Machine Learning: Science and Technology

ACCEPTED MANUSCRIPT • OPEN ACCESS

Stochastic Gradient Descent with Random Label Noises: Doubly Stochastic Models and Inference Stabilizer

To cite this article before publication: Haoyi Xiong *et al* 2023 *Mach. Learn.: Sci. Technol.* in press https://doi.org/10.1088/2632-2153/ad13ba

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2023 The Author(s). Published by IOP Publishing Ltd**.

As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by/4.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 152.78.1.77 on 22/01/2024 at 13:15

Stochastic Gradient Descent with Random Label Noises: Doubly Stochastic Models and Inference Stabilizer

Haoyi Xiong[1], Xuhong Li[1], Boyang Yu[1], Dongrui Wu[2], Zhanxing Zhu[3] and Dejing Dou[1]

[1] Baidu Inc., Beijing, China
[2] Huazhong University of Science and Technology, Wuhan, China
[3] Peking University, Beijing, China
Corresponding Author: Haoyi Xiong (xhyccc@gmail.com)

Abstract. Random label noises (or observational noises) widely exist in practical machine learning settings. While previous studies primarily focus on the affects of label noises to the performance of learning, our work intends to investigate the implicit regularization effects of the label noises, under mini-batch sampling settings of stochastic gradient descent (SGD), with assumptions that label noises are unbiased. Specifically, we analyze the learning dynamics of SGD over the quadratic loss with unbiased label noises, where we model the dynamics of SGD as a stochastic differentiable equation (SDE) with two diffusion terms (namely a *Doubly Stochastic Model*). While the first diffusion term is caused by mini-batch sampling over the (label-noiseless) loss gradients, as in many other works on SGD [1, 2], our model investigates the second noise term of SGD dynamics, which is caused by mini-batch sampling over the label noises, as an implicit regularizer. Our theoretical analysis finds such implicit regularizer would favor some convergence points that could stabilize model outputs against perturbation of parameters (namely *inference stability*). Though similar phenomenon have been investigated by Blanc *et al.* [3], our work doesn't assume SGD as an Ornstein-Uhlenbeck like process and achieve a more generalizable result with convergence of approximation proved.

To validate our analysis, we design two sets of empirical studies to analyze the implicit regularizer of SGD with unbiased random label noises for deep neural networks training and linear regression. Our first experiment studies *the noisy self-distillation tricks* for deep learning, where student networks are trained using the outputs from well-trained teachers with additive unbiased random label noises. Our experiment shows that the implicit regularizer caused by the label noises tend to select models with improved inference stability. We also carry out experiments on SGD-based linear regression with unbiased label noises, where we plot the trajectories of parameters learned in every step and visualize the effects of implicit regularization. The results backup our theoretical findings.

Accepted Manuscript

Ethical Statements

None procedures performed in studies involving human participants or animals. All authors have made contribution to the manuscript. Dr. Haoyi Xiong contributed in the original research ideas, formulated the research problems, proposed part of algorithms, wrote the manuscript and involved in discussions. Dr. Xuhong Li contributed in the original research ideas, conducted part of experiments, and wrote part o the manuscript. Mr. Boyang Li helped in part of mathematical proofs. Dongrui Wu and Zhanxing Zhu involved in discussions. Prof. Dejing Dou oversaw the progress of research. All procedures were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

1. Introduction

Stochastic Gradient Descent (SGD) has been widely used as an effective way to train deep neural networks with large datasets [4]. While the mini-batch sampling strategy was firstly proposed to lower the cost of computation per iteration, it has been considered to incorporate an implicit regularizer preventing the learning process from converging to the local minima with poor generalization performance [5, 1, 6, 7, 8]. To interpret such implicit regularization, one can model SGD as gradient descent (GD) with gradient noises caused by mini-batch sampling [9]. Studies have demonstrated the potentials of such implicit regularization or gradient noises to improve the generalization performance of learning from both theoretical [10, 11, 12, 13] and empirical aspects [1, 7, 8]. In summary, gradient noises keep SGD away from converging to the sharp local minima that generalizes poorly [1, 12, 13] and would select a flat minima [14] as the outcome of learning.

In this work, we aim at investigating the influence of random label noises to the implicit regularization under mini-batch sampling of SGD. To simplify our research, we assume the training dataset as a set of vectors $\mathcal{D} = \{x_1, x_2, x_3, \ldots, x_N\}$.The label $\tilde{y}_i$ for every vector $x_i \in \mathcal{D}$ is the noisy response of the true neural network $f^*(x)$ such that

$$\tilde{y}_i = y_i + \varepsilon_i,\ y_i = f^*(x_i),\ \text{and } \mathbb{E}[\varepsilon_i] = 0,\ \text{var}[\varepsilon_i] = \sigma^2\ , \tag{1}$$

where the label noise ε_i is assumed to be an independent zero-mean **random variable**. In our work, the random label noises can be either (1) drawn from probability distributions before training steps (but re-sampled by mini-batch sampling of SGD) or (2) realized by the random variables per training iteration [15]. Thus learning is to estimate $\widehat{\theta}$ in $f(x, \widehat{\theta})$ for approximating $f^*(x)$, such that

$$\widehat{\theta} \leftarrow \underset{\forall \theta \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \tilde{L}_i(\theta) := \frac{1}{2N} \sum_{i=1}^{N} \left(f(x_i, \theta) - \tilde{y}_i \right)^2 \right\}. \tag{2}$$

Note that we denote $L_i^*(\theta) = \frac{1}{2}(f(x_i, \theta) - y_i)^2$ as the loss based on a noiseless sample in this work. Inspired by [14, 1], our work studies how unbiased label noises ε_i $(1 \leq i \leq N)$

Accepted Manuscript

Symbols	Definitions and Equations	
x_i, and $y_i = f^*(x_i)$	the ith data point and true label	(1)
$\tilde{y}_i$, and ε_i	the ith noisy label and the label noise.	(1)
$f(x, \theta)$	the output of neural network with parameter θ and input x.	(2)
$\widehat{\theta}$	the estimator of parameters of a neural network.	(2)
L_i^*	$L_i(\theta) = \frac{1}{2}(f(x_i, \theta) - y_i)^2$ the loss based on a noiseless sample.	(2)
$\tilde{L}_i$	$\tilde{L}_i(\theta) = \frac{1}{2}(f(x_i, \theta) - \tilde{y}_i)^2$ the loss based on a noisy sample.	(2)
	SGD without assumptions on label noises	
B_k	the mini-batch of samples drawn by the kth step of SGD	(3)
$b = \lvert B_k \rvert$	the constant batch size of B_k	(3)
θ_k	the kth step of SGD .	(3)
V_k	SGD noise caused by mini-batch sampling of loss gradients.	(5)
η	the learning rate of SGD.	(3)
$\Theta(t)$	the continuous-time dynamics of SGD.	(6)
z_k	the random vector of standard Gaussian.	(9)
$W(t)$	the Brownian motion over time.	(6)
$\bar{\theta}_k$	the kth step of discrete-time approximation to $\Theta(t)$.	(7)
	SGD with Unbiased Label Noises (ULN)	
θ_k^{ULN}	the kth step of SGD with unbiased label noises .	(8)
ξ_k^*	SGD noises thru. mini-batch sampling of TRUE loss gradients.	(8)
ξ_k^{ULN}	SGD noise thru. mini-batch sampling of unbiased label noises.	(8)
Σ_N^{SGD}	the covariance matrix of the TRUE loss gradients.	(15),(16),(17)
Σ_N^{ULN}	the covariance matrix based on unbiased label noises.	(15),(16),(17)
$\Theta^{\mathrm{ULN}}(t)$	the continuous-time doubly stochastic model.	(16)
$\bar{\theta}_k^{\mathrm{ULN}}$	the kth step of discrete-time doubly stochastic model.	(17)
$W_1(t)$, and $W_2(t)$	two independent Brownian motions over time.	(16)
z_k, and z_k'	two independent random vectors of standard Gaussian.	(9),(17).
$\Theta^{\mathrm{LNL}}(t)$	the continuous-time dynamics under Label-Noiseless settings.	(20)

Table 1: Key Symbols and Definitions

would affect the "selection" of $\widehat{\theta}$ from possible solutions, in the viewpoint of learning dynamics [16] of SGD under mini-batch sampling [17, 18, 12]. For symbols used in this paper, please refer to Table 1.

1.1. Backgrounds: SGD Dynamics and Implicit Regularization

To analyze the SGD algorithm solving the problem in Eq (2), we follow settings in [17] and consider SGD as an algorithm that, in the kth iteration with the estimate θ_k, it randomly picks up a b-length subset of samples from the training dataset i.e., $B_k \subset \mathcal{D}$, and estimates the mini-batch stochastic gradient $\frac{1}{b}\sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k)$, then updates the

Accepted Manuscript

estimate for θ_{k+1} based on θ_k, as follow

$$\theta_{k+1} \leftarrow \left(\theta_k - \frac{\eta}{|B_k|}\sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k)\right) \ , \tag{3}$$

where η refers to the step-size of SGD. Furthermore, we can derive the mini-batch sampled loss gradients into the combination of the full batch loss gradient and the noise, such that

$$\begin{aligned} &\frac{\eta}{|B_k|}\sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k) \\ =&\frac{\eta}{N}\sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \left(\frac{\eta}{N}\sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \frac{\eta}{|B_k|}\sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k)\right) \\ =&\frac{\eta}{N}\sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \sqrt{\eta} V_k(\theta_k) \ , \end{aligned} \tag{4}$$

where η refers to the step-size of SGD, and $V_k(\theta_k)$ refers to a stochastic gradient noise term caused by mini-batch sampling. The noise would converge to zero with increasing batch size, as follow

$$V_k(\theta_k) = \sqrt{\eta}\left(\frac{1}{N}\sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \frac{1}{|B_k|}\sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k)\right) \to \mathbf{0}_d, \quad \text{as} \quad B \to N \ . \tag{5}$$

With $\mathbf{d}t = \eta \to 0$ and the constant batch size $b = |B_k|$, SGD algorithm would diffuse to a continuous-time dynamics $\Theta(t)$ with a stochastic differential equation (SDE), with weak convergence [17, 19], as follow,

$$\mathbf{d}\Theta = -\frac{1}{N}\sum_{i=1}^{N} \nabla \tilde{L}_i(\Theta)\mathbf{d}t + \left(\frac{\eta}{b}\tilde{\Sigma}_N^{\text{SGD}}(\Theta)\right)^{\frac{1}{2}} \mathbf{d}W(t) \tag{6}$$

where $W(t)$ is a standard Brownian motion in $\mathbb{R}^d$, and let us define $\Sigma_N^{\text{SGD}}(\Theta)$ as the sample covariance matrix of loss gradients $\nabla L_i(\Theta)$ for $1 \leq i \leq N$. Please note that, for detailed derivations to obtain above continuous-time approximation and the assumptions, please refer to [17]. We follow [17] and do not make low-rank assumptions on $\tilde{\Sigma}_N^{\text{SGD}}(\Theta)$. Through Euler discretization [17, 11], one can approximate SGD as $\bar{\theta}_k$ such that

$$\begin{aligned} &\bar{\theta}_{k+1} \leftarrow \bar{\theta}_k - \frac{\eta}{N}\sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\bar{\theta}_k) + \sqrt{\eta}\xi_k(\bar{\theta}_k), \text{ and} \\ &\xi_k(\bar{\theta}_k) = \left(\frac{\eta}{b}\tilde{\Sigma}_N^{\text{SGD}}(\bar{\theta}_k)\right)^{\frac{1}{2}} z_k, \ z_k \sim \mathcal{N}(0, \mathbf{I}_d) \ . \end{aligned} \tag{7}$$

The implicit regularizer of SGD is $\xi_k(\bar{\theta}_k) = \left(\frac{\eta}{b}\tilde{\Sigma}_N^{\text{SGD}}(\bar{\theta}_k)\right)^{\frac{1}{2}} z_k$ which is data-dependent and controlled by the learning rate η and batch size B [20]. [10, 11, 12] discussed SGD for varational inference and enabled novel applications to samplers [21, 22]. To understand

Accepted Manuscript

the effect to generalization performance, [1, 20] studied the escaping behavior from the sharp local minima [8] and convergence to the flat ones. [23] discover the way that SGD could find a flat local minimum from information-theoretical perspectives and propose a novel regularizer to improve the performance. Finally, [24] studied regularization effects to linear DNNs and our previous work [18] proposed new multiplicative noises to interpret SGD and obtain stronger theoretical properties.

1.2. Our Contributions

In this work, we assume the unbiased random label noises ε_i $(1 \leq i \leq N)$ and the mini-batch sampler of SGD are independent. When the random label noises have been drawn from probability distributions prior to the training procedure, SGD re-samples the label noises and generates a new type of data-dependent noises, in addition to the stochastic gradient noises of label-noiseless losses, through re-sampling label-noisy data and averaging label-noisy loss gradients of random mini-batchs [25, 26].

Our analysis shows that under mild conditions, with gradients of label-noisy losses, SGD might incorporate an additional data-dependent noise term, complementing with the stochastic gradient noises [17, 18] of label-noiseless losses, through re-sampling the samples with label noises [26] or dynamically adding noises to labels over iterations [15]. We consider such noises as an implicit regularization caused by unbiased label noises, and interpret the effects of such noises as a solution selector of learning procedure. More specifically, this work has made unique contributions as follow.

1.2.1. Doubly Stochastic Models We reviewed the preliminaries [17, 27, 12, 18] and extended the analytical framework in [17] to interpret the effects of unbiased label noises as an additional implicit regularizer on top of the continuous-time dynamics of SGD. Through discretizing the continuous-time dynamics of label-noisy SGD, we write discrete-time approximation to the learning dynamics, denoted as θ_k^{ULN} for $k = 1, 2, \ldots,$ as

$$\theta_{k+1}^{\mathrm{ULN}} \leftarrow \theta_k^{\mathrm{ULN}} - \frac{\eta}{N} \sum_{i=1}^{N} \nabla L_i^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^{\mathrm{ULN}}(\theta_k^{\mathrm{ULN}}), \tag{8}$$

where $L_i^*(\theta) = (f(x_i, \theta) - f^*(x_i))^2$ refers to the *label-noiseless loss function* with sample x_i and the true (noiseless) label y_i, the noise term $\xi_k^*(\theta)$ refers to the stochastic gradient noise [17] of label-noiseless loss function $L_i^*(\theta)$, then we can obtain the new implicit regularizer caused by the unbiased label noises (ULN) for $\forall \theta \in \mathbb{R}^d$, which can be approximated as follow

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\eta\sigma^2}{bN} \sum_{i=1}^{N} \nabla_\theta f(x_i, \theta) \nabla_\theta f(x_i, \theta)^\top \right)^{\frac{1}{2}} z_k, \text{ and } z_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) \ , \tag{9}$$

where z_k refers to a random noise vector drawn from the standard Gaussian distribution, θ_k refers to the parameters of network in the kth iteration, $(\cdot)^{1/2}$ refers to the Chelosky

Accepted Manuscript

decomposition of the matrix, $\nabla_\theta f(x_i, \theta) = \partial f(x_i, \theta)/\partial\theta$ refers to the gradient of the neural network output for sample x_i over the parameter θ_k, and B and η are defined as the batch size and the learning rate of SGD respectively. Obviously, the strength of such implicit regularizer is controlled by σ^2, B and η.

Section 3 formulates the algorithm of SGD with unbiased random label noises as a stochastic dynamics based on two noise terms (**Proposition** 1), derives the Continuous-time and Discrete-time Doubly Stochastic Models from SGD algorithms (**Definitions** 1 and 2), and provides approximation error bounds (**Proposition** 2). Proofs of two propositions are provided in Appendix A and Appendix B

1.2.2. Inference Stabilizer as Implicit Regularizer The regularization effects of unbiased random label noises should be

$$\mathbb{E}_{z_k} \left\|\xi_k^{\mathrm{ULN}}(\theta_k)\right\|_2^2 \approx \frac{\eta\sigma^2}{bN} \sum_{i=1}^{N} \|\nabla_\theta f(x_i, \theta_k)\|_2^2 = \frac{\eta\sigma^2}{bN} \sum_{i=1}^{N} \left\|\frac{\partial}{\partial\theta} f(x_i, \theta_k)\right\|_2^2 , \tag{10}$$

where $\nabla_\theta f(x, \theta)$ refers to the gradient of f over θ and the effects are controlled by the batch size B and the variance of label noises σ^2. Similar results have been obtained by assuming the deep learning algorithms have been driven by an Ornstein-Uhlenbeck like process [3], while our work does not rely on such assumption but is all based on our proposed *Doubly Stochastic Models*.

Section 4 analyzes the implicit regularization effects of unbiased random label noises for SGD, where we conclude the implicit regularizer as a controller of the neural network gradient norm $\frac{1}{N}\sum_{i=1}^{N} \|\nabla_\theta f(x_i, \theta)\|_2^2$ in the dynamics (**Proposition** 3). We then offer **Remarks** 2, 3 and 4 to characterize the behaviors of SGD with unbiased label noises: (1) SGD would escape the local minimums with higher gradient norms, due to the larger perturbation driven by the implicit regularizer, (2) the strength of implicit regularization effects is controlled by the learning rate η and batch size b, and (3) it is possible to tune the performance of SGD through adding and controlling the unbiased label noises, as low neural network gradient norms usually correspond to flat loss landscapes.

We validate our findings through of a series of experiments. In Section 5, we apply *self-distillation with unbiased label noises* [28, 29] for training deep neural networks. Through the teacher-student learning paradigm, the well-trained model beneficially escapes local minima by learning from its own noisy outputs. In Section 6, we visualize implicit regularization effects using SGD-based linear regression with unbiased label noises. We observe a Gaussian-like distribution, centered at the solution for linear regression, with its (co-)variance determined by the covariance of data samples, the learning rate, and the batch size. Collectively, the results of these experiments substantiate our theory.

1.2.3. Significance of Our Contributions This work establishes a novel framework that provides an in-depth understanding of the influence of unbiased label noises in Stochastic Gradient Descent (SGD). The study highlights the implicit regularization

Accepted Manuscript

effects within the learning process and demonstrates how these effects can be quantified through doubly stochastic models. By assuming the independence of two noise sources and using globally bounded assumptions, the paper brings new theoretical insights on SGD dynamics, which are reinforced by empirical evidence obtained from Deep Neural Networks (DNN) and linear regression models. Moreover, it provides a comprehensive empirical studies on the impacts of additive noise during the self-distillation process and SGD-based linear regression. Therefore, this study significantly contributes to the learning space by elucidating the underlying mechanisms of SGD and broadening our understanding of implicit regularization effects due to unbiased label noises, especially in the context of complex machine learning algorithms.

2. Related Work

SGD Implicit Regularization for Ordinary Least Square (OLS) The most recent and relevant work in this area is [27, 30], where the same group of authors studied the implicit regularization of gradient descent and stochastic gradient descent for OLS. They investigated an implicit regularizer of ℓ_2-norm alike on the parameter, which regularizes OLS as a Ridge estimator with decaying penalty. Prior to these efforts, F. Bach and his group have studied the convergence of gradient-based solutions for linear regression with OLS and regularized estimators under both noisy and noiseless settings in [31, 32, 33].

Langevin Dynamics and Gradient Noises With similar agendas, [10, 11, 12] studied limiting behaviors of SGD (or steady-state of dynamics) from the perspectives of Bayesian/variational inference. They also promoted novel applications to stochastic gradient MCMC samplers [21, 22]. Through connecting $\Sigma_N^{\mathrm{SGD}}(\theta)$ to the loss Hessian $1/N\nabla^2 L_i(\theta)$ in near-convergence regions, [1] studied the escaping behavior from the sharp local minima, while [8] discussed this issue in large-batch training settings. Furthermore, [20] discussed how learning rates and batch sizes would affect the generalization performance and flatness of optimization results. Finally, [24] studied the implicit regularization on linear neural networks and [18] proposed a new multiplicative noise model to interpret the gradient noises with stronger theoretical properties.

Self-Distillation and Noisy Students Self-distillation [28, 34, 35, 29] has been examined as an effective way to further improve the generalization performance of well-trained models. Such strategies enable knowledge distillation using the well-trained ones as teacher models and optionally adding noises (e.g., dropout, stochastic depth, and label smoothing or potentially the label noises) onto training procedure of student models [15].

Deep Learning with Label Noises The influence of label noises on deep learning has been explored in several studies, including [3, 36, 37]. Some of these findings align with our work; for example, [3, 36] achieved similar results, through different approaches. [3] also viewed SGD as a dynamical system and made similar observations by establishing

Accepted Manuscript

stronger assumptions (please refer to the *Discussion on the Relevant Work*). Bar et al. [36] applied a spectral analysis to the learned mapping of networks and provided theoretical justification for the observed robustness to label noise. They associated typical smoothness regularization to the suppression of high-frequency components potentially caused by label noises. On a different note, a survey by Song et al. [37] reviews deep learning algorithms for label noise. The studies featured in this survey predominantly view label noises as a negative influencer of Deep Neural Network (DNN) performance, proposing robust training algorithms as a countermeasure. Our work, however, is more concerned with the potential advantages conferred by unbiased label noises. This distinct aspect sets our work apart from other works.

Discussion on the Relevant Work Compared to above works, we still make contributions in above three categories. First of all, this work characterizes the implicit regularization effects of label noises to SGD dynamics. Compared to [27, 30] working on linear regression, our proposed doubly stochastic model could be used to explained the learning dynamics of SGD with label noises for nonlinear neural networks. Even from linear regression perspectives [27, 30, 33], we precisely measured the gaps between SGD dynamics with and without label noises and provide an new example with numerical simulation to visualize the implicit regularization effects.

Compared to [38, 29], our analysis emphasized the role of the implicit regularizer caused by label noises for model selection, where models with high inferential stability would be selected. [39] is the most relevant work to us, where authors studied the early stopping of gradient descent under label noises via neural tangent kernel (NTK) [40] approximation. Our work made the analyze for SGD without assumptions for approximation such as NTK.

In addition to NTK assumption, [3] assumes the deep learning algorithms are driven by an Ornstein-Uhlenbeck (OU) like process and obtains similar results as the inference stabilizer (the third result of our research), while our work makes contribution through proposing *Doubly Stochastic Models* and reach the conclusion in a different way. We also provide yet the first empirical results and evidences, based on commonly-used DNN architectures and benchmark datasets, to visualize the effects of implicit regularizers caused by the unbiased label noises in real-world settings.

Please note that an earlier manuscript [41] from us has been put on OpenReview with discussion, where external reviewers demonstrated their concerns–part of results has been investigated in [3] and we didn't provide the results in a strong form (e.g., theorems or proofs). Hereby, this work shifts the main contributions from implicit regularization of label noises to the *doubly stochastic models* with approximation error bounds and proofs. The implicit regularization effects could be estimated via doubly stochastic models directly without the assumption of OU process. To best of our knowledge, this work is the first to understand the effects of unbiased label noises to SGD dynamics, by addressing technical issues including implicit regularization, OLS, self-distillation, model selection, and the stability inference results.

Accepted Manuscript

3. Double Stochastic Models for SGD with Unbiased Random Label Noises

In this section, we present SGD with unbiased random label noises, derive the Continuous-time/Discrete-time Doubly Stochastic Models, and provide convergence of approximation between models.

3.1. Modeling Unbiased Label Noises in SGD

In our research, SGD with Unbiased Random Label Noises refers to an iterative algorithm that updates the estimate incrementally from initialization θ_0^{ULN}. With mini-batch sampling and unbiased random label noises, in the kth iteration, SGD algorithm updates the estimate θ_k^{ULN} using the stochastic gradient $\tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}})$ through a gradient descent rule , such that

$$\theta_{k+1}^{\mathrm{ULN}} \leftarrow \theta_k^{\mathrm{ULN}} - \eta \tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}}) \ , \tag{11}$$

Specifically, in the kth iteration, SGD randomly picks up a batch of sample $B_k \subseteq \mathcal{D}$ to estimate the stochastic gradient, as follow

$$\begin{aligned} \eta \tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}}) &= \frac{\eta}{|B_k|} \sum_{x_i \in B_k} \nabla \tilde{L}_i(\theta_k^{\mathrm{ULN}}) \\ &= \frac{\eta}{|B_k|} \sum_{x_i \in B_k} \big(\big(f(x_i, \theta_k^{\mathrm{ULN}}) - y_i \big) - \varepsilon_i \big) \cdot \nabla_\theta f(x_i, \theta_k^{\mathrm{ULN}}) \\ &= \frac{\eta}{N} \sum_{i=1}^{N} \nabla L_i^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta} \xi_k^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta} \xi_k^{\mathrm{ULN}}(\theta_k^{\mathrm{ULN}}), \end{aligned} \tag{12}$$

where $\nabla L_i^*(\theta)$ for $\forall \theta \in \mathbb{R}^d$ refers to the loss gradient based on the label-noiseless sample (x_i, y_i) and $y_i = f^*(x_i)$, $\xi_k^*(\theta)$ refers to stochastic gradient noises [17] through mini-batch sampling over the gradients of label-noiseless samples, and $\xi_k^{\mathrm{ULN}}(\theta)$ is an additional noise term caused by the mini-batch sampling and the Unbiased Random Label Noises, such that

$$\begin{aligned} \nabla L_i^*(\theta) &= \frac{\partial \ (f(x_i, \theta) - f^*(x_i))^2}{\partial \theta \quad 2} = (f(x_i, \theta) - f^*(x_i)) \cdot \nabla f(x_i, \theta) \ , \\ \xi_k^*(\theta) &= \frac{\sqrt{\eta}}{|B_k|} \sum_{x_j \in B_k} \left(\nabla L_j^*(\theta) - \frac{1}{N} \sum_{i=1}^{N} \nabla L_i^*(\theta) \right), \\ \xi_k^{\mathrm{ULN}}(\theta) &= -\frac{\sqrt{\eta}}{|B_k|} \sum_{x_j \in B_k} \varepsilon_j \cdot \nabla_\theta f(x_j, \theta) \ . \end{aligned} \tag{13}$$

Proposition 1 (Mean and Variance of the Two Noise Terms)**.** *The mean and variance of the noise terms $\xi_k^*(\theta)$ and $\xi_k^{\mathrm{ULN}}(\theta)$ should be the vector-value functions as follow*

$$\begin{aligned} &\mathbb{E}_{B_k}[\xi_k^*(\theta)] = \mathbf{0}_d, && \textit{and } \mathrm{Var}_{B_k}[\xi_k^*(\theta)] = \frac{\eta}{|B_k|} \Sigma_N^{\mathrm{SGD}}(\theta) \\ &\mathbb{E}_{B_k, \varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] = \mathbf{0}_d, && \textit{and } \mathrm{Var}_{B_k \varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] = \frac{\eta}{|B_k|} \Sigma_N^{\mathrm{ULN}}(\theta) \ . \end{aligned} \tag{14}$$

Accepted Manuscript

The two matrix-value functions $\Sigma_N^{\mathrm{SGD}}(\theta)$ and $\Sigma_N^{\mathrm{ULN}}(\theta)$ over $\theta \in \mathbb{R}^d$ characterize the variance of noise vectors. When we assume the label noises and mini-batch sampling are independent, there has

$$\begin{aligned}\Sigma_N^{\mathrm{SGD}}(\theta) &= \frac{1}{N}\sum_{j=1}^{N}\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N} L_i^*(\theta)\right)\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N} L_i^*(\theta)\right)^\top \\ \Sigma_N^{\mathrm{ULN}}(\theta) &= \frac{\sigma^2}{N}\sum_{j=1}^{N}\nabla_\theta f(x_j,\theta)\nabla_\theta f(x_j,\theta)^\top \quad \textit{as} \quad \mathrm{var}[\varepsilon_j] = \sigma^2 \ .\end{aligned} \tag{15}$$

The two noise terms $\xi_k^*(\theta)$ and $\xi_k^{\mathrm{ULN}}(\theta)$ that are controlled by the learning rate and the batch size would largely influence the SGD dynamics. Please refer to Appendix A for proofs.

With the mean and variance of two noise terms, we can easily formulate the learning dynamics of SGD with unbiased label noises as follows.

3.2. Doubly Stochastic Models and Approximation

We consider the SGD algorithm with unbiased random label noises in the form of gradient descent with additive data-dependent noise, such that $\theta_{k+1} = \theta_k - \frac{\eta}{N}\sum_{i=1}^{N}\nabla\tilde{L}_i(\theta_k) + \sqrt{\eta}\tilde{V}_k(\theta_k)$. To simplify the model and analysis involved, we assume

A1. The two noise terms $\xi_k^*(\theta_k)$ and $\xi_k^{\mathrm{ULN}}(\theta_k)$ are independent.

By the making above assumption, the complexities of considering their potential interaction or correlation can be bypassed. Furthermore, our later experiments (in Sections 5 and 6) would show the robustness and correctness of our model and analysis based on this assumption. In this way, when $\eta \to 0$, we can follow the analysis in [19] to derive the diffusion process of SGD with unbiased random label noises, denoted as $\Theta^{\mathrm{ULN}}(t)$ over continuous-time $t \geq 0$. We define the *Doubly Stochastic Models* that characterizes the continuous-time dynamics of SGD with unbiased label noises as follows.

Definition 1 (Continuous-Time Doubly Stochastic Models)**.** *Given an SGD algorithm θ_k^{ULN} defined and specified in Section 3.1, with $\eta = \mathbf{d}t$, we assume $B_k = B$ for $k = 1, 2, 3\ldots$ and formulate its continuous-time dynamics as*

$$\begin{aligned}\mathbf{d}\Theta^{\mathrm{ULN}} = -\frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\Theta^{\mathrm{ULN}})\mathbf{d}t &+ \left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\Theta^{\mathrm{ULN}})\right)^{\frac{1}{2}}\mathbf{d}W_1(t) \\ &+ \left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\Theta^{\mathrm{ULN}})\right)^{\frac{1}{2}}\mathbf{d}W_2(t) \ ,\end{aligned} \tag{16}$$

where $W_1(t)$ and $W_2(t)$ refer to two independent Brownian motions over time, $\mathbf{d}t = \eta$ and $\Theta^{\mathrm{ULN}}(0) = \theta_0^{\mathrm{ULN}}$.

Obviously, we can obtain the discrete-time approximation [17, 11] to the SGD dynamics as follows.

Accepted Manuscript

Definition 2 (Discrete-Time Doubly Stochastic Models)**.** *We denote* $\bar{\theta}_k^{\mathrm{ULN}}$ *for* $k = 1, 2, \ldots$ *as the discrete-time approximation to the* Doubly Stochastic Models for SGD with Unbiased Label Noises*, which in the* k^{th} *iteration behaves as*

$$\bar{\theta}_{k+1}^{\mathrm{ULN}} \leftarrow \bar{\theta}_k^{\mathrm{ULN}} - \frac{\eta}{N}\sum_{i=1}^{N} \nabla L_i^*(\bar{\theta}_k^{\mathrm{ULN}}) + \sqrt{\eta}\left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\bar{\theta}_k^{\mathrm{ULN}})\right)^{\frac{1}{2}} z_k + \sqrt{\eta}\left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\bar{\theta}_k^{\mathrm{ULN}})\right)^{\frac{1}{2}} z_k', \tag{17}$$

where z_k *and* z_k' *are two independent* d*-dimensional random vectors drawn from a standard* d*-dimensional Gaussian distribution* $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ *per iteration independently, and* $\bar{\theta}_0^{\mathrm{ULN}} = \Theta^{\mathrm{ULN}}(0)$.

The convergence between $\bar{\theta}_k^{\mathrm{ULN}}$ and $\Theta^{\mathrm{ULN}}(t)$ is tight when $t = k\eta$ and the convergence bound is as follow.

Proposition 2 (Convergence of Approximation)**.** *Let* $T \geq 0$*. Let* $\Sigma_N^{\mathrm{SGD}}(\theta)$ *and* $\Sigma_N^{\mathrm{ULN}}(\theta)$ *be the two diffusion matrices defined in Eq. 15. Assume that*

A2. There exists some $M > 0$ *such that* $\max_{i=1,2,\ldots,N}\{(\|\nabla L_i^*(\theta)\|_2)\} \leq M$ *and* $\max_{i=1,2,\ldots,N}\{(\|\nabla_\theta f(x_i;\theta)\|_2)\} \leq M$*;*

A3. There exists some $L > 0$ *such that* $\nabla L_i(\theta)$ *and* $\nabla_\theta f(x,\theta)$ *for* $\forall x \in \mathcal{D}$ *are Lipschitz continuous with bounded Lipschitz constant* $L > 0$ *uniformly for all* $i = 1, 2, ..., N$*.*

The continuous-time dynamics of SGD with unbiased label noises (ULN), denoted as $\Theta^{\mathrm{ULN}}(t)$ *in Eq.* (16)*, is with order* 1 *strong approximation to the discrete time SGD dynamics* $\bar{\theta}_k^{\mathrm{ULN}}$ *in Eq.* (17)*. I.e., there exist a constant* C *independent on* η *but depending on* σ^2*,* L *and* M *such that*

$$\mathbb{E}\|\Theta^{\mathrm{ULN}}(k\eta) - \bar{\theta}_k^{\mathrm{ULN}}\|^2 \leq C\eta^2, \quad \textit{for all } 0 \leq k \leq \lfloor T/\eta \rfloor. \tag{18}$$

Please refer to Appendix B for proofs.

Note that the boundedness and smoothness assumptions made in this proposition have been also used in series of previous studies [17, 2, 42, 43, 44]. We believe the assumptions are reasonable under the deep learning settings. For example, the Rectified Linear Unit (ReLU) activation function [45], commonly used in DNNs, has a globally bounded derivative, being either 0 or 1. The boundedness of DNN also depends on the norms of weight matrices, while well-trained DNNs are often with norm-bounded weight matrices [46, 47, 48, 49]. Techniques like weight normalization or layer normalization can further maintain this boundedness.

Remark 1. *With above strong convergence bound for approximation, we can consider* $\bar{\theta}_k^{\mathrm{ULN}}$ *– the solution of Eq.* (17) *– as a tight approximation to the SGD algorithm with unbiased label noises based on the same initialization. A tight approximation to the noise term* $\xi_k^{\mathrm{ULN}}(\theta)$ *(defined in Eq.* (27)*) could be as follow*

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\theta)\right)^{\frac{1}{2}} z_k', \ \textbf{\textit{and}}\ z_k' \sim \mathcal{N}(\mathbf{0_d}, \mathbf{I_d}) \ . \tag{19}$$

Accepted Manuscript

We use such discrete-time iterations and approximations to the noise term $\xi_k^{\mathrm{ULN}}(\theta)$ to interpret the implicit regularization behaviors of the SGD with unbiased label noises algorithm θ_k^{ULN} accordingly.

4. Implicit Regularization Effects to Neural Networks

In this section, we use our model to interpret the regularization effects of *SGD with unbiased label noises* for general neural networks, without assumptions on the structures of neural networks.

4.1. Implicit Regularizer Influenced by Unbiased Random Label Noises

Compared the stochastic gradient with unbiased random label noises $\tilde{g}_k(\theta)$ and the stochastic gradient based on the label-noiseless losses, we find an additional noise term $\xi_k^{\mathrm{ULN}}(\theta)$ as the **implicit regularizer**.

To interpret $\xi_k^{\mathrm{ULN}}(\theta)$, we first define the diffusion process of SGD based on Label-NoiseLess losses i.e., $L_i^*(\theta)$ for $1 \leq i \leq N$ as

$$\mathbf{d}\Theta^{\mathrm{LNL}} = -\frac{1}{N}\sum_{i=1}^{N} \nabla L_i^*(\Theta^{\mathrm{LNL}})\mathbf{d}t + \left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\Theta^{\mathrm{LNL}})\right)^{\frac{1}{2}} \mathbf{W}(t) \ . \tag{20}$$

Through comparing $\Theta^{\mathrm{ULN}}(t)$ with $\Theta^{\mathrm{LNL}}(t)$, the effects of $\xi_k^{\mathrm{ULN}}(\Theta)$ over contin-uous-time form should be $\sqrt{\eta/B}(\Sigma_N^{\mathrm{ULN}}(\Theta))^{1/2}\mathbf{d}W(t)$. Then, in discrete-time, we could get results as follow.

Proposition 3 (The implicit regularizer $\xi_k^{\mathrm{ULN}}(\theta)$)**.** *The implicit regularizer of SGD with unbiased random label noises could be approximated as follow,*

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\sigma^2\eta}{bN}\sum_{i=1}^{N} \nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)^{\frac{1}{2}} z_k, \ \textit{and } z_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) \ . \tag{21}$$

In this way, we can estimate the expected regularization effects of the implicit regularizer $\|\xi_k^{\mathrm{ULN}}(\theta)\|_2$ as follow,

$$\mathbb{E}_{z_k}\|\xi_k^{\mathrm{ULN}}(\theta)\|_2^2 = \frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2 \, . \tag{22}$$

Please refer to Appendix C for proofs.

We thus conclude that the effects of implicit regularization caused by unbiased random label noises for SGD is proportional to $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ the average gradient norm of the neural network $f(x,\theta)$ over samples.

Accepted Manuscript

4.2. Understanding the Unbiased Label Noises as an Inference Stabilizer

Here we extend the existing results on SGD [1, 50] to understand **Proposition 3** and obtain remarks as follows.

Remark 2 (Inference Stability)**.** *In the partial derivative form, the gradient norm could be written as*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\|\frac{\partial}{\partial\theta}f(x_i,\theta)\|_2^2 \tag{23}$$

characterizes the variation of neural network output $f(x,\theta)$ based on samples x_i (for $1 \leq i \leq N$) over the parameter interpolation around the point θ. Lower $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ leads to higher stability of neural network $f(x,\theta)$ outputs against the (random) perturbations over parameters.

Remark 3 (Escape and Converge)**.** *When the noise $\xi_k^{\mathrm{ULN}}(\theta)$ is θ-dependent (section 4 would present a special case that $\xi_k^{\mathrm{ULN}}(\theta)$ is θ-independent with OLS), we follow [1] and suggest that the implicit regularizer helps SGD escape from the point $\tilde{\theta}$ with high neural network gradient norm $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\tilde{\theta})\|_2^2$, as the scale of noise $\xi_k^{\mathrm{ULN}}(\tilde{\theta})$ is large. Reciprocally, we follow [50] and suggest that when the SGD with unbiased random label noises converges, the algorithm would converge to a point θ^* with small $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta^*)\|_2^2$. Similar results have been obtained in [3] when assuming the deep learning algorithms are driven by an Ornstein–Uhlenbeck process.*

Remark 4 (Performance Tuning)**.** *Considering $\eta\sigma^2/B$ as the coefficient balancing the implicit regularizer and vanilla SGD, one can regularize/penalize the SGD learning procedure with the fixed η and B more fiercely using a larger σ^2. More specifically, we could expect to obtain a regularized solution with lower $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ or higher inference stability of neural networks, as regularization effects become stronger when σ^2 increases.*

5. Experiments on Self-Distillation with Unbiased Label Noises

The goal of this experiment is to understand **Proposition 3** and*Remarks. 3 & 4*, i.e., examining (1) whether the unbiased label noises would lower the gradient norm of the neural networks; (2) whether such unbiased label noises would improve the performance of neural networks; and (3) whether one can carry out performance tuning through controlling the variances of unbiased label noises, all in real-world deep learning settings.

5.1. Experiments Design

To evaluate SGD with unbiased label noises, we design a set of novel experiments based on *self-distillation with unbiased label noises*. In addition to learn from noisy labels directly, our experiment intends to train a (student) network from the noisy outputs of a (teacher) network in a quadratic regression loss, where the student network has been

Accepted Manuscript

initialized from weights of the teacher one and unbiased label noises are given to the soft outputs of the teacher network randomly.

We aim to directly measure the gradient norm $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ of the neural network after every epoch to testify the SGD implicit regularization effects of unbiased label noises (i.e., **Proposition 3**). The performance comparisons among the teacher network, the student network (trained with unbiased label noises), and the student network (trained noiselessly) demonstrate the advantage of unbiased label noises in SGD for regression tasks (i.e., *Remarks. 3 & 4*).

Particularly, we design a set of novel experiments based on *self-distillation with unbiased label noises* and elaborate in which way the proposed *SGD with unbiased label noises* fits the settings of *self-distillation with unbiased label noises*. Further, we introduce the goal of our empirical experiments with a list of expected evidences, then present the experiment settings for the empirical evaluation. Finally, we present the experiment results with solid evidence to validate our proposals in this work.

5.2. Noisy Self-Distillation

Given a well-trained model, Self-Distillation algorithms [28, 35, 29, 34] intend to further improve the performance of a model through learning from the "soft label" outputs (i.e., logits) of the model (as the teacher). Furthermore, some practices found that the self-distillation could be further improved through incorporating certain randomness and stochasticity in the training procedure, namely noisy self-distillation, so as to obtain better generalization performance [34, 29]. In this work, we study two well-known strategies for additive noises as follow.

(i) *Gaussian Noises.* Given a pre-trained model with $\mathbb{L}$-dimensional logit output, for every iteration of self-distillation, this simple first draws random vectors from a $\mathbb{L}$-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}_\mathbb{L},\sigma^2\mathbf{I}_\mathbb{L})$, then adds the vectors to the logit outputs of the model. It makes the student model learn from the noisy outputs. Note that in our analysis, we assume the output of the model is single dimension while, in self-distillation, the logit labels are with multiple dimensions. Thus, the diagonal matrix $\sigma^2\mathbf{I}_\mathbb{L}$ now refers to the complete form of the variances and σ^2 controls the scale of variances of noises.

(ii) *Symmetric Noises..* Basically, this strategy is derived from [15] that generates noises through randomly swapping the values of logit output among the $\mathbb{L}$ dimensions. Specifically, in every iteration of self-distillation, given a swap-probability p, every logit output (denoted as y here) from the pre-trained model, and every dimension of logit output denoted as y_l, the strategy in probability p swaps the logit value in the dimension that corresponds to y_l with any other dimension $y_{m\neq l}$ in equal prior (i.e., in $(\mathbb{L}-1)^{-1}$ probability). In the rest $1-p$ probability, the strategy remains the original logit output there. In this way, the new noisy label $\tilde{y}$

Accepted Manuscript

is with expectation $\mathbb{E}[\tilde{y}]$ as follow,

$$\mathbb{E}[\tilde{y}_l] = (1-p) \cdot y_l + \frac{p \cdot \sum_{\forall m \neq l} y_m}{\mathbb{L} - 1} \tag{24}$$

This strategy introduces explicit bias to the original logit outputs. However, when we consider the expectation $\mathbb{E}[\tilde{y}]$ as the innovative soft label, then the random noise around the new soft label is still unbiased as $\mathbb{E}[\tilde{y} - \mathbb{E}[\tilde{y}]] = 0$ for all dimensions. Note that this noise is not the symmetric noises studied for robust learning [51].

Thus, literately, our proposed *SGD with Unbiased Label Noises* settings well fit the practice of noisy-self distillation.

5.3. Datasets and DNN Models

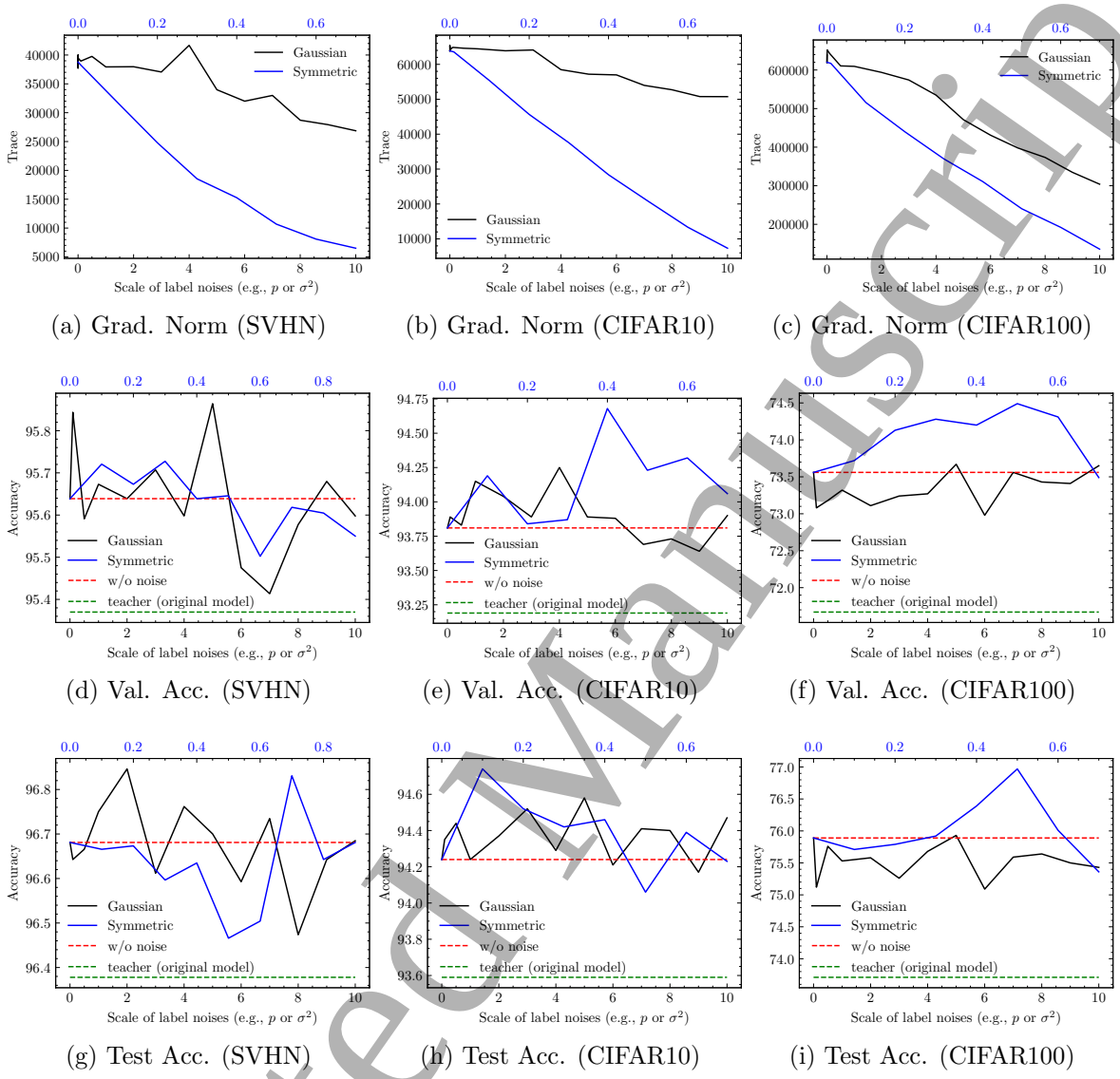
We choose the ResNet-56 [52], one of the most practical deep models, for conducting the experiments on three datasets: SVHN [53], CIFAR-10 and CIFAR-100 [54]. We follow the standard training procedure [52] for training a teacher model (original model). Specifically we train the model from scratch for 200 epochs and adopt the SGD optimizer with batch size 64 and momentum 0.9. The learning rate is set to 0.1 at the beginning of training and divided by 10 at 100th epoch and 150th epoch. A standard weight decay with a small regularization parameter (10^{-4}) is applied. As for noiseless self-distillation, we follow the standard procedure [55] for distilling knowledge from the teacher to a student of the same network structure. The basic training setting is the same as training the teacher model.

For the settings of noisy self-distillation, we divide the original training set into a new training set (80%) and a validation set (20%). For clarity, we also present the results using varying scales of unbiased label noises on all three sets, where the original training set is used for training.

5.4. Experiment Results

Figure 1 presents the results of above two methods with increasing scales of noises, i.e., increasing σ^2 for Gaussian noises and increasing p for Symmetric noises. In Figure 1(a)–(c), we demonstrate that the gradient norms of neural networks $\frac{1}{N}\sum_{i=1}^{N} \|\nabla_\theta f(x_i, \theta)\|_2^2$ decrease with growing σ^2 and p for two strategies. The results backup our theoretical investigation , which means the model would be awarded high inferential stability, as the variation of neural network outputs against the potential random perturbation in parameters has been reduced by the regularization. In Figure 1(d)–(f) and (g)–(i), we plot the validation and testing accuracy of the models obtained under noisy self-distillation. The results show that (1) student models are with lower gradient norms of neural networks $\frac{1}{N}\sum_{i=1}^{N} \|\nabla_\theta f(x_i, \theta)\|_2^2$ than teacher models, the gradient norm further decreases with increasing scale of noises (i.e., σ^2 and P); (2) some of models have been improved through noisy self-distillation compared to the teacher model, while noisy self-distillation could obtain better performance than noiseless self-distillation; and (3) it is

Accepted Manuscript



(a) Grad. Norm (SVHN) (b) Grad. Norm (CIFAR10) (c) Grad. Norm (CIFAR100)

(d) Val. Acc. (SVHN) (e) Val. Acc. (CIFAR10) (f) Val. Acc. (CIFAR100)

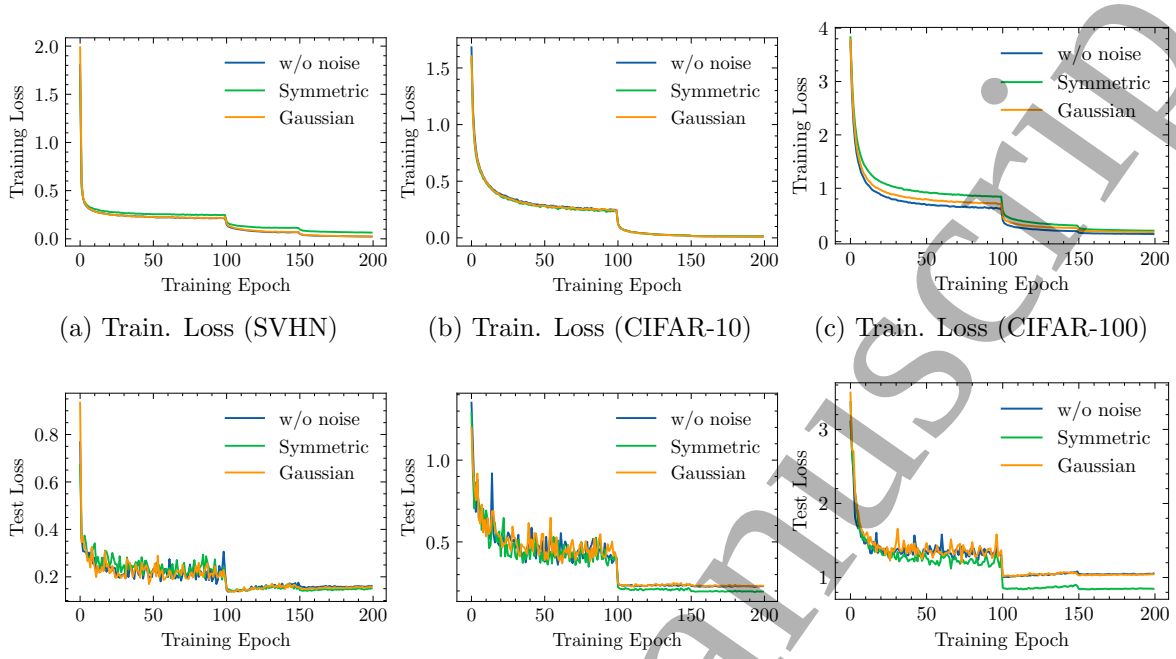
(g) Test Acc. (SVHN) (h) Test Acc. (CIFAR10) (i) Test Acc. (CIFAR100)

Figure 1: Gradient Norms (corresponding to the inference stability), Validation Accuracy, and Testing Accuracy in Noisy Self-Distillation using ResNet-56 with varying scale of label noises (e.g., p and σ^2) based on SVNH, CIFAR-10 and CIFAR-100 Datasets.

possible to select noisily self-distilled models using validation accuracy for better overall generalization performance (in testing dataset). All results here are based on 200 epochs of noisy self-distillation.

We show the evolution of training and test losses during the entire training procedure, and compare the settings of adding no label noises, symmetric and Gaussian noises for noisy self-distillation. Figure 2 presents the results on the three datasets, i.e., SVHN, CIFAR-10 and CIFAR-100 with the optimal scales of label noises on validation sets. It shows all algorithms would finally converge to a local minima with a training loss near to zero, while the local minimas searched by the SGD with Symmetric noise would be flatter with better generalization performance (especially for CIFAR-100 dataset).

Accepted Manuscript



(a) Train. Loss (SVHN) (b) Train. Loss (CIFAR-10) (c) Train. Loss (CIFAR-100)

(d) Val. Loss (SVHN) (e) Val. Loss (CIFAR-10) (f) Val. Loss (CIFAR-100)

Figure 2: Training and Validation Loss per Epoch during the Training Procedure

6. Experiments on Linear Regression with Unbiased Label Noises

To validate our findings in linear regression settings, we carry out numerical evaluation using synthesize data to simply visualize the dynamics over iteration of SGD algorithms with label-noisy OLS and label-noiseless OLS.

6.1. Linear Regression with Unbiased Label Noises

We here hope to see how unbiased label noises would influence SGD iterates for ordinary linear regression (OLS), where a simple quadratic loss function is considered for OLS, such that

$$\widehat{\beta}_{\text{OLS}} \leftarrow \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \tilde{L}_i(\beta) := \frac{1}{2N} \sum_{i=1}^{N} \left(x_i^\top \beta - \tilde{y}_i \right)^2 \right\} , \tag{25}$$

where samples are generated through $\tilde{y}_i = x_i^\top \beta^* + \varepsilon_i$, $\mathbb{E}[\varepsilon_i] = 0$ and $\text{var}[\varepsilon_i] = \sigma^2$. Note that in this section, we replace the notation of θ with β to present the parameters of linear regression models.

Let us combine Eq. (25) and Eq. (11). We write the SGD for Ordinary Least Squares with Unbiased Label Noises as the iterations β_k^{ULN} for $k = 1, 2, 3 \ldots$ as follow

$$\beta_{k+1}^{\text{ULN}} \leftarrow \beta_k^{\text{ULN}} - \frac{\eta}{N} \sum_{i=1}^{N} \nabla L_i^*(\beta_k^{\text{ULN}}) + \sqrt{\eta} \xi_k^*(\beta_k^{\text{ULN}}) + \sqrt{\eta} \xi_k^{\text{ULN}}(\beta_k^{\text{ULN}}), \tag{26}$$

where $\nabla L_i^*(\beta)$ for $\forall \theta \in \mathbb{R}^d$ refers to the loss gradient based on the label-noiseless sample

Accepted Manuscript

(x_i, y_i) and $y_i = x_i^\top \beta^*$, $\xi_k^*(\beta)$ refers to stochastic gradient noises caused by mini-batch sampling over the gradients of label-noiseless samples, and $\xi_k^{\mathrm{ULN}}(\beta)$ is an additional noise term caused by the mini-batch sampling and the unbiased label noises, such that

$$\begin{aligned}
\nabla L_i^*(\beta) &= \frac{\partial}{\partial \beta} \frac{(x_i^\top \beta - x_i^\top \beta^*)^2}{2} = x_i x_i^\top (\beta - \beta^*), \\
\xi_k^*(\beta) &= \frac{\sqrt{\eta}}{|B_k|} \sum_{x_j \in B_k} \left(\nabla L_j^*(\beta) - \frac{1}{N} \sum_{i=1}^{N} L_i^*(\beta) \right), \\
\xi_k^{\mathrm{ULN}}(\beta) &= -\frac{\sqrt{\eta}}{|B_k|} \sum_{x_j \in B_k} \varepsilon_j \cdot x_j \ .
\end{aligned} \tag{27}$$

We denote the *sample covariance matrix* of N samples as

$$\bar{\Sigma}_N = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^\top \ . \tag{28}$$

According to *Remark 1* for implicit regularization in the general form, we can write the implicit regularizer of SGD with the random label noises for OLS as,

$$\xi_k^{\mathrm{ULN}}(\beta) \approx \left(\frac{\eta \sigma^2}{B} \bar{\Sigma}_N \right)^{\frac{1}{2}} z_k' = \left(\frac{\eta \sigma^2}{bN} \sum_{i=1}^{N} x_i x_i^\top \right)^{\frac{1}{2}} z_k', \text{ and } z_k' \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \tag{29}$$

which is unbiased $\mathbb{E}[\xi_k^{\mathrm{ULN}}(\beta)] = \mathbf{0_d}$ with invariant covariance structure, and is independent with β (the location) and k (the time).

Let us combine **Proposition 2** and linear regression settings, we obtain the continuous-time dynamics for linear regression with unbiased label noises, denoted as $\beta^{\mathrm{ULN}}(t)$. According to [33, 56], we can see SGD and its continuous-time dynamics for noiseless linear regression (denoted as β_k^{LNL} and $\beta^{\mathrm{LNL}}(t)$) would asymptotically converge to the optimal solution β^*. As the additional noise term ξ^{ULN} is unbiased with an invariant covariance structure, when $t \to \infty$, we can simply conclude that $\lim_{t\to\infty} \mathbb{E}\ \beta^{\mathrm{ULN}}(t) = \lim_{t\to\infty} \mathbb{E}\ \beta^{\mathrm{LNL}}(t) = \beta^*$, $\lim_{t\to\infty} \mathbf{d}\beta^{\mathrm{ULN}}(t) = (\frac{\eta\sigma^2}{B}\bar{\Sigma}_N)^{1/2} \mathbf{d}W(t)$. By definition of a distribution from a stochastic process, we could conclude $\beta^{\mathrm{ULN}}(t)$ converges to a stationary distribution, such that $\beta^{\mathrm{ULN}}(t) \sim \mathcal{N}(\beta^*, \frac{\eta\sigma^2}{B}\bar{\Sigma}_N)$, as $t \to \infty$. .

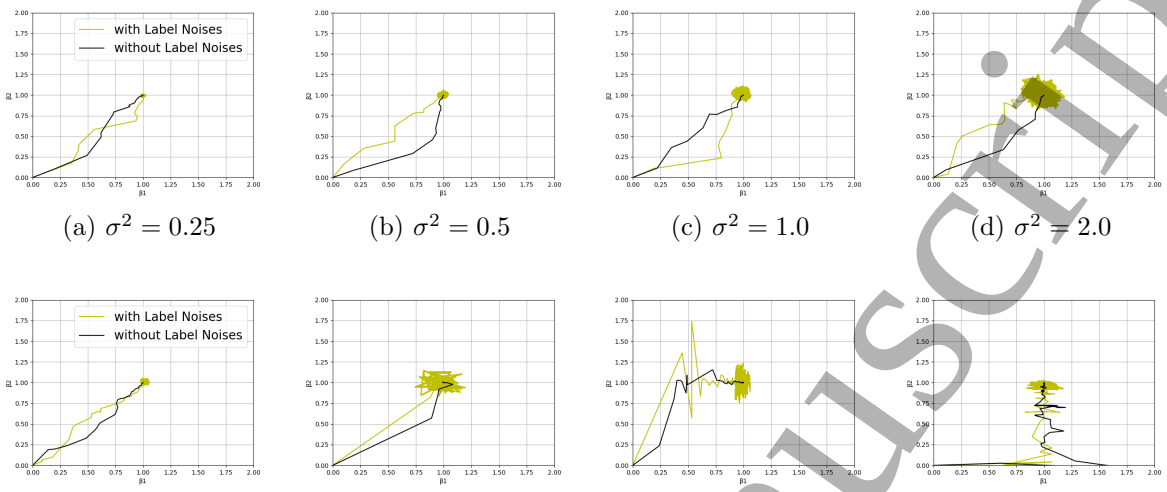
Remark 5. *Thus, with $k \to \infty$, the SGD algorithm for OLS with unbiased label noises would converge to a distribution of Gaussian-alike as follow*

$$\lim_{k\to\infty} \mathbb{E}\ [\beta_k^{\mathrm{ULN}}] = \beta^*, \textit{and} \lim_{k\to\infty} \mathrm{Var}\ [\beta_k^{\mathrm{ULN}}] = \frac{\eta\sigma^2}{B} \bar{\Sigma}_N, \tag{30}$$

The span and shape of the distribution are controlled by σ^2 and $\bar{\Sigma}_N$ when η and B are constant.

In this experiment, we hope to evaluate above remark using numerical simulations, so as to testify (1) whether the trajectories of β_k^{ULN} converges to a distribution of $\mathcal{N}(\beta^*, \frac{\eta\sigma^2}{B}\bar{\Sigma}_N)$; (2) whether the shape of convergence area could be controlled by the sample covariance matrix the data $\bar{\Sigma}_N$ of; and (3) whether the size of convergence area could be controlled by the variance of label noises σ^2.

Accepted Manuscript



(a) $\sigma^2 = 0.25$ (b) $\sigma^2 = 0.5$ (c) $\sigma^2 = 1.0$ (d) $\sigma^2 = 2.0$

(e) $\Sigma_{1,2} = 10 \cdot \mathbf{I}_d$ (f) $\Sigma_{1,2} = 100 \cdot \mathbf{I}_d$ (g) $\Sigma_{1,2} = \Sigma^{\text{Ver}}$ (h) $\Sigma_{1,2} = \Sigma^{\text{Hor}}$

Figure 3: Trajectories of SGD over OLS with and without Unbiased Random Label Noises using various $\tilde{\sigma}^2$ and $\Sigma_{1,2}$ settings for (noisy) random data generation. For Figures (a)–(d), the experiments are setup with a fixed $\Sigma_{1,2} = [[20, 0]^\top, [0, 20]^\top]$ and varying $\tilde{\sigma}^2$. For Figures (e)–(h), the experiments are setup with a fixed $\tilde{\sigma}^2 = 0.5$ and varying $\Sigma_{1,2}$, where we set $\Sigma^{\text{Ver}} = [[10, 0]^\top, [0, 100]^\top]$ and $\Sigma^{\text{Hor}} = [[100, 0]^\top, [0, 10]^\top]$ to shape the converging distributions.

6.2. Experiment Setups

In our experiments, we use 100 random samples realized from a 2-dimension Gaussian distribution $X_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{1,2})$ for $1 \leq i \leq 100$, where $\Sigma_{1,2}$ is an symmetric covariance matrix controlling the random sample generation. To add the noises to the labels, we first draw 100 copies of random noises from the normal distribution with the given variance $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, then we setup the OLS problem with $(X_i, \tilde{y}_i)$ pairs using $\tilde{y}_i = X_i^\top \beta^* + \varepsilon_i$ and $\beta^* = [1, 1]^\top$ and various settings of σ^2 and $\Sigma_{1,2}$. We setup the SGD algorithms with the fixed learning rate $\eta = 0.01$, and batch size $B = 5$, with the total number of iterations $K = 1,000,000$ to visualize the complete paths.

6.3. Experiment Results

Figure 3 presents the results of numerical validations. In Figure 3(a)–(d), we gradually increase the variances of label noises σ^2 from 0.25 to 2.0, where we can observe (1) SGD over label-noiseless OLS converges to the optimal solution $\beta^* = [1.0, 1.0]^\top$ in a fast manner, (2) SGD over OLS with unbiased random label noises would asymptotically converge to a distribution centered at the optimal point, and (3) when σ^2 increases, the span of the converging distribution becomes larger. In Figure 3(e)–(h), we use four settings of $\Sigma_{1,2}$, where we can see (4) no matter how $\Sigma_{1,2}$ is set for OLS problems, the SGD with unbiased random label noises would asymptotically converge to a distribution

Accepted Manuscript

centered at the optimal point.

Compared the results in (e) with(f), we can find that, when the trace of $\Sigma_{1,2}$ increases, the span of converging distributions would increase. Furthermore, (5) the shapes of converging distributions depend on $\Sigma_{1,2}$. In Figure 3(g), when we place the principal component of $\Sigma_{1,2}$ onto the vertical axis (i.e., $\Sigma^{\text{Ver}} = [[10, 0]^\top, [0, 100]^\top]$), the distribution lays on the vertical axis principally. Figure 3(h) demonstrates the opposite layout of the distribution, when we set $\Sigma^{\text{Hor}} = [[100, 0]^\top, [0, 10]^\top]$ as $\Sigma_{1,2}$. The scale and shape of the converging distribution backups our theoretical investigation in Eq (29).

Note that the unbiased random label noises are added to the labels prior to the learning procedure. In this setting, it is the mini-batch sampler of SGD that re-samples the noises and influences the dynamics of SGD through forming the implicit regularizer.

7. Discussion and Conclusion

In this work, we have advanced the understanding of the impact of random label noises in the context of stochastic gradient descent (SGD) under mini-batch sampling. Unlike most prior studies [57],[39] focusing on performance degradation due to label noises or corrupted labels, we intend study to the implicit regularization effects caused by these random label noises.

By adopting a dynamical systems viewpoint of SGD, our work decomposes SGD with unbiased label noises, into three distinct components – $\nabla L^*(\theta)$, the true gradient of label-noiseless loss functions; $\xi_k^*(\theta)$, noise introduced via mini-batch sampling on the label-noiseless loss gradients; and $\xi_k^{\text{ULN}}(\theta)$, noise influenced by both random label noises and mini-batch sampling. The unpublished noise term, $\xi_k^{\text{ULN}}(\theta)$ emerges as an implicit regularizer, contributing to a lower neural network gradient norm, a signature of higher model stability against random parameter perturbations.

Our extensive experiments offer robust testimony for our theoretic analyses. Notably, experiments involving self-distillation with deep neural networks substantiate the role of implicit regularizer in lowering the gradient norm of neural networks and enhancing model stability. This process, enhanced by iteratively adding noises to the outputs of teacher models, eventually results in resolution with better generalization performance. Complementary evidence comes from our linear regression analysis, where the derived SGD dynamics are vividly manifested in the observed trajectories of SGD-based linear regression under the influence of random label noises. These empirical observations echo our theoretical findings, affirming the critical controlling role of data distribution, learning rate, and batch size for the convergence behavior.

To sum up, our work underscores the significance of understanding and harnessing the influence of random label noises in SGD. The emergence of the implicit regularizer, $\xi_k^{\text{ULN}}(\theta)$, could help shape future algorithms to harness noise in a positive way, leading to improved model stability and performance. This study thus opens up new avenues for future research in designing more effective and robust learning procedures.

Accepted Manuscript

References

[1] 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 *ICML*, pages 7654–7663, 2019.

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

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

[4] Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8), 1991.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[6] Stanisław Jastrzebski, 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.

[7] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1731–1741. Curran Associates, Inc., 2017.

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

[9] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[10] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[11] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

[12] Wenqing Hu, Zhanxing Zhu, Haoyi Xiong, and Jun Huan. Quasi-potential as an implicit regularizer for the loss function in the stochastic gradient descent. *arXiv preprint arXiv:1901.06054*, 2019.

[13] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

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

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[16] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Reporesentations (ICLR)*, 2014.

[17] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

[18] 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 (ICML)*, 2020.

[19] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of

Accepted Manuscript

nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1):3–32, 2019.

[20] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

[21] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2019.

[22] Haoyi Xiong, Kafeng Wang, Jiang Bian, Zhanxing Zhu, Cheng-Zhong Xu, Zhishan Guo, and Jun Huan. Sphmc: Spectral hamiltonian monte carlo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5516–5524, 2019.

[23] Zhiwei Jia and Hao Su. Information-theoretic local minima characterization and regularization. In *International Conference on Machine Learning*, pages 4773–4783. PMLR, 2020.

[24] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

[25] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018.

[26] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *AAAI*, 2018.

[27] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. *AISTATS*, 2019.

[28] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3713–3722, 2019.

[29] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation: A simple way for better generalization. *arXiv preprint arXiv:2006.12000*, 2020.

[30] Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, 2020.

[31] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[32] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, pages 2294–2340, 2019.

[33] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020.

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[35] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, 2020.

[36] Oshrat Bar, Amnon Drory, and Raja Giryes. A spectral perspective of dnn robustness to label noise. In *International Conference on Artificial Intelligence and Statistics*, pages 3732–3752. PMLR, 2022.

[37] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[38] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *In International Conference on Learning Representations (ICLR)*, 2016.

Accepted Manuscript

[39] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

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

[41] Haoyi Xiong, Xuhong Li, Boyang Yu, Dejing Dou, Dongrui Wu, and Zhanxing Zhu. Implicit regularization effects of unbiased random label noises with {sgd}, 2021.

[42] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[43] Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5492–5501. PMLR, 2018.

[44] 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.

[45] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[46] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018.

[47] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[48] Charles H Martin and Michael W Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 505–513. SIAM, 2020.

[49] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5):054124, 2022.

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

[51] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019.

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

[53] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[54] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

[55] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[56] Jonas Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(4):1–25, 2021.

[57] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313, 2018.

Accepted Manuscript

[58] Vivek S Borkar and Sanjoy K Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of optimization theory and applications*, 100(3):499–513, 1999.

[59] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

Accepted Manuscript

Appendix A. Proof of Proposition 1

Proof. As the mini-batch B_k are randomly, independently, and uniformly drawn from the full set sample $\mathcal{D}$, thus for $\forall\theta \in \mathbb{R}^d$ and $\forall x_j \in B_k$, there has

$$\mathbb{E}_{x_j\in B_k}\left[\nabla L_j^*(\theta)\right] = \frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta) \Rightarrow \ \mathbb{E}_{B_k}\ \left[\frac{1}{|B_k|}\sum_{x_j\in B_k}\nabla L_j^*(\theta)\right] = \frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta)\ . \tag{A.1}$$

In this way, we can derive that

$$\begin{aligned}\mathbb{E}_{B_k}[\xi_k^*(\theta)] &= \mathbb{E}_{B_k}\ \sqrt{\eta}\left(\frac{1}{|B_k|}\sum_{i=1}^{k}\nabla_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)\\ &= \sqrt{\eta}\left(\mathbb{E}_{B_k}\left[\frac{1}{|B_k|}\sum_{i=1}^{k}\nabla_j^*(\theta)\right] - \frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right) = \mathbf{0_d}\ .\end{aligned} \tag{A.2}$$

Further, for any $\theta \in \mathbb{R}^d$ there has

$$\begin{aligned}\mathrm{Var}[\xi_k^*(\theta)] &= \mathbb{E}_{B_k}\ \left(\xi_k^*(\theta) - \mathbb{E}_{B_k}\xi_k^*(\theta)\right)\left(\xi_k^*(\theta) - \mathbb{E}_{B_k}\xi_k^*(\theta)\right)^\top\\ &= \mathbb{E}_{B_k}\left[\xi_k^*(\theta)\xi_k^*(\theta)^\top\right]\\ &= \frac{\eta}{|B_k|^2}\sum_{x_j\in B_k}\mathrm{Var}[\nabla L_j^*(\theta)]\\ &= \frac{\eta}{|B_k|N}\sum_{j=1}^{N}\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)^\top\\ &= \frac{\eta}{|B_k|}\Sigma_N^{\mathrm{SGD}}(\theta),\end{aligned} \tag{A.3}$$

where $\Sigma_N^{\mathrm{SGD}}(\theta)$ is defined as Eq. (15).

Similarly, as the mini-batch B_k are randomly, independently, and uniformly drawn from the full set sample $\mathcal{D}$, there for $\forall\theta \in \mathbb{R}^d$, there has

$$\mathbb{E}_{x_j\in B_k}\left[\nabla_\theta f(x_j,\theta)\right] = \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\ . \tag{A.4}$$

Thus, there has

$$\mathbb{E}_{B_k,\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] = -\frac{\sqrt{\eta}}{|B_k|}\sum_{x_i\in B_k}\left\{\mathbb{E}_{x_i\in B_k}\left[\nabla f(x_i,\theta)\right]\cdot\mathbb{E}_{\varepsilon_i}[\varepsilon_i]\right\} = \mathbf{0_d}\ . \tag{A.5}$$

Accepted Manuscript

Again, by the definition, there has

$$\begin{aligned}\mathrm{Var}_{B_k,\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] &= \mathbb{E}_{B_k,\varepsilon_i}\left[\xi_k^{\mathrm{ULN}}(\theta)\xi_k^{\mathrm{ULN}}(\theta)^\top\right] \\ &\quad \text{Let us assume } B_k \text{ and } \varepsilon_i \text{ for } 1\le i\le N \text{ are independent.} \\ &= \frac{\eta}{|B_k|^2}\sum_{x_j\in B_k}\mathbb{E}_{B_k\varepsilon_i}\left[(\varepsilon_i\cdot\nabla_\theta f(x_i,\theta)-\mathbb{E}_{B_k\varepsilon_i}(\varepsilon_i\cdot\nabla_\theta f(x_i,\theta))^2\right] \\ &\quad \text{As } \mathrm{Var}[\varepsilon_i]=\sigma^2 \text{ for } 1\le i\le N, \text{ there has} \\ &= \frac{\eta\sigma^2}{|B_k|N}\sum_{i=1}^N \nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top = \frac{\eta\sigma^2}{|B_k|}\Sigma_N^{\mathrm{ULN}}(\theta)\ .\end{aligned} \tag{A.6}$$

where $\Sigma_N^{\mathrm{ULN}}(\theta)$ is defined as Eq. (15). □

Appendix B. Proof of Proposition 2

Proof. We show that, as $\eta\to 0$, the discrete iteration $\bar\theta_k$ of Eq. (17) in strong norm and on finite–time intervals is close to the solution of the SDE. (16). The main techniques follow [58], but [58] only considered the case when $\Sigma_N^{\mathrm{SGD}}(\theta)$ and $\Sigma_N^{\mathrm{ULN}}(\theta)$ are constants.

Let $C_1(\theta)=\sqrt{\frac{1}{B}\Sigma_N^{\mathrm{SGD}}(\theta)}$, $C_2(\theta)=\sqrt{\frac{1}{B}\Sigma_N^{\mathrm{ULN}}(\theta)}$ and $L^*(\theta)=\frac{1}{N}\sum_{i=1}^N L_i^*(\theta)$. Let $\widehat\Theta(t)$ be the process defined by the integral form of the stochastic differential equation

$$\begin{aligned}\widehat\Theta(t)-\widehat\Theta(0) = &-\int_0^t \nabla L^*(\widehat\Theta_{\lfloor\frac{s}{\eta}\rfloor\eta})\mathbf{d}s+\sqrt{\eta}\int_0^t C_1(\widehat\Theta(\lfloor\frac{s}{\eta}\rfloor\eta))\mathbf{d}W_1(s) \\ &+\sqrt{\eta}\int_0^t C_2(\widehat\Theta(\lfloor\frac{s}{\eta}\rfloor\eta))\mathbf{d}W_2(s)\ ,\ \widehat\Theta(0)=\theta_0\ .\end{aligned} \tag{B.1}$$

Here for a real positive number $a>0$ we define $\lfloor a\rfloor=\max\{k\in\mathbb{N}_+, k<a\}$. From Eq.(B.1) we see that we have, for $k=0,1,2,...$

$$\begin{aligned}\widehat\Theta((k+1)\eta)-\widehat\Theta(k\eta) = &-\eta\nabla L^*(\widehat\Theta(k\eta))-\sqrt{\eta}C_1(\widehat\Theta(k\eta))(W_1((k+1)\eta)-W_1(k\eta)) \\ &-\sqrt{\eta}C_2(\widehat\Theta(k\eta))(W_2((k+1)\eta)-W_2(k\eta)).\end{aligned} \tag{B.2}$$

Since $\sqrt{\eta}(W_1((k+1)\eta)-W_1(k\eta))\sim\mathcal{N}(0,\eta^2 I)$ and $\sqrt{\eta}(W_2((k+1)\eta)-W_2(k\eta))\sim\mathcal{N}(0,\eta^2 I)$, we could let $\eta z_{k+1}=\sqrt{\eta}(W_1((k+1)\eta)-W_1(k\eta))$ and $z'_{k+1}=\sqrt{\eta}(W_2((k+1)\eta)-W_2(k\eta))$, where z_{k+1} and z'_{k+1} are the i.i.d. Gaussian sequences in (16). From here, we see that

$$\widehat\Theta(k\eta)=\bar\theta_k^{\mathrm{ULN}}\ , \tag{B.3}$$

where $\bar\theta_k^{\mathrm{ULN}}$ is the solution to (17).

We then try to bound $\widehat\Theta_t$ in Eq. (B.1) and $\Theta^{\mathrm{ULB}}(t)$ in Eq. (16). Finally we could obtain the error estimation of $\bar\theta_k^{\mathrm{ULN}}=\widehat\Theta(k\eta)$ and $\Theta^{\mathrm{ULN}}(k\eta)$ by simply set $t=k\eta$. Since we assumed that $\nabla L_i^*(\theta)$ and $\nabla_\theta f(x,\theta)$ are L–Lipschitz continuous, we get

$$\|C_1(\theta_1)-C_1(\theta_2)\|_2=\sqrt{\frac{1}{bN}\sum_{i=1}^N\|\nabla L_i^*(\theta_1)-\nabla L_i^*(\theta_2)\|_2^2}\le L\|\theta_1-\theta_2\|_2 \tag{B.4}$$

Accepted Manuscript

since the batch size $b \geq 1$. In the same way,

$$\|C_2(\theta_1) - C_2(\theta_2)\|_2 = \sqrt{\frac{\sigma^2}{bN}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta_1) - \nabla_\theta f(x_i,\theta_2)\|_2^2} \leq \sigma L\|\theta_1 - \theta_2\|_2 \tag{B.5}$$

since the batch size $B \geq 1$. Thus $C_1(\theta)$ and C_2 are both L–Lipschitz continuous. Take a difference between (B.1) and (16) we get

$$\begin{aligned}\widehat{\Theta}(t) - \Theta^{\mathrm{ULN}}(t) = &-\int_0^t [\nabla L^*(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}s \\ &+ \sqrt{\eta}\int_0^t [C_1(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor)) - C_1(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s) \\ &+ \sqrt{\eta}\int_0^t [C_2(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor)) - C_2(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_2(s).\end{aligned} \tag{B.6}$$

We can estimate

$$\begin{aligned}&\|\nabla L^*(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2\|\nabla L^*(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|\nabla L^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s)\|_2^2\ ,\end{aligned} \tag{B.7}$$

where we used the inequality derived from L-Lipschitz. Similarly, we estimate

$$\begin{aligned}&\|C_1(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_1(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2\|C_1(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_1(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|C_1(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2L^2\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s))\|_2^2\ .\end{aligned} \tag{B.8}$$

In the same way, based on the inequality derived from L-Lipschitz, we can also have

$$\begin{aligned}&\|C_2(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_2(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2\|C_2(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_2(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|C_2(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \Theta^{\mathrm{ULN}}(s))\|_2^2 \\ \leq&2\sigma^2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2\sigma^2L^2\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s))\|_2^2\ .\end{aligned} \tag{B.9}$$

On the other hand, from (B.6), the Itô's isometry [59] and Cauchy–Schwarz

Accepted Manuscript

inequality we have

$$\begin{aligned}\mathbb{E}|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)|^2 \leq & 2\mathbb{E}\left\|\int_0^t[\nabla L(^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta))-\nabla L^*(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}s\right\|_2^2 \\ & +2\eta\mathbb{E}\left\|\int_0^t[C_1(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor})-C_1(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s)\right\|^2 \\ & +2\eta\mathbb{E}\left\|\int_0^t[C_2(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor})-C_2(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s)\right\|^2 \\ \leq & 2\int_0^t\mathbb{E}\left\|\nabla L(^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta))-\nabla L^*(\Theta^{\mathrm{ULN}}(s))\right\|_2^2\mathbf{d}s \\ & +2\eta\int_0^t\mathbb{E}\left\|C_1(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor})-C_1(\Theta^{\mathrm{ULN}}(s))\right\|^2\mathbf{d}s \\ & +2\eta\int_0^t\mathbb{E}\left\|C_2(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor})-C_2(\Theta^{\mathrm{ULN}}(s))\right\|^2\mathbf{d}s\ .\end{aligned} \tag{B.10}$$

Combining Eqs. (B.7), (B.8), (B.9) and (B.10) we obtain that

$$\begin{aligned}&\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2 \\ &\leq 2\int_0^t\left(2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2L^2\mathbb{E}\|\Theta_{\lfloor\frac{s}{\eta}\rfloor\eta}-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s \\ &+2\eta\int_0^t\left(2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2L^2\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s \\ &+2\eta\int_0^t\left(2\sigma^2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2\sigma^2L^2\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s\ . \\ &=4(1+\eta+\eta\sigma^2)L^2\cdot \\ &\left(\int_0^t\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2\mathbf{d}s+\int_0^t\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\mathbf{d}s\right)\ .\end{aligned} \tag{B.11}$$

Since we assumed that there is an $M>0$ such that $\max_{i=1,2\ldots,N}(|\nabla L_i^*(\theta)|)\leq M$, we conclude that $|\nabla L^*(\theta)|\leq\frac{1}{N}\sum_{i=1}^N|\nabla L_i^*(\theta)|\leq M$ and

$$\|C_1(\theta)\|_2\leq\frac{1}{\sqrt{b}}\sqrt{\frac{1}{N}\sum_{i=1}^N\|\nabla L_i^*(\theta)\|_2^2}\leq M, \tag{B.12}$$

$$\|C_2(\theta)\|_2\leq\frac{1}{\sqrt{b}}\sqrt{\frac{\sigma^2}{N}\sum_{i=1}^N\|\nabla_\theta f(x_i,\theta)\|_2^2}\leq\sigma M \tag{B.13}$$

since $B\geq 1$. By Eq. (16), the Itô's isometry [59], the Cauchy-Schwarz inequality and $0\leq s-\lfloor\frac{s}{\eta}\rfloor\eta\leq\eta$ we know that

Accepted Manuscript

$$
\begin{aligned}
&\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\\
=&\mathbb{E}\left\|-\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\nabla L^*(\Theta_u)\mathbf{d}u+\sqrt{\eta}\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}C_1(\Theta_u)\mathbf{d}W_1(u)+\sqrt{\eta}\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}C_2(\Theta_u)\mathbf{d}W_2(u)\right\|_2^2\\
\leq&2\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\nabla L^*(\Theta_u)\mathbf{d}u\right\|_2^2+2\eta\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}C_1(\Theta_u)\mathbf{d}W_1(u)\right\|_2^2+2\eta\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}C_2(\Theta_u)\mathbf{d}W_2(u)\right\|_2^2\\
\leq&2\mathbb{E}\left(\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\|\nabla L^*(\Theta_u)\|\,\mathbf{d}u\right)^2+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\mathbb{E}\|C_1(\Theta_u)\|_2^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\mathbb{E}\|C_2(\Theta_u)\|_2^2\mathbf{d}u\\
\leq&2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\mathbb{E}|\nabla L^*(\Theta_u)|^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\mathbb{E}\|C_1(\Theta_u)\|_2^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^{s}\mathbb{E}\|C_2(\Theta_u)\|_2^2\mathbf{d}u\\
\leq&2\eta^2M^2+2\eta^2M^2+2\eta^2\sigma^2M^2=(4+2\sigma^2)\eta^2M^2\ .
\end{aligned}
\tag{B.14}
$$

Combining (B.14) and (B.11) we obtain

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2\\
\leq&4(1+\eta+\eta\sigma^2)L^2\cdot\left(\int_0^t\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2\mathbf{d}s+(4+2\sigma^2)\eta^2M^2t\right)\ .
\end{aligned}
\tag{B.15}
$$

Set $T>0$ and $m(t)=\max_{0\leq s\leq t}\mathbb{E}|\widehat{\Theta}_s-\Theta^{\mathrm{ULN}}(s))|^2$, noticing that $m(\lfloor\frac{s}{\eta}\rfloor\eta)\leq m(s)$ (as $\lfloor\frac{s}{\eta}\rfloor\eta\leq s$), then the above gives for any $0\leq t\leq T$,

$$
m(t)\leq 4(1+\eta+\eta\sigma^2)L^2\cdot\left(\int_0^t m(s)\mathbf{d}s+(4+2\sigma^2)\eta^2M^2T\right)\ . \tag{B.16}
$$

By Gronwall's inequality we obtain that for $0\leq t\leq T$,

$$
m(t)\leq 4(1+\eta+\eta\sigma^2)(4+2\sigma^2)L^2\eta^2M^2Te^{4(1+\eta+\eta\sigma^2)L^2t}. \tag{B.17}
$$

Suppose $0<\eta<1$, then there is a constant C which is independent on η s.t.

$$
\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2\leq m(t)\leq C\eta^2. \tag{B.18}
$$

Set $t=k\eta$ in (B.18) and make use of (B.3), we finish the proof. □

Appendix C. Proof of Proposition 3

Proof. To obtain Eq (22), we can use the simple vector-matrix-vector products transform that, for the random vector v and symmetric matrix A there has $\mathbb{E}_v[v^\top Av]=$

Accepted Manuscript

$\text{trace}(A\mathbb{E}[vv^\top])$, such that

$$\begin{aligned}
\mathbb{E}_{z_k}\|\xi_k^{\text{ULN}}(\theta)\|_2^2 &= \mathbb{E}_{z_k}\left[\xi_k^{\text{ULN}}(\theta)^\top \xi_k^{\text{ULN}}(\theta)\right] \\
&\approx \frac{\eta\sigma^2}{b}\mathbb{E}_{z_k}\left[z_k^\top\left(\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)z_k\right] \\
&= \frac{\eta\sigma^2}{b}\text{trace}\left(\left(\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)\mathbb{E}_{z_k}[z_k z_k^\top]\right) \\
&\text{as } \mathbb{E}_{z_k}[z_k z_k^\top] = \mathbf{I}_d \ \textit{for} \ z_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) \\
&= \frac{\eta\sigma^2}{b}\text{trace}\left(\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right) \\
&= \frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2 .
\end{aligned} \tag{C.1}$$

□

Accepted Manuscript