+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\mathbf{\Theta}_K}|L(\theta)|\\&\le\rho\mathcal{W}^{(1)}(P|_{\mathbf{\Theta}_K},\hat{P}|_{\mathbf{\Theta}_K})+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\mathbf{\Theta}_K}|L(\boldsymbol{\theta})|\\&\le\frac{2C(\delta)\alpha^2\tilde{C}K\Gamma}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n}}\mathrm{tr}(\Sigma^{-1})+\sup_{\theta\in\Theta}|L(\theta)|\cdot\frac{2p}{K\sqrt{2\pi}}e^{-K^2/2}\\&\triangleq C_1\alpha^2K+C_2\frac{p}{Ke^{K^2/2}},\end{aligned}$$

where

$$C_1\triangleq\frac{2C(\delta)\tilde{C}\Gamma}{3\lambda_{\min}(\mathbf{H}_{\mathcal{S}})\sqrt{n}}\mathrm{tr}(\Sigma^{-1}),$$
$$C_2\triangleq\sup_{\theta\in\Theta}|L(\theta)|\cdot\sqrt{\frac{2}{\pi}}.$$

Let $K\triangleq\sqrt{2\log(\frac{C_2p}{C_1\alpha})}$. We have

$$|(L(P)-L_{\mathcal{S}}(P))-(L(\hat{P})-L_{\mathcal{S}}(\hat{P}))|\le C_1\alpha^2\left(\sqrt{2\log\Big(\frac{C_2p}{C_1\alpha}\Big)}+\sqrt{2\log\Big(\frac{C_2p}{C_1\alpha}\Big)}^{-1}\right).$$

5.3. Proof of Lemma 3. Let $\bar{P}=N(\theta^*,\Sigma)$. By definition,

$$\begin{aligned}&D_{\mathrm{KL}}(\hat{P}\|\sigma(\mathcal{S})^{\perp})\le D_{\mathrm{KL}}(\hat{P}\|\bar{P})\\&=\frac{1}{2}\int_{\theta\in\Theta}-\log\frac{|\Sigma_{\mathcal{S}}|}{|\Sigma|}+(\theta-\theta^*_{\mathcal{S}})^\top(\Sigma^{-1}-\Sigma_{\mathcal{S}}^{-1})(\theta-\theta^*_{\mathcal{S}})\\&\quad+2(\theta-\theta^*_{\mathcal{S}})^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*)+(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*)\mathrm{d}\theta\\&=-\frac{1}{2}\log|\Sigma^{-1}\Sigma_{\mathcal{S}}|+\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_{\mathcal{S}}-I)+\frac{1}{2}(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*).\end{aligned}$$

Let $0<a_*\le a_1\le\cdots\le a_k\le 1\le a_{k+1}\le\cdots\le a_p$ be the eigenvalues of $\mathbf{M}_{\mathcal{S}}\triangleq\Sigma^{-1}\Sigma_{\mathcal{S}}$; thus

$$D_{\mathrm{KL}}(\hat{P}\|\bar{P})=\frac{1}{2}\sum_{i=1}^{p}(-\log a_i+a_i-1)+\frac{1}{2}(\theta^*_{\mathcal{S}}-\theta^*)^\top\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*).$$

Since $-\log(1-x^{1/2})+(1-x^{1/2})-1$ is convex for $x\in(0,(1-a_*)^2)$ and $-\log(1+x^{1/2})+(1+x^{1/2})-1$ is concave for $x>0$,

$$-\log(1-x^{1/2})+(1-x^{1/2})-1<\frac{-\log a_*+a_*-1}{(1-a_*)^2}x,$$

$$
\begin{aligned}
&-\log(1+x^{1/2})+(1+x^{1/2})-1\\
&<\frac{1}{2(1+\sqrt{x_0})}(x-x_0)\\
&\quad-\log(1+\sqrt{x_0})+(1+\sqrt{x_0})-1,
\end{aligned}
$$

where $x_0 = V_2/(p-k)$. Therefore,

$$
\sum_{i=1}^{k} -\log a_i + a_i - 1 \le \frac{-\log a_* + a_* - 1}{(1-a_*)^2} V_1, \tag{10}
$$

$$
\begin{aligned}
\sum_{i=k+1}^{p} -\log a_i + a_i - 1 \le& -(p-k)\log(1+\sqrt{\frac{V_2}{p-k}})\\
&+(p-k)\sqrt{\frac{V_2}{p-k}}\\
\le& V_2,
\end{aligned} \tag{11}
$$

where $V_1 = \sum_{i=1}^{k}(a_i-1)^2, V_2 = \sum_{i=k+1}^{p}(a_i-1)^2$. Combine (10) and (11) to get

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}} D_{\mathrm{KL}}(\hat{P}\|\sigma(\mathcal{S})^{\perp})\\
&\le \frac{1}{2}\max\Big\{\frac{-\log a_* + a_* - 1}{1-a_*}, 1\Big\}M\\
&\quad+\frac{1}{2}(\theta^*_{\mathcal{S}}-\theta^*)^{\top}\Sigma^{-1}(\theta^*_{\mathcal{S}}-\theta^*).
\end{aligned}
$$

The final result follows the Chebyshev's inequality.

5.4. Proof of Proposition 3. (i): Since $\mathbf{u}_t$ is uniformly bounded, $\exists \mathbf{C} \in \mathbb{R}^{p\times p}, \mathbf{C} \succ 0$ such that $\mathrm{Cov}(\mathbf{u}_t) \prec \mathbf{C}$ holds for any t. Then we have

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{\theta}_T) &= \sum_{t=0}^{T-1}\alpha^2(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^{T-t-1}\\
&\quad\times \mathrm{Cov}(\mathbf{u}_t)(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^{T-t-1}\\
&\le T\alpha^2\mathbf{C} = \mathcal{O}(T\alpha^2).
\end{aligned}
$$

(ii): Let ϕ_x be the characteristic function of $\boldsymbol{x}$. Thus

$$
\begin{aligned}
&\phi_{\boldsymbol{\theta}_T-\mathbb{E}[\boldsymbol{\theta}_T]}(\mathbf{s})\\
&\quad=\prod_{t=0}^{T-1}\phi_{\boldsymbol{u}_t}(\alpha(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s})\\
&\quad=\prod_{t=0}^{T-1}(1-\alpha^2\mathbf{s}^{\top}(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\\
&\qquad\times\mathrm{Cov}(\mathbf{u}_t)(\mathbf{I}-\alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{s}\\
&\qquad+o(\alpha^2\|\mathbf{s}\|_2^2))\\
&\quad=1-\mathbf{s}^{\top}\mathrm{Cov}(\boldsymbol{\theta}_T)\mathbf{s}+o(\|\mathbf{s}\|_2^2\alpha^2),
\end{aligned}
$$

By the proof of (i), $\phi_{\boldsymbol{\theta}_\infty}(\mathbf{s}) \to 1-\mathbf{s}^{\top}\mathrm{Cov}(\boldsymbol{\theta}_\infty)\mathbf{s}$ as $\max\alpha_t \to 0$; thus $(\sum_{t=0}^{T-1}\alpha_t^2)^{-1/2}(P(\alpha)-\hat{P}(\alpha)) \overset{\text{law}}{\to} \mathbf{0}$.

(iii): Without loss of generality, assume that the eigenvector direction of $\mathbf{H}_{\mathcal{S}}$ is consistent with the coordinate axis. Set event $A = \{\boldsymbol{\theta} \mid |\boldsymbol{\theta}[i]-\boldsymbol{\theta}^*_{\mathcal{S}}[i]| \le K\sqrt{\boldsymbol{\Sigma}[i][i]}, i=1,\dots,p\}$.We have

$$
\begin{aligned}
&\mathcal{W}^{(1)}(P|_{\Theta_K},\hat{P}|_{\Theta_K})\\
&\quad=\inf_{F_{\boldsymbol{\theta}_1}=F_{P|_{\Theta_K}},F_{\boldsymbol{\theta}_2}=F_{\hat{P}|_{\Theta_K}}}\mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_1\\
&\quad\le\inf_{F_{\boldsymbol{\theta}_1}=F_P,F_{\boldsymbol{\theta}_2}=F_{\hat{P}}}\mathbb{E}_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}[\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_1\\
&\qquad\cdot\chi_A(\boldsymbol{\theta}_1)\cdot\chi_A(\boldsymbol{\theta}_2)]\\
&\quad\le\inf_{F_{\theta_1}=F_P,F_{\theta_2}=F_{\hat{P}}}\sum_{i=1}^{p}\int_{\theta^*_{\mathcal{S}}[i]-K\sqrt{\boldsymbol{\Sigma}[i][i]}}^{\theta^*_{\mathcal{S}}[i]+K\sqrt{\boldsymbol{\Sigma}[i][i]}}\\
&\qquad\cdot|F_{P_i}(x)-F_{\hat{P}_i}(x)|\mathrm{d}x\\
&\quad\le 2K\sum_{i=1}^{p}\sqrt{\boldsymbol{\Sigma}[i][i]}\cdot\tilde{C}\mathbb{E}\big|\boldsymbol{\theta}[i]/\sqrt{\Sigma[i][i]}\big|^3\\
&\quad\le 2\tilde{C}K\Big(\sum_{i=1}^{q}(\boldsymbol{\Sigma}[i][i])^{-1}\\
&\qquad\cdot\Big(\sum_{t=0}^{T-1}\alpha^3(1-\alpha\tilde{\lambda}_{\min}(\mathbf{H}_{\mathcal{S}}))^{3t}\Gamma\Big)[i]\\
&\qquad+\sum_{i=q+1}^{p}(\boldsymbol{\Sigma}[i][i])^{-1}\cdot\Big(\sum_{t=0}^{T-1}\alpha^3\mathbb{E}|\boldsymbol{u}_t[i]|^3\Big)\Big)\\
&\quad\le\tilde{C}'K\Big(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{\min}}+\frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\Big),
\end{aligned}
$$

where the third inequality is obtained by the Berry-Essen inequality.

5.5. Proof of Lemma 4. We have

$$
\begin{aligned}
&|(L(P)-L_{\mathcal{S}}(P))-(L(\hat{P})-L_{\mathcal{S}}(\hat{P}))|\\
&=|\mathbb{E}_{\theta\sim P}L(\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}}L(\boldsymbol{\theta})|\\
&\le|\mathbb{E}_{\boldsymbol{\theta}\sim P|_{\boldsymbol{\Theta}_K}}L(\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{\theta}\sim\hat{P}|_{\boldsymbol{\Theta}_K}}L(\boldsymbol{\theta})|\\
&\quad+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K}|L(\theta)|\\
&\le\frac{C(\delta)}{\sqrt{n}}\mathcal{W}^{(1)}(P|_{\boldsymbol{\Theta}_K},\hat{P}|_{\boldsymbol{\Theta}_K})\\
&\quad+\max\{P(A^c),\hat{P}(A^c)\}\cdot\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_K}|L(\boldsymbol{\theta})|\\
&\le\frac{2}{\sqrt{n}}C(\delta)\tilde{C}K\Big(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{\min}}+\frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\Big)\\
&\quad+\sup_{\theta\in\Theta}|L(\theta)|\cdot\frac{2p}{K\sqrt{2\pi}}e^{-K^2/2}\\
&\triangleq C_1K+C_2\frac{p}{Ke^{K^2/2}},
\end{aligned}
$$

where

$$
C_1 \triangleq \frac{2}{\sqrt{n}}C(\delta)\tilde{C}K\Big(\frac{\alpha\boldsymbol{\Gamma}}{3\tilde{\lambda}_{\min}}+\frac{\sum_{i=1}^{T}\alpha_t^3}{\sum_{i=1}^{T}\alpha_t^2}\Big),
$$

$$C_2 \triangleq \sup_{\theta\in\Theta} |L(\theta)| \cdot \sqrt{\frac{2}{\pi}}.$$

Let $K \triangleq \sqrt{2\log(\frac{C_2 p}{C_1})}$. We have

$$|(L(P) - L_{\mathcal{S}}(P)) - (L(\hat{P}) - L_{\mathcal{S}}(\hat{P}))| \leq C_1(\sqrt{2\log\Big(\frac{C_2 p}{C_1}\Big)} + \sqrt{2\log\Big(\frac{C_2 p}{C_1}\Big)}^{-1}).$$

5.6. Proof of Proposition 5. (i) Additive noise insertion: By substituting $\mathbf{u}_t$ in the proof of Proposition 1 with $\mathbf{u} - \boldsymbol{\eta}_t$, our conclusion directly follows $\mathrm{Cov}(\mathbf{u} - \boldsymbol{\eta}_t) = \mathrm{Cov}(\mathbf{u}) + \mathrm{Var}(\boldsymbol{\eta}_0[1])\mathbf{I}$.

(ii) Multiplicative noise insertion: The dynamics of SGD with multiplicative noise are

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}^{(t)} \odot g_{B_t} \\ &= (I - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(t)})\boldsymbol{\theta}_t - \alpha\boldsymbol{\gamma}^{(t)} \odot \mathbf{u}_t.\end{aligned}$$

Thus,

$$\boldsymbol{\theta}_T = \sum_{t=0}^{T-1}\prod_{i=t+1}^{T-1}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(i)}) \cdot \alpha\boldsymbol{\gamma}^{(i)} \odot \mathbf{u}_t + \prod_{t=1}^{T}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(t)})\theta_0.$$

By taking the covariance of $\boldsymbol{\theta}_T$, we have

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\theta}_T) =& \mathbb{E}_{\boldsymbol{\gamma}_T,\mathbf{u}}[(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(T)})\mathrm{Cov}(\boldsymbol{\theta}_{T-1}) \\ & (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}} \odot \boldsymbol{\gamma}^{(T)})] + \mathrm{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) \\ =& (\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\mathrm{Cov}(\boldsymbol{\theta}_{T-1})(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}}) \\ & + \mathrm{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) + \mathcal{O}(\alpha^2\mathrm{Cov}(\boldsymbol{\theta}_{T-1})).\end{aligned}$$

Thus,

$$\begin{aligned}&\lim_{\alpha\to 0}\alpha^{-1}\mathrm{Cov}(\boldsymbol{\theta}_T) \\ &= \lim_{\alpha\to 0}\alpha^{-1}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})\mathrm{Cov}(\boldsymbol{\theta}_{T-1})(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}}) \\ &\quad + \alpha^{-1}\mathrm{Cov}(\alpha\boldsymbol{\gamma}_T \odot \mathbf{u}_t) \\ &= \lim_{\alpha\to 0}\alpha\sum_{t=0}^{T}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}'(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t,\end{aligned}$$

where $\mathbf{C}' = (\mathbf{C} + (\mathbb{E}\boldsymbol{\gamma}_0[1]^2 - 1)\mathrm{diag}(\mathbf{C}))$. Setting $T = \infty$, we have

$$\lim_{\alpha\to 0}\alpha^{-1}\mathrm{Cov}(\boldsymbol{\theta}'_\infty) = \alpha\sum_{t\geq 0}(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t\mathbf{C}'(\mathbf{I} - \alpha\mathbf{H}_{\mathcal{S}})^t.$$

6. Empirical evidence

In Proposition 1, given that gradient noise is generally non-Gaussian, we have proven that the limiting distribution of $\mathcal{U}$-SGD is Gaussian-like. To validate these statements from an empirical perspective, we conduct systematic numerical experiments to examine the Gaussianity of limiting distributions of non-Gaussian SGDs. Numerical results suggest that, for convex loss functions, SGD with non-Gaussian gradient noise yields a Gaussian limiting parameter distribution. Visualization shows that the limiting distribution becomes more Gaussian-like as the learning rate gets smaller. Moreover, such Gaussianity is observed to hold for more general cases.

Experimental settings. To visualize the limiting parameter distribution, we consider three instances of $\mathcal{U}$-SGD on three loss functions on $\mathbb{R}^2$. The loss function $f(\cdot)$ is chosen from

$$\begin{aligned}f_1(\boldsymbol{\theta}) &\triangleq \frac{1}{2}\boldsymbol{\theta}^\top\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\boldsymbol{\theta}, \\ f_2(\boldsymbol{\theta}) &= \boldsymbol{\theta}[1]\log(\boldsymbol{\theta}[1]) + (1,1)^\top\boldsymbol{\theta} \cdot \log((1,1)^\top\boldsymbol{\theta}), \\ f_3(\boldsymbol{\theta}) &= f_1(\boldsymbol{\theta})\log(f_1(\boldsymbol{\theta})).\end{aligned}$$

The learning rate α takes values from $\{0.1, 0.01, 0.001\}$. The gradient noise $\{\mathbf{u}_t\}$ is chosen from $\{\{\mathbf{u}_t^{(i)}\}\}_{i=1}^3$. For different choices of α, $\{\mathbf{u}_t^{(i)}\}$ are i.i.d., and they are generated as follows:

$$\begin{aligned}\{\mathbf{u}_0^{(1)}[1], \mathbf{u}_0^{(1)}[2]\} &\overset{\text{i.i.d.}}{\sim} 0.01 \cdot \mathrm{U}(-0.5, 0.5), \\ \{\mathbf{u}_0^{(2)}[1], \mathbf{u}_0^{(2)}[2]\} &\overset{\text{i.i.d.}}{\sim} (\mathrm{Exp}(100) - 0.01), \\ \{\mathbf{u}_0^{(3)}[1], \mathbf{u}_0^{(3)}[2]\} &\overset{\text{i.i.d.}}{\sim} 0.01 \cdot (\mathrm{Bin}(4, 0.5) - 2).\end{aligned}$$

$\mathrm{U}(-0.5, 0.5)$ denotes the uniform distribution on $(-0.5, 0.5)$, $\mathrm{Exp}(100)$ denotes the exponential distribution with a rate parameter equal to 100, and $\mathrm{Bin}(4, 0.5)$ denotes the binomial distribution with the number of trials equal to 4 and the probability of success equal to 0.5. The mean and covariance of these distributions are $\mathbf{0}$ and $0.01^2\mathbf{I}$.

These distributions are chosen to represent bounded and well-behaved distributions, exponential-tailed continuous distributions, and discrete distributions, respectively.

In all these cases, we ran the following experiments. In each episode, the parameter is initialized at $\boldsymbol{\theta}_0 = (1,1)^\top$. In the t-th iterations, we update the parameter as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\nabla l(\boldsymbol{\theta}_t)$. After running T iterations (we set $T = 10^4$), we collect $\boldsymbol{\theta}_T$. We run 10^4 episodes to obtain $\{\boldsymbol{\theta}_T^i\}_{i=1}^{10^4}$, a sample set of the limiting parameter distribution. Then, we visualize the empirical limiting distribution and perform a Henze-Zirkler multivariate

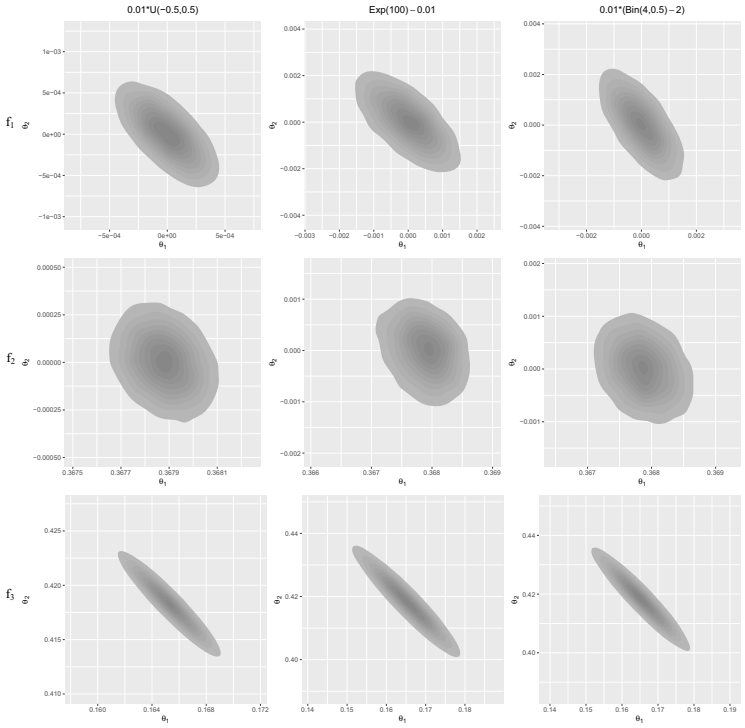

Fig. 1. For each gradient noise implementation (including adding uniform, exponential and binomial gradient noise) and each loss function f_1, f_2, f_3, experiments are run with $\alpha = 0.1$ and $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution with a 2D-kernel density plot. The scatter plots are contained in Appendix.

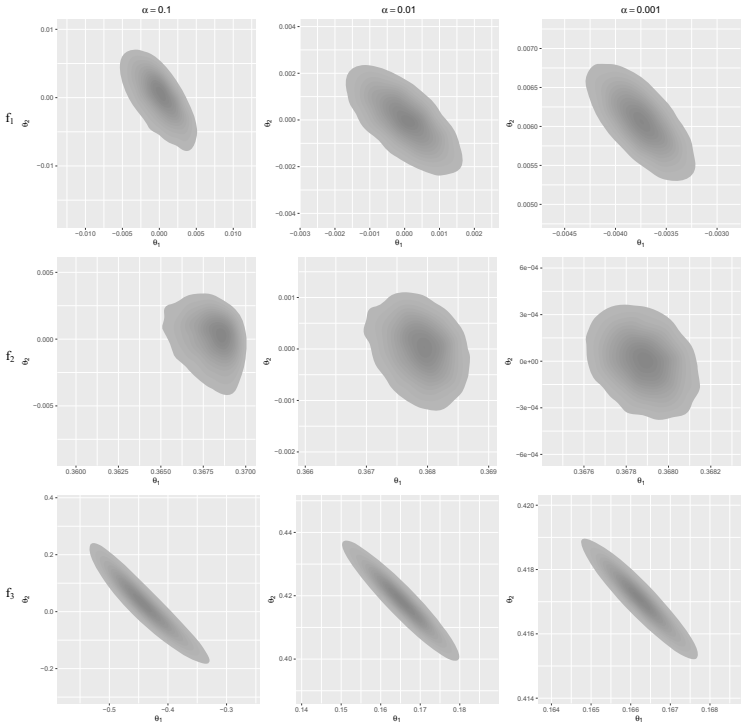

Fig. 2. The gradient noise is fixed to be exponential. For each loss function f_1, f_2, f_2, the experiments are run with $\alpha \in \{0.1, 0.01, 0.001\}$ and a fixed initial value $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution with a 2D-kernel density plot. The scatter plots are contained in the Appendix.

normality test to examine whether $\{\boldsymbol{\theta}_T^i\}_{i=1}^{10^4}$ follows a two-dimensional Gaussian distribution. The whole process is repeated 30 times. Similar procedures are conducted on two neural networks and different datasets, namely MNIST and CIFAR-10 (LeCun and Cortes, 2010; Krizhevsky and Hinton, 2009). The optimizer is the standard SGD without weight decay or momentum. We train from a fixed initialization until a high, stable training accuracy is achieved, and then we collect the model parameters. This process is repeated 3×10^3 times to get $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{3\times 10^3}$. To check the marginal-Gaussianity of $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{3000}$, we perform a Pearson's Gaussian (D'Agostino and Pearson, 1973) test on the projections of each dimension. For marginals with p-values lower than 0.01, we reject the null hypothesis that the marginal parameters are Gaussian at a confidence level of 99%; otherwise, we accept the null hypothesis. Implementation details are included in the supplementary material.

Numerical results. Visualization evidence coincides with our theoretical findings. As shown in Fig. 1, for convex loss functions (not necessarily quadratic), non-Gaussian SGD with a fixed initialization still leads to a Gaussian-like limiting distribution, which disperses around the minima. SGDs with uniformly distributed and binomially distributed gradient noise belong to the $\mathcal{U}$-SGD family, and they result in a visually more Gaussian limiting distribution than the exponentially noisy SGD does. Figure 2 shows, for exponentially noisy SGD, that the limiting distribution tends toward a Gaussian distribution as α gets smaller. In conclusion, our statement that 'non-Gaussian SGD has a Gaussian-like limiting distribution' is extended to more general cases, where the loss functions are not necessarily quadratic and gradient noise is exponential-tailed. In all cases, the p-values of the Henze-Zirkler multivariate normality test suggest that there is no statistically significant evidence against the null hypothesis that the limiting distribution is Gaussian. To further examine the two-dimensional Gaussianity of the limiting distribution, the aforementioned procedures with a random initialization $\{\boldsymbol{\theta}_0[1], \boldsymbol{\theta}_0[2]\} \overset{\text{i.i.d.}}{\sim} \mathrm{U}(0,1)$ are repeated 30 times. For each initialization, we perform the Henze-Zirkler multivariate normality test on the limiting distributions. We then collect the p-values of each repetition. As we can see in Fig. 3, there is no statistically significant evidence against the null hypothesis that the limiting distribution is Gaussian.

For neural network cases, a representative result is displayed in Fig. 4. Clearly, the p-values of each marginal are nearly uniformly distributed on $[0, 1]$. About 6.7% (MNIST) and 1.5% (CIFAR-10) of the marginal p-values are lower than 0.01. This evidence shows that the marginal-Gaussianity holds for most of the dimensions

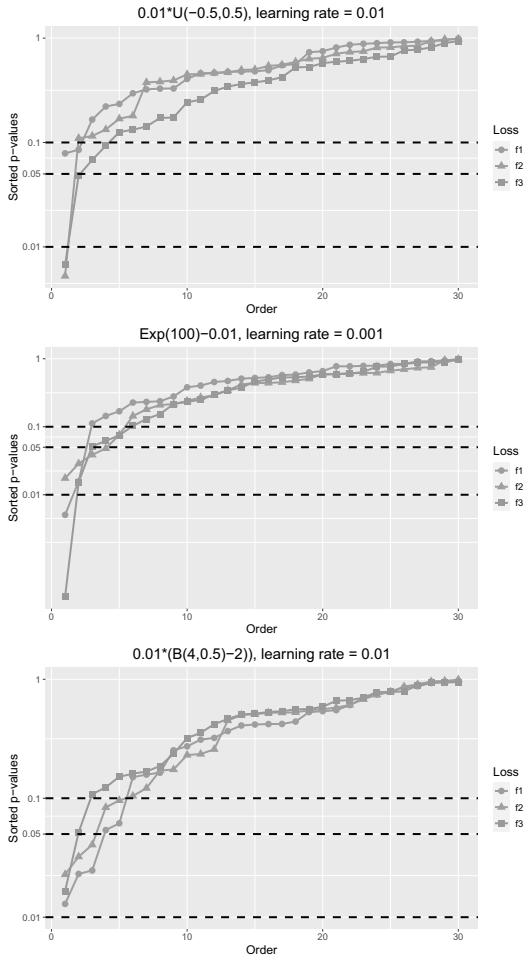

Fig. 3. For loss functions f_1, f_2, f_3, we perform SGDs with uniformly, exponentially and binomially distributed gradient noise, and set $\alpha = 0.01$. At a confidence level of 0.99, about $29/30$ out of 30 repetitions fail to provide statistically significant evidence against the two-dimensional Gaussianity of the limiting parameter distributions.

and strongly suggests that the limited distributions of parameters are Gaussian-like. More experiments with randomized initialization along with the table of p-values are reported in Appendix.

7. Conclusions

In this paper, we propose a novel Gaussian approximation framework to develop generalization error bounds for the $\mathcal{U}$-SGD family. Our general process is two-fold. We prove that the limiting parameter distribution $P(\alpha)$ tends to be Gaussian as $\alpha \to 0$, even when the gradient noise is non-Gaussian. This result is numerically validated to hold for more general cases, and it enables us to establish a Gaussian approximation $\hat{P}(\alpha)$ with an $\mathcal{O}(\alpha)$ approximation error. Then, we bound the deviation term (3) and the KL-divergence term (4), respectively. The

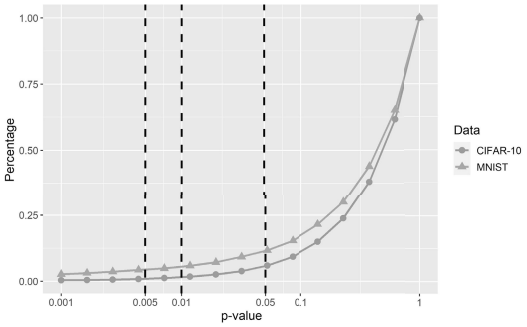

Fig. 4. For a given threshold (horizontal axis), we calculate the percentage (vertical axis) of marginals with p-values smaller than the threshold. The horizontal axis of the lower figure is log-scaled.

combination of these results leads to a tight generalization error bound at an order of $\mathcal{O}((1+\sqrt{\log(p\sqrt{n})})/\sqrt{n})$. On top of that, we propose additive and multiplicative noise insertion methods to improve the generalization performance. Admittedly, a grossly small learning rate greatly loosens the KL-divergence bound (4). In future works, we will attempt to refine the generalization error bound and develop hybrid noise insertion methods.

References

Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R.R. and Wang, R. (2019). On exact computation with an infinitely wide neural net, *in* H. Wallach *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, New York, pp. 8141–8150.

Bartlett, P.L., Foster, D.J. and Telgarsky, M.J. (2017). Spectrally-normalized margin bounds for neural networks, *in* I. Guyon *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, New York, pp. 6240–6249.

Bottou, L. (1998). On-line learning and stochastic approximations, *in* L. Bottou (Ed) *On-line Learning in Neural Networks*, Cambridge University Press, pp. 9–42.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization, *Journal of Machine Learning Research* **2**: 499–526, DOI: 10.1162/153244302760200704.

Chen, H., Mo, Z., Yang, Z. and Wang, X. (2019). Theoretical investigation of generalization bound for residual networks, *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Francisco*, pp. 2081–2087, DOI: 10.24963/ijcai.2019/288.

Chen, Y., Chang, H., Meng, J. and Zhang, D. (2019). Ensemble neural networks (enn): A gradient-free stochastic method, *Neural Networks* **110**: 170–185.

D'Agostino, R. and Pearson, E.S. (1973). Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b_1}$, *Biometrika* **60**(3): 613–622.

Dieuleveut, A., Durmus, A. and Bach, F. (2017). Bridging the gap between constant step size stochastic gradient descent and markov chains, *arXiv:* 1707.06386.

Elisseeff, A., Evgeniou, T. and Pontil, M. (2005). Stability of randomized learning algorithms, *Journal of Machine Learning Research* **6**: 55–79.

Feng, Y., Gao, T., Li, L., Liu, J.-G. and Lu, Y. (2020). Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation, *Communications in Mathematical Sciences* **18**(1).

Hardt, M., Recht, B. and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent, *in* M.F. Balcan and K.Q. Weinberger (Eds), *Proceedings of The 33rd International Conference on Machine Learning, New York, USA,* pp. 1225–1234.

He, F., Liu, T. and Tao, D. (2019). Control batch size and learning rate to generalize well: Theoretical and empirical evidence, *in* H. Wallach *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, New York, pp. 1143–1152.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Poceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H. and Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, DOI:10.1109/CVPR.2018.00286.

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

Krizhevsky, A. and Hinton, G. (2009). *Learning Multiple Layers of Features from Tiny Images*, Master's thesis, Department of Computer Science, University of Toronto.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database, `http://yann.lecun.com/exdb/mnist/`.

Li, H., Xu, Z., Taylor, G., Studer, C. and Goldstein, T. (2018). Visualizing the loss landscape of neural nets, *in* S. Bengio *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, New York, pp. 6389–6399.

Li, J., Luo, X. and Qiao, M. (2020). On generalization error bounds of noisy gradient methods for non-convex learning, *arXiv*: 1902.00621.

Li, Q., Tai, C. and Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations, *Journal of Machine Learning Research* **20**(40): 1–47.

Li, X., Lu, J., Wang, Z., Haupt, J. and Zhao, T. (2019). On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond, *arXiv*: 1806.05159.

Ljung, L., Pflug, G. and Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems*, Birkhäuser, Basel, Switzerland.

London, B., Huang, B., Taskar, B. and Getoor, L. (2014). PAC-Bayesian Collective Stability, *in* S. Kaski and J. Corander (Eds), *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland*, pp. 585–594.

Mandt, S., Hoffman, M.D. and Blei, D.M. (2017). Stochastic gradient descent as approximate Bayesian inference, *Journal of Machine Learning Research* **18**(1): 4873–4907.

McAllester, D.A. (1999). PAC-Bayesian model averaging, *Proceedings of the 12th Annual Conference on Computational Learning Theory, COLT '99, New York, NY, USA*, pp. 164–170, DOI: 10.1145/307400.307435.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A. and Roy, D. M. (2019). Information-theoretic generalization bounds for sgld via data-dependent estimates, *in* H. Wallach *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, New York, pp. 11015–11025.

Panigrahi, A., Somani, R., Goyal, N. and Netrapalli, P. (2019). Non-gaussianity of stochastic gradient noise, *arXiv*: 1910.09626.

Qian, Y., Wang, Y., Wang, B., Gu, Z., Guo, Y. and Swaileh, W. (2022). Hessian-free second-order adversarial examples for adversarial learning, *arXiv:* 2207.01396.

Rong, Y., Huang, W., Xu, T. and Huang, J. (2020). DropEdge: Towards deep graph convolutional networks on node classification, *arXiv*: 1907.10903.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, International Conference on Learning Representations 2020, `https://openreview.net/pdf?id=Hkx1qkrKPr`.

Simsekli, U., Sagun, L. and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks, *in* K. Chaudhuri and R. Salakhutdinov (Eds), *Proceedings of the 36th International Conference on Machine Learning*, New York, pp. 5827–5837.

Sulaiman, I.M., Kaelo, P., Khalid, R. and Nawawi, M.K.M. (2024). A descent generalized rmil spectral gradient algorithm for optimization problems, *Internationl Jouranl of Applied Mathematics and Computer Science* **34**(2): 225–233, DOI: 10.61822/amcs-2024-0016.

Sutskever, I., Martens, J., Dahl, G. and Hinton, G. (2013). On the importance of initialization and momentum in deep learning, *in* S. Dasgupta and D. McAllester (Eds), *Proceedings of the 30th International Conference on Machine Learning, Atlanta, USA,* pp. 1139–1147.

Villani, C. (2008). *Optimal Transport: Old and New*, Grundlehren der Mathematischen Wissenschaften, Springer, Berlin/Heidelberg.

Weinan, E., Ma, C. and Wang, Q. (2019). A priori estimates of the population risk for residual networks, *arXiv:* 1903.02154.

Welling, M. and Teh, Y.W. (2011). Bayesian learning via stochastic gradient Langevin dynamics, *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, Madison, WI, USA*, p. 681–688.

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V. and Zhu, Z. (2020). On the noisy gradient descent that generalizes as SGD, *International Conference on Machine Learning,* PMLR 119: 10367–10376.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization, *arXiv*: 1611.03530.

Hao Chen is currently a PhD student in the School of Big Data at the University of Science and Technology of China (USTC). He received his BS degree in mathematics from USTC in 2017. His main research interests include machine learning theory and its applications.

Zhanfeng Mo is currently a PhD student in the School of Computer Science and Engineering of Nanyang Technological University, Singapore. He received his BS degree in statistics from University of Science and Technology of China in 2020. His main research interests include theory and algorithms of statistical machine learning.

Zhouwang Yang is a professor in the School of Mathematical Sciences at the University of Science and Technology of China (USTC). He received his BS, MS degree and PhD degrees in mathematics from USTC in 1997, 2000 and 2005, respectively. He worked at Seoul National University as a postdoctoral researcher from 2006 to 2007. He was a visiting scholar in the School of Industrial and Systems Engineering (ISyE) at the Georgia Institute of Technology from 2010–2011. His main research interests include geometric modeling and processing, data-driven optimization modeling, and the mathematical theory of machine learning.

Appendix
Details of experiments

Experimental settings. Our experiments on neural networks are conducted on different models and different datasets, namely MNIST (LeCun and Cortes, 2010) and CIFAR-10 (Krizhevsky and Hinton, 2009). On the MNIST dataset, we train a three-layer network (Model 1) with (784×200 FC)-ReLU-(200×200 FC)-ReLU-(200×10 FC), where FC denotes a fully connected layer. We use the optimizer of SGD with batch_size $= 200$ and learning_rate $= 0.01$ for the network. For the CIFAR-10 dataset, we use a convolution network (Model 2) with ($3 \times 6\, 5 \times 5$C)-ReLU-MP2-($6 \times 16\, 5 \times 5$C)-ReLU-MP2-(400×120 FC)-ReLU-(120×84 FC)-ReLU-(84×10 FC), where (5×5C) denotes a 5×5 convolution layer and MP2 denotes a 2×2 max pooling layer. The optimizer of SGD is used again but the settings changes to batch_size $= 4$ and learning_rate $= 0.001$. Experiments are executed as follows:

1. Initialize the model at a fixed point in the vicinity of the optima. In each experiments, we get this fixed point by training 5 epochs on Model 1 and 10 epochs in Model 2 with a Xavier and Kaiming initialization (He *et al.*, 2015).

Table A1. For each experiment, we calculate the percentages of dimensions with marginal p-values smaller than $0.1, 0.05$ and 0.01, respectively. For a marginal with a p-value smaller than $\delta \in (0, 1)$, we can reject the null hypothesis that this marginal follows a Gaussian distribution at a confidence level of $1 - \delta$.

Percentage	≤ 0.1	≤ 0.05	≤ 0.01
MNIST Exp. 1	10.8%	5.7%	1.5%
MNIST Exp. 2	11.8%	6.9%	2.6%
MNIST Exp. 3	12.3%	7.2%	2.7%
MNIST Exp. 4	12.6%	7.5%	3.2%
CIFAR-10 Exp. 1	8.3%	3.9%	0.3%
CIFAR-10 Exp. 2	8.8%	4.1%	0.3%
CIFAR-10 Exp. 3	10.1%	4.4%	0.4%
CIFAR-10 Exp. 4	10.2%	4.4%	0.4%

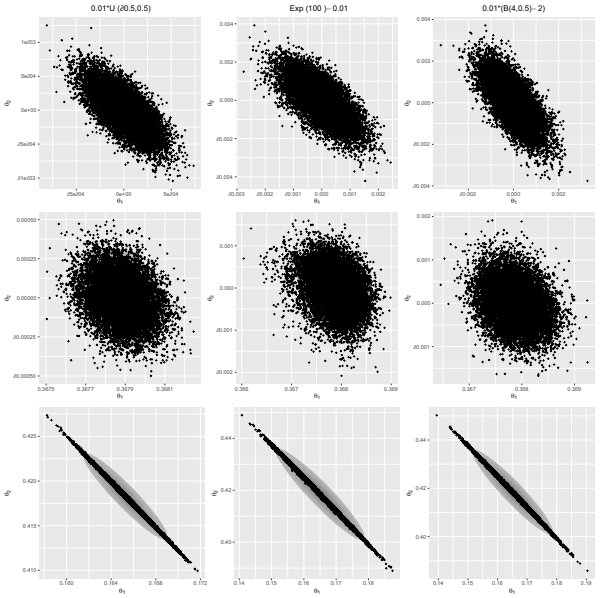

Fig. A1. For each gradient noise implements (including adding uniform, exponential and binomial gradient noise) and each loss function f_1, f_2 and f_3, experiments are run with $\alpha = 0.01$ and $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution by a 2D-kernel density plot.

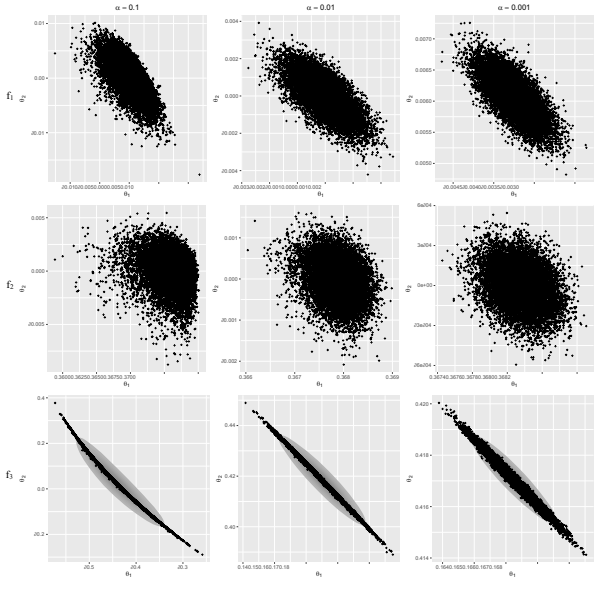

Fig. A2. The gradient noise is fixed to be exponential. For each loss function f_1, f_2, f_2, the experiments are run with $\alpha \in \{0.1, 0.01, 0.001\}$ and a fixed initialization $\boldsymbol{\theta}_0 = (1, 1)^\top$. We visualize the empirical limiting distribution by a 2D-kernel density plot.

2. Train the models until the training loss and accuracy are stable. We train 30 epochs on Model 1 and 50 epochs on Model 2.
3. Repeat the second step for 3000 times and collect the parameters of the final epochs. We obtain $\{\boldsymbol{\theta}^{(i)}_{\text{MNIST}}\}_{i=1}^{3000}, \{\boldsymbol{\theta}^{(i)}_{\text{CIFAR10}}\}_{i=1}^{3000}$.
4. Take MNIST for example; for each marginal $j = 1, \ldots, p_{\text{MNIST}}$ with $p_{\text{MNIST}} = 198800$, we perform the Person test on $\{\boldsymbol{\theta}^{(i)}_{\text{MNIST}}[j]\}_{i=1}^{3000}$ to check where marginal-Gaussianity holds for the j-th dimension. This results in 198800 marginal p-values. At a confidence level of $1 - \delta$, we reject the null hypothesis that the j-th marginal is Gaussian if the corresponding p-value is less than δ. The same procedures are conducted on CIFAR-10.

Experimental results. For a given threshold, we calculate the percentage of marginals with p-values smaller than the threshold. The horizontal axis of the lower figure is log-scaled. Table A1 shows that the marginal-Gaussianity holds for most of the dimensions and strongly suggests that the limited distributions of parameters are Gaussian-like.

Received: 3 June 2024
Revised: 26 August 2024
Accepted: 16 October 2024