Fast Mixing of Stochastic Gradient Descent with Normalization and Weight Decay

Anonymous Author(s)
Affiliation
Address
email

Abstract

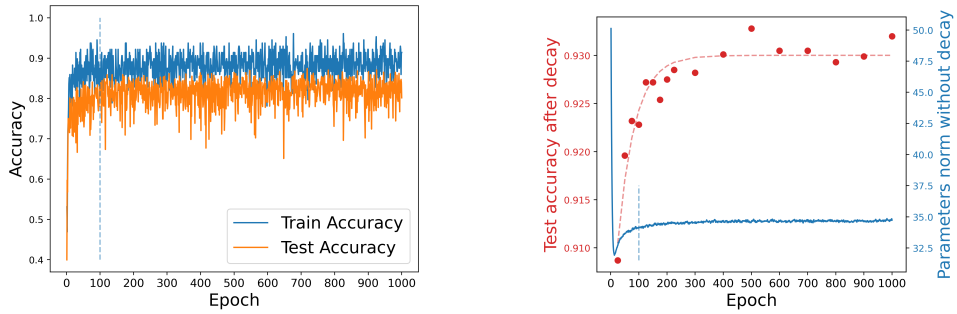
We prove the Fast Equilibrium Conjecture [26], *i.e.*, Stochastic Gradient Descent (SGD) for scale-invariant loss (*e.g.*, networks with various normalization schemes) with learning rate η and weight decay factor λ mixes in function space in $\tilde{\mathcal{O}}(\frac{1}{\lambda\eta})$ steps, under two standard assumptions: (1) the noise covariance matrix is non-degenerate and has constant trace and (2) all minimizers of the loss form a connected and smooth manifold. The analysis uses the framework from [27] and shows that for every $T > 0$, the iterates of SGD with learning rate η and weight decay factor λ on the scale-invariant loss converge in distribution in $\Theta\left(\eta^{-1}\lambda^{-1}(T + \ln(\lambda/\eta))\right)$ iterations as $\eta\lambda \to 0$ (while simultaneously satisfying $\eta \lesssim \lambda$ and $\lambda \lesssim 1$). Moreover, the evolution of limiting distribution can be described by an Ito's stochastic differential equation (SDE) which mixes to a unique distribution in time independent of η.

1 Introduction

Normalization methods such as Batch Normalization [14], Group Normalization [39], Layer Normalization [2] and Weight Normalization [32] have facilitated the efficient training of modern deep neural networks, and in the extensive tool case stochastic gradient gradient (SGD) with weight decay (WD) is one of the main workhorses. However, the interplay between normalization, gradient noise and WD can move the training dynamics far away from standard optimization viewpoint, *e.g.*, deep neural networks with normalization can be trained with an exponentially increasing learning rate (LR) schedule [25].

There is a line of research [3, 17] suggesting that small LRs (or equivalently large batch sizes) might hurt generalization since they lead to the so-called *sharp minimizers*. It is argued that noise can help avoid sharp minima [11, 17, 22, 15, 9], so large LRs and small batches are preferred as they increase the noise during training. However, later Li et al. [26] observed that normalized models trained using SGD+WD can generalize equally well with small LRs, given longer training time after the training loss converges. Such discrepancies imply an incomplete characterization of what happens during training and especially the affect of the gradient noise. To explain the effectiveness of small LRs, Li et al. [26] theoretically identified an intrinsic learning rate parameter equal to $\eta\lambda$ (η is the LR and λ is the WD factor) which governs the convergence rate of the norm of the weights. Based on empirical evidence, they further proposed the *fast equilibrium conjecture* which states that the mixing time of the normalized model is roughly $\widetilde{O}\left(1/(\eta\lambda)\right)$, much faster than existing upper bounds. This conjecture is important because there is some unexplained mechanism towards better generalization happening during the mixing process — the training accuracy/loss plateaus but the test accuracy

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Train and test accuracy for CIFAR-10 training with $\eta = 0.8$, $\lambda = 5 \cdot 10^{-4}$.

(b) Test accuracy after LR decay and the total norm of parameters before LR decay.

Figure 1: The train and test accuracy plateaus after parameter norm convergence within 100 epochs, but the generalization of SGD iterate after LR decay keeps improving. Figure 1a shows the train and test accuracy of scale invariant PreResNet trained with standard data augmentation on CIFAR-10. Each red dot in Figure 1b represents the test accuracy of model which decays LR to 10^{-3} at the corresponding epoch. The test accuracy is evaluated until achieving full training accuracy after LR decay.

after LR decay still improves (see Figure 1 for a demonstration). The fast mixing time can provide theoretical guidelines to decide the proper time of LR decay and thus saves training time.

Before stating the fast equilibrium conjecture, let us first specify the problem settings. We consider a family of models F_x whose output is scale-invariant to its parameters x, i.e., $F_x(z) = F_{Cx}(z)$ for any parameters x, input data z and constant $C > 0$. This scale-invariance property is commonly induced by various normalization schemes, and makes the loss function for any data (or batch of data) also scale invariant to x. Let Ξ be the number of training samples and the empirical loss $L(x) = \frac{1}{\Xi}\sum_{i=1}^{\Xi} L_i(x)$ where each L_i is the scale invariant loss over the i-th sample.[1] We optimize L using SGD+WD. The dynamics is formalized as follows:

$$x_{\eta,\lambda}(k+1) = (1-\eta\lambda)x_{\eta,\lambda}(k) - \eta(\nabla L(x_{\eta,\lambda}(k)) + \sqrt{\Xi}\sigma_{\xi_k}(x_{\eta,\lambda}(k))) \tag{1}$$

where η is the learning rate (LR), λ is the WD factor, $x_{\eta,\lambda}(k) \in \mathbb{R}^D$ denote the parameters after k iterations, $x_{\eta,\lambda}(0) = x_0$ is the initialization of the parameters, and $\sigma(x) = [\sigma_1(x), \ldots, \sigma_\Xi(x)] \in \mathbb{R}^{D\times\Xi}$ is the deterministic noise function with each $\sigma_i(x) := \frac{1}{\sqrt{\Xi}}(\nabla L_i(x) - \nabla L(x))$. Here each ξ_k is sampled uniformly from $\{1, \ldots, \Xi\}$, so $\mathbb{E}_{\xi_k}[\sigma_{\xi_k}(x)] = 0$ for all $x \in \mathbb{R}^D$ and k.

As a common approach to analyzing SGD, the canonical stochastic differential equation (SDE) approximation of SGD+WD in Equation (1) is

$$\mathrm{d}X_{\eta,\lambda}(t) = -\eta\nabla L(X_{\eta,\lambda}(t))\mathrm{d}t - \eta\lambda X_{\eta,\lambda}(t)\mathrm{d}t - \eta\sigma(X_{\eta,\lambda}(t))\mathrm{d}W(t) \tag{2}$$

where $\{W(t)\}_{t\geq 0}$ is the Ξ-dimensional Brownian motion.

Conjecture 1.1 (Li et al. [26])**.** Suppose $X_{\eta,\lambda}(t)$ is a solution of Equation (2), then $F_{X_{\eta,\lambda}(t)}$ converges to an equilibrium in $\widetilde{O}(1/(\lambda\eta))$ time.

This convergence rate is much faster than the standard $e^{-\Theta(1/\eta)}$ global mixing time of Langevin dynamics and that's how the conjecture gets the name. Li et al. [26] proved that the norm of $X_{\eta,\lambda}(t)$ will converge in $\widetilde{O}(1/(\lambda\eta))$ time but failed to show convergence to the equilibrium. Another gap is that the analysis in Li et al. [26] only works for the SDE approximation in Equation (2), and does not bridge between the discrete-time and continuous-time dynamics.

In this paper, we provide a partial proof to this conjecture.[2]

[1] When using Batch Normalization, the loss is only defined for a batch of data, and we can think L_i as the loss for the i-th possible batch.

[2] We treat the expected loss landscape, the distribution of the stochastic loss and initialization as constant and only focus on the convergence rate depending on η and λ.

Theorem 1.2 (Informal version of Theorem 6.7)**.** *Let* $\overline{x}_{\eta,\lambda} := \frac{x_{\eta,\lambda}}{\|x_{\eta,\lambda}\|_2}$ *and* $x_{\eta,\lambda}(0) \equiv x_0$ *for all* $\eta, \lambda > 0$ *for SGD+WD in* (1)*. Under technical assumptions, there is a unique distribution* π *such that in the limit of* $\eta\lambda \to 0$ *and* $T \to \infty$*, the distribution of* $\overline{x}_{\eta,\lambda}(\lfloor \frac{\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)}{4\eta\lambda} \rfloor)$ *can be arbitrarily "close" to* π *in the sense of weak convergence for large enough* T*.*

The **main contributions** of this paper are summarized as follows:

1. We give a SDE-based characterization (Theorem 5.4) for the limiting dynamics of SGD+WD for a scale-invariant stochastic loss in the limit of $\eta\lambda \to 0$ with $\eta = O(\lambda)$ and $\lambda = O(1)$. In detail, we show the direction of SGD iterate at step $\frac{1}{4\eta\lambda}\ln\left(\frac{2\lambda}{\eta}(e^{2T}-1)+1\right)$ converges in distribution to some random variable Y_T, where Y_T is the solution of some SDE depending on the stochastic loss and initialization. This result can not be proved by Li et al. [27] directly as there is no minimizer in the usual sense, not to mention manifold. The key proof technique is to introduce a novel time-rescaling (Equation (9)) to make the convergence result in [16] applicable.

2. We show SGD without WD for stochastic scale invariant loss has the same limiting dynamics as SGD+WD, but is exponentially slowed down (see Theorem 4.2). This is consistent to the empirical observation [41, 26] that turning on WD for SGD for scale invariant loss helps generalization.

3. Under the assumption of all minimizers forming a compact manifold and noise being non-degenerate in the tangent space of the manifold, we show that from any initialization, the limiting diffusion process converges to a unique stationary distribution (see Theorem 6.6 and Theorem 6.7). This delivers a partial proof to the Fast Equilibrium Conjecture in [26].

4. Though our convergence result is asymptotic, we verify in simplified settings that the predicted phenomena happens with LR η and WD λ of practical scale (see Section 7 for details of experiments). We also show empirically that the mixing process exists in practical settings, and it is beneficial for generalization.

2 Related Work

Scale-invariance. For scale-invariant loss functions, existing works have proved the convergence of GD and SGD with WD [1]. It has been shown that scale-invariance enables robust and efficient training of SGD with WD [28]. Scale-invariance also brings about the interesting equivalence between the effect of WD and LR schedules studied in [12, 41, 25]. Moreover, for SGD with WD, η being the LR and λ being the WD factor, the norm of the parameters will converge to $(\lambda/\eta)^{1/4}$ [26, 35, 36], and this induces the intrinsic LR which is equal to $\eta\lambda$ [26]. These observations are crucial to our derivations in the following sections.

SDE approximation. Continuous-time tools such as SDE have been a popular lens for studying optimization algorithms including SGD [29, 23, 33, 24, 4, 34]. Many interesting properties of SGD have been discovered through this approach [19, 8, 42, 26, 40, 31, 38]. Recently, [27] proposed a framework to study the implicit bias of SGD by showing that the iterates converge in distribution to a limiting diffusion as LR goes to 0. Using this framework, we exhibit the limiting dynamics of SGD+WD for infinitesimal intrinsic LR.

3 Preliminary

We first clarify the notations. We denote by $\mathbb{N}$ the set of all nonnegative integers and $\mathbb{R}_+$ the set of all nonnegative real numbers. For any $k \in \mathbb{N}$, we denote by $\mathcal{C}^k$ the set of all k times continuously differentiable functions. For any vector $u \in \mathbb{R}^D$, we denote its i-th coordinate by u_i. For any mapping $F : \mathbb{R}^D \to \mathbb{R}^D$, we denote the *Jacobian* of F at x by $\partial F(x) \in \mathbb{R}^{D\times D}$ where the (i,j)-th entry is $\partial_j F_i(x)$. We also use $\partial F(x)[u]$ and $\partial^2 F(x)[u,v]$ to denote the first and second order directional derivatives of F at x along the derivation of u (and v). With a slight abuse of notation, we view $\partial^2 F$ as a linear mapping on $\mathbb{R}^{D^2}$ such that $\partial^2 F(x)[A] = \sum_{i,j=1}^D \partial^2 F(x)[e_i, e_j]A_{ij}$, for any $A \in \mathbb{R}^{D\times D}$. For any submanifold $\Gamma \subset \mathbb{R}^D$ and $x \in \Gamma$, we denote by $T_x(\Gamma)$ the tangent space of Γ at x. We denote by $\mathbb{1}_\xi \in \mathbb{R}^\Xi$ the one-hot vector where the ξ-th coordinate is 1, and $\mathbb{1}$ denotes the all 1 vector.

3.1 Loss function and manifold of local minimizers

Since this paper deals with scale-invariant loss, most sets in our analysis appear naturally as cones.

Definition 3.1 (Cone)**.** We say $K \subset \mathbb{R}^D$ is a *cone* if and only if $0 \notin K$ and $\forall \alpha > 0$, $\alpha K \subset K$.

Assumption 3.2. The loss function is $L(x) = \frac{1}{\Xi}\sum_{i=1}^{\Xi} L_i(x)$, where each L_i is the loss on the i-th sample and is scale-invariant, that is, $L_i(x) = L_i(\alpha x)$ for all $x \in \mathbb{R}^D \setminus \{0\}$ and $\alpha > 0$.

Recall that each $\sigma_i(x) = \frac{1}{\sqrt{\Xi}}(\nabla L_i(x) - \nabla L(x))$, so the noise function σ is also scale-invariant. We further define $\Sigma(x) = \sigma(x)\sigma(x)^\top$, i.e., the noise covariance at x. With vast overparametrization in modern machine models, multiple minimizers can exist and form a manifold [5, 6], which motivates the following assumption.

Assumption 3.3. The loss function $L : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}$ is a $\mathcal{C}^3$ function. Γ is a $(D-M)$-dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^D$ for some integer $0 \le M \le D$, where each $x \in \Gamma$ is a local minimizer of L and $\mathrm{rank}(\nabla^2 L(x)) = M$ for all $x \in \Gamma$. Since we assumed $L(x)$ is scale invariant in Assumption 3.2, we also assume Γ is a cone.

Next, we formalize the definition of the gradient flow (GF) with respect to L and its associated limiting map, which plays a key role in our analysis.

Definition 3.4. For any initialization $x \neq 0$, we define the GF governed by $-\nabla L$ as $\phi(x,t) = x - \int_0^t \nabla L(\phi(x,s))\mathrm{d}s$ for $t \ge 0$, and its associated limiting map $\Phi(x) = \lim_{t\to\infty}\phi(x,t)$ whenever the limit exists.

Assumption 3.5. U is a cone and an open neighborhood of Γ satisfying that GF starting in U converges to some point in Γ, i.e., $\forall x \in U$, $\Phi(x) \in \Gamma$.

Li et al. [27] established several important properties of Φ by relating the derivatives of Φ to those of L, and in particular, we recall the following characterization of $\partial\Phi$.

Lemma 3.6 (Lemma 4.2, [27])**.** *For any $x \in \Gamma$, $\partial\Phi(x) \in \mathbb{R}^{D\times D}$ is the projection matrix onto the tangent space $T_x(\Gamma)$.*

The following properties of L and Φ are peculiar in our setting thanks to scale-invariance.

Lemma 3.7. *Under Assumption 3.2, for any $\alpha > 0$ and $x \in \mathbb{R}^D \setminus \{0\}$, it holds that $\nabla^k L(\alpha x) = \alpha^{-k}\nabla^k L(x)$ for any $k \in \mathbb{N}$. Moreover, for any $x \in U$ where U is defined in Assumption 3.5, we have $\partial\Phi(\alpha x) = \partial\Phi(x)$ and $\partial^2\Phi(\alpha x) = \alpha^{-1}\partial^2\Phi(x)$. Also, $\partial\Phi(x)x = x$ for any $x \in \Gamma$.*

3.2 Limiting diffusion on the manifold of local minimizers

We recap the notion of *Katzenberger processes* proposed in [27] and the characterization of the corresponding limiting diffusion based on Katzenberger's theorems in [16].

For each $n \in \mathbb{N}$, let $A_n : \mathbb{R}_+ \to \mathbb{R}_+$ and $B_n : \mathbb{R}_+ \to \mathbb{R}_+$ be two non-decreasing functions with $A_n(0) = B_n(0) = 0$, and $\{Z_n(t)\}_{t\ge0}$ a $\mathbb{R}^\Xi$-valued stochastic process. Given loss function $L : \mathbb{R}^D \to \mathbb{R}$ and noise function $\sigma : \mathbb{R}^D \to \mathbb{R}^{D\times\Xi}$, we consider the following stochastic process:

$$X_n(t) = X_n(0) + \int_0^t \sigma(X_n(s))\mathrm{d}Z_n(s) + \int_0^t X_n(t)\mathrm{d}B_n(s) - \int_0^t \nabla L(X_n(s))\mathrm{d}A_n(s) \tag{3}$$

where $X_n(0) \in U$. Further define $Y_n(t) = X_n(t) - \phi(X_n(0), A_n(t)) + \Phi(X_n(0))$.

Theorem 3.8 ([27])**.** *Fix some initialization $X_n(0) \equiv x_0 \in U$ for the stochastic process in Equation (3). Suppose $\{X_n(t)\}_{t\ge0}$ is a Katzenberger process satisfying as $n \to \infty$,*

- *A_n increases infinitely fast, i.e., $\forall \epsilon > 0, \inf_{t\ge0}(A_n(t+\epsilon) - A_n(t)) \to \infty$;*
- *$B_n(t)$ converges to $c \cdot t$ in uniform metric for some constant c.*
- *Z_n converges in distribution to $(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)W$ in Skorokhod metricwhere W is a Ξ-dimension standart Brownian motion;*

If SDE (4) has a global solution Y with $Y(0) = \Phi(x_0)$, then for any $t > 0$, as $n \to \infty$, $Y_n(t)$ converges in distribution to $Y(t)$.

$$Y(t) = \Phi(x_0) + \int_0^t c\partial\Phi(Y(s))Y(s)\mathrm{d}s + \int_0^t \partial\Phi(Y(s))\sigma(Y(s))\mathrm{d}W(s) + \int_0^t \frac{1}{2}\partial^2\Phi(Y(s))[\Sigma(Y(s))]\mathrm{d}s. \tag{4}$$

4 Warm-up: Simultaneous Limit Case

We first consider the setting that $\eta, \lambda \to 0$ simultaneously with $\frac{\lambda}{\eta} = C$ for some constant $C \geq 0$. We fix $x_{\eta,\lambda}(0) = x_0$ for some x_0. Now, define $X_{\eta,\lambda}(t) = x_{\eta,\lambda}(\lfloor t/\eta^2 \rfloor)$, which is roughly equivalent to SDE (2) with $1/\eta^2$ times acceleration. We rewrite the discrete-time update of $x_{\eta,\lambda}$ as

$$\mathrm{d}X_{\eta,\lambda}(t) = -\nabla L(X_{\eta,\lambda})\mathrm{d}A_{\eta,\lambda}(t) - \frac{\lambda}{\eta}X_{\eta,\lambda}\mathrm{d}B_{\eta,\lambda}(t) - \sigma(X_{\eta,\lambda})\mathrm{d}Z_{\eta,\lambda}(t) \tag{5}$$

where $A_{\eta,\lambda}, B_{\eta,\lambda}$ and $Z_{\eta,\lambda}$ are defined by

$$A_{\eta,\lambda}(t) = \eta\lfloor t/\eta^2 \rfloor, \quad B_{\eta,\lambda}(t) = \eta^2\lfloor t/\eta \rfloor, \quad Z_{\eta,\lambda}(t) = \eta \sum_{k=1}^{\lfloor t/\eta^2 \rfloor} \sqrt{\Xi}\left(\mathbb{1}_{\xi_k} - \frac{1}{\Xi}\mathbb{1}\right). \tag{6}$$

Note that $A_{\eta,\lambda}(t)$ is roughly t/η which becomes very large for small η, thus the negative gradient part will drive $X_{\eta,\lambda}(t)$ rapidly towards the manifold Γ and force $X_{\eta,\lambda}(t)$ to stay close to Γ after that. On the other hand, as $\eta \to 0$, $B_{\eta,\lambda}(t)$ will converge to t and $Z_{\eta,\lambda}$ will weakly converge to a Brownian motion, and these terms make up the slow dynamics of SGD. More precisely, we have the following lemma summarizing the properties of these integrators, which shows Equation (5) is a valid Katzenberger process (Equation (6) in Li et al. 27).

Lemma 4.1. *Let $A_{\eta,\lambda}, B_{\eta,\lambda}$ and $Z_{\eta,\lambda}$ be as defined in Equation (6). Then as $\eta, \lambda \to 0$ with $\frac{\lambda}{\eta} \equiv C$, we have (1) $A_{\eta,\lambda}$ increases infinitely fast, (2) $B_{\eta,\lambda}(t)$ converges to t in uniform metric and (3) $Z_{\eta,\lambda}$ converges in distribution to $(I_\Xi - \frac{1}{\Xi^2}\mathbb{1}\mathbb{1}^\top)^{1/2}W$ in Skorokhod metric where $\{W(t)\}_{t\geq 0}$ is the Ξ-dimensional standard Brownian motion.*

Therefore, a direct application of Theorem 3.8 yields the following.

Theorem 4.2. *Under Assumption 3.2, 3.3 and 3.5, fix some $x_0 \in U$ and let $x_{\eta,\lambda}(0) \equiv x_0$ for all $\eta, \lambda > 0$ for SGD with weight decay (1). Consider*

$$\mathrm{d}Y_C(t) = -C\partial\Phi(Y_C)Y_C\mathrm{d}t + \frac{1}{2}\partial^2\Phi(Y_C)[\Sigma(Y_C)]\mathrm{d}t - \partial\Phi(Y_C)\sigma(Y_C)\mathrm{d}W(t) \tag{7}$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^\Xi$. If SDE (7) has a global solution Y_0 (for $C = 0$) with $Y_0(0) = \Phi(x_0)$ and Y_0 never leaves U, then for any $t > 0, C \geq 0$, the following holds:

1. *SDE (7) has a global solution Y_C in U with $Y_C(0) = \Phi(x_0)$. Moreover, $Y_C(t) \overset{d}{=} Y_0\left(\frac{e^{4Ct}-1}{4C}\right)e^{-Ct}$;*
2. *$x_{\eta,\lambda}(\lfloor t/\eta^2 \rfloor)$ converges in distribution to $Y_C(t)$ as $\lambda, \eta \to 0$ with $\frac{\lambda}{\eta} \equiv C$.*

Remark 4.3. *Y_0 won't leave Γ (and thus U) if Γ is a closed manifold.*

Remark 4.4. *Theorem 4.2 shows that when there's no WD, the limiting diffusion is exponentially slower than that with WD in the regime of LR η and WD λ converging to zero with fixed ratio.*

5 Limiting diffusion for general case

In the previous section, we showed that the limiting diffusion exists when η and λ go to zero with a fixed ratio. However, the situation is more complicated in the general case, *i.e.*, the intrinsic LR

$\eta\lambda \to 0$ while $\frac{\eta}{\lambda}$ is upper bounded by some constant. A concrete example is $\eta \to 0$ and λ being fixed.

Since the norm has no effect on the loss value but only affects the speed, we need to consider the dynamics of the direction of the weights. Though we do not have the a closed-form expression for the norm of the iterates in general, we can approximate it by considering a special case where the trace of the noise covariance is constant. In specific, recall the canonical SDE approximation in Equation (2). Li et al. [26] proved that the dynamics of $\|X_{\eta,\lambda}(t)\|_2^2$ is

$$\mathrm{d}\|X_{\eta,\lambda}(t)\|_2^2 = -\left[2\eta\lambda\|X_{\eta,\lambda}(t)\|_2^2 + \frac{\eta^2}{\|X_{\eta,\lambda}(t)\|_2^2}\operatorname{tr}\left(\Sigma\left(\frac{X_{\eta,\lambda}(t)}{\|X_{\eta,\lambda}(t)\|_2}\right)\right)\right]\mathrm{d}t.$$

Suppose $\operatorname{tr}(\Sigma(X)) \equiv 1/\|X\|_2^2$, then $\mathrm{d}\|X_{\eta,\lambda}(t)\|_2^2 = -2\eta\lambda\|X_{\eta,\lambda}(t)\|_2^2\mathrm{d}t + \frac{\eta^2}{\|X_{\eta,\lambda}(t)\|_2^2}\mathrm{d}t$, which admits a closed-form solution:

$$\|X_{\eta,\lambda}(t)\|_2^4 = \frac{\eta}{2\lambda} + e^{-4\lambda\eta t}\left(\|X_{\eta,\lambda}(0)\|_2^4 - \frac{\eta}{2\lambda}\right). \tag{8}$$

This implies that the norm of the weights at the equilibrium is of order $(\eta/\lambda)^{1/4}$. Moreover, Equation (8) reflects the scaling of the norm of the iterates in terms of η and λ as we will see later.

5.1 Continuous-time analysis for SDE

We first consider the continuous-time case based on the SDE approximation in Equation (2) since the derivation for this case is cleaner compared to that of the discrete-time case, and the discrete-time analysis indeed mimics the continuou-time counterpart. In a nutshell, Theorem 5.1 shows that the limiting diffusion exists for SDE with some suitable non-linear time rescaling. The main result is summarized in the following theorem.

Here the key idea is that we only need to track the direction of the iterates. To do so, we need to normalize the iterates properly. However, when the trace of the noise covariance is not constant, in general it is hard to find a close-form solution for $\|X_{\eta,\lambda}(t)\|_2$. Fortunately, the norm dynamics for the special case in Equation (8) turns out to be a perfect surrogate. Specifically, we consider a scaling function $R_{\eta,\lambda}(t)$ defined as

$$R_{\eta,\lambda}(t) = \left(\frac{\eta}{2\lambda} + e^{-4\eta\lambda t}\left(1 - \frac{\eta}{2\lambda}\right)\right)^{1/4}. \tag{9}$$

Next, to rescale the time, we define $\tau_{\eta,\lambda} : [0,\infty) \to [0,\infty)$ by

$$\tau_{\eta,\lambda}(t) = \int_{s=0}^{t} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s = \frac{1}{2}\ln\left(1 + (e^{4\eta\lambda t} - 1)\frac{\eta}{2\lambda}\right) \tag{10}$$

where the second equality follows from a direct calculation. Then we have the following theorem (see Appendix C for the proof).

Theorem 5.1. *Under Assumption 3.2, 3.3 and 3.5, fix some $x_0 \in U$ and let $X_{\eta,\lambda}(0) \equiv x_0$ for all $\eta, \lambda > 0$ in the SDE (2). Let $R_{\eta,\lambda}(t)$ and $\tau_{\eta,\lambda}(t)$ be as defined in (9) and (10) respectively. If SDE (11) has a global solution Y with $Y(0) = \Phi(x_0)$ and Y never leaves U, then $X_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(T))/R_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(T))$ converges in distribution to $Y(T)$ as $\eta\lambda \to 0$ while simultaneously satisfying $\eta < 2\lambda < c$ for some constant c,*

$$\mathrm{d}Y(t) = -\frac{1}{2}\partial\Phi(Y(t))Y(t)\mathrm{d}t + \frac{1}{2}\partial^2\Phi(Y(t))[\Sigma(Y(t))]\mathrm{d}t + \partial\Phi(Y(t))\sigma(Y(t))\mathrm{d}W(t) \tag{11}$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^{\Xi}$ and $\tau_{\eta,\lambda}^{-1}(T) = \frac{\ln\left(\frac{2\lambda}{\eta}(e^{2T}-1)+1\right)}{4\eta\lambda}$.

Remark 5.2. *The additional constraint $\eta < 2\lambda$ when $\eta\lambda \to 0$ can be relaxed to $\eta < C\lambda$ for any constant $C > 0$. It suffices to note that for SDE (2), $X_{\eta,\lambda}$ with $X_{\eta,\lambda}(0) = x_0$ and $X_{C^{-1/2}\eta,C^{1/2}\lambda}$ with $X_{C^{-1/2}\eta,C^{1/2}\lambda}(0) = C^{-1/4}x_0$ have the same trajectories, up to a rescaing of $C^{-1/4}$.*

5.2 Discrete-time analysis for SGD

Now, we proceed to analyze SGD+WD by mimicking the continuous-time behavior. Specifically, we view $R_{\eta,\lambda}(k)$ as an approximation of the norm of $x_{\eta,\lambda}(k)$, and consider the rescaled version of Equation (1) denoted by $\hat{x}_{\eta,\lambda}(k) = x_{\eta,\lambda}(k)/R_{\eta,\lambda}(k)$. Further define $\widehat{X}_{\eta,\lambda}(t) = \hat{x}_{\eta,\lambda}(\lfloor t \rfloor)$, and it can be shown that (see Appendix C for the derivation)

$$\begin{aligned}\widehat{X}_{\eta,\lambda}(t) = &\int_{s=0}^{t} \left[-\eta\lambda \widehat{X}_{\eta,\lambda}(s) + (1-\eta\lambda)\left(\frac{R_{\eta,\lambda}(s)}{R_{\eta,\lambda}(s+1)} - 1\right)\widehat{X}_{\eta,\lambda}(s)\right] \mathrm{d}\lfloor s \rfloor \\ &- \int_{s=0}^{t} \frac{\eta}{R_{\eta,\lambda}(s)R_{\eta,\lambda}(s+1)} \nabla L(\widehat{X}_{\eta,\lambda}(s))\mathrm{d}\lfloor s \rfloor \\ &- \int_{s=0}^{t} \frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(s)R_{\eta,\lambda}(s+1)} \sigma(\widehat{X}_{\eta,\lambda}(s))\mathrm{d}\sum_{i=1}^{\lfloor s \rfloor}\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right). \end{aligned} \tag{12}$$

Next, we introduce the time rescaling through the $\tau_{\eta,\lambda}(\cdot)$ defined in Equation (10). Denote $\tilde{t} = \tau_{\eta,\lambda}(t)$, so $t = \tau_{\eta,\lambda}^{-1}(\tilde{t})$. Define $\widetilde{X}_{\eta,\lambda}(\tilde{t}) = \widehat{X}_{\eta,\lambda}(t)$. We can further rewrite Equation (12) as

$$\widetilde{X}_{\eta,\lambda}(\tilde{t}) = \int_{\tilde{s}=0}^{\tilde{t}} -\nabla L(\widetilde{X}_{\eta,\lambda}(\tilde{s}))\mathrm{d}A_{\eta,\lambda}(\tilde{s}) - \int_{\tilde{s}=0}^{\tilde{t}} \widetilde{X}_{\eta,\lambda}(\tilde{s})\mathrm{d}B_{\eta,\lambda}(\tilde{s}) - \int_{\tilde{s}=0}^{\tilde{t}} \sigma(\widetilde{X}(\tilde{s}))\mathrm{d}Z_{\eta,\lambda}(\tilde{s}) \tag{13}$$

where $A_{\eta,\lambda}$, $B_{\eta,\lambda}$ and $Z_{\eta,\lambda}$ are defined by

$$A_{\eta,\lambda}(\tilde{t}) = \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \frac{\eta}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}, \tag{14}$$

$$B_{\eta,\lambda}(\tilde{t}) = \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \eta\lambda - (1-\eta\lambda)\left(\frac{R_{\eta,\lambda}(i)}{R_{\eta,\lambda}(i+1)} - 1\right), \tag{15}$$

$$Z_{\eta,\lambda}(\tilde{t}) = \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right). \tag{16}$$

The convergence of $A_{\eta,\lambda}$, $B_{\eta,\lambda}$ and $Z_{\eta,\lambda}$ are summarized in the following lemma.

Lemma 5.3. *Let $A_{\eta,\lambda}$, $B_{\eta,\lambda}$ and $Z_{\eta,\lambda}$ be as defined in Equation (14), (15) and (16) respectively. Then as $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c, we have (1) $A_{\eta,\lambda}$ increases infinitely fast, (2) $B_{\eta,\lambda}(t)$ converges to $\frac{t}{2}$ in uniform metric and (3) $Z_{\eta,\lambda}$ converges in distribution to $(I_\Xi - \frac{1}{\Xi^2}\mathbb{1}\mathbb{1}^\top)^{1/2}W$ in Skohorod metric where $\{W(t)\}_{t\geq 0}$ is the Ξ-dimensional standard Brownian motion.*

Therefore, $\widetilde{X}_{\eta,\lambda}(\tilde{t})$ is also a valid Katzenberger process, then applying Theorem 3.8 yields:

Theorem 5.4. *Under Assumption 3.2, 3.3 and 3.5, fix some $x_0 \in U$ and let $x_{\eta,\lambda}(0) \equiv x_0$ for all $\eta > 0$ for SGD+WD (1). If SDE (11) has a global solution Y with $Y(0) = \Phi(x_0)$ and Y never leaves U, then for any $T > 0$, $x_{\eta,\lambda}(\lfloor \tau_{\eta,\lambda}^{-1}(T) \rfloor)/R_{\eta,\lambda}(\lfloor \tau_{\eta,\lambda}^{-1}(T) \rfloor)$ converges in distribution to $Y(T)$ as $\eta\lambda \to 0$ while simultaneously satisfying $\eta < 2\lambda < c$ for some constant c.*

6 Mixing to Equilibrium

Now we proceed to study the ergodicity of the limiting diffusion (11). Omitted proofs of this section are delayed to Appendix D. We need the following assumption.

Assumption 6.1. For any $x \in \Gamma$ with *unit norm*, $\mathrm{tr}(\Sigma(x)) = 1$.[3]

[3] Here 1 can be replaced by any positive constant and the proof still works.

First, we can characterize the norm of the limiting dynamics in the following lemma.

Lemma 6.2. *For SDE (11), it holds that* $\mathrm{d}\|Y(t)\|_2^2 = -\|Y(t)\|_2^2\mathrm{d}t + \mathrm{tr}(\Sigma(Y(t)))\mathrm{d}t$.

Note that, by Lemma 6.2, Assumption 6.1 implies that $\mathrm{d}\|Y(t)\|_2^2 = \left(-\|Y(t)\|_2^2 + \frac{1}{\|Y(t)\|_2^2}\right)\mathrm{d}t$, which is deterministic and has a closed-form solution:

$$\|Y(t)\|_2^2 = \sqrt{1 + e^{-2t}(\|Y(0)\|_2^4 - 1)}. \tag{17}$$

Therefore, regardless of the initialization, $\|Y(t)\|_2$ will always (monotonically) converge to 1 as $t \to \infty$. In this case, we only need to consider the mixing of the direction of $Y(t)$, i.e. $Y(t)/\|Y(t)\|_2$.

Lemma 6.3. *Under Assumption 6.1, for SDE (11), define* $\overline{Y}(t) = Y(t)/\|Y(t)\|_2$. *Then*

$$\mathrm{d}\overline{Y}(t) = \frac{\partial^2\Phi(\overline{Y}(t))[\Sigma(\overline{Y}(t))] - \overline{Y}(t)}{2 + 2e^{-2t}(\|Y(0)\|_2^4 - 1)}\mathrm{d}t + \frac{\partial\Phi(\overline{Y}(t))\sigma(\overline{Y}(t))}{\sqrt{1 + e^{-2t}(\|Y(0)\|_2^4 - 1)}}\mathrm{d}W(t). \tag{18}$$

To study the ergodicity of the above normalized diffusion process, we need some additional assumptions on the manifold Γ and the noise covariance.

Assumption 6.4. The loss L is C^∞. For any $r > 0$, Γ_r is a compact connected C^∞-manifold of dimension $D - M - 1$.

Assumption 6.5. For any $x \in \Gamma$, $\mathrm{rank}(\partial\Phi(x)\sigma(x)) = D - M - 1$. Note $x^\top\partial\Phi(x)\sigma(x) = 0$, the assumption is equivalent to $\mathrm{span}\{\partial\Phi(x)\sigma(x)\} = T_x(\Gamma_{|x|})$

Theorem 6.6. *Under Assumption 6.1, 6.4 and 6.5, starting from any initialization* $Y(0) \in \Gamma_1$, *the distribution of* $\overline{Y}(t)$ *converges to a unique stationary distribution* π *on* $\overline{\Gamma}$ *in total variation distance.*

Theorem 6.7 (Mixing of SGD+WD)**.** *Under the setting of Theorem 5.4 and Theorem 6.6, fix any* $x_0 \in U$ *and let* $x_{\eta,\lambda}(0) \equiv x_0$ *for all* $\eta, \lambda > 0$ *for SGD + WD in (1). Let* π *be the unique stationary distribution on* Γ *given in Theorem 6.6. Let* f *be any bounded continuous test function. Then for any* $\epsilon > 0$, *there exists some* $T > 0$ *and* $\delta > 0$ *such that when* $\eta\lambda < \delta$ *with* $\eta < 2\lambda < c$, *we have*

$$|\mathbb{E}[f(x_{\eta,\lambda}(\lfloor\tau_{\eta,\lambda}^{-1}(T)\rfloor)/R_{\eta,\lambda}(\lfloor\tau_{\eta,\lambda}^{-1}(T)\rfloor))] - \mathbb{E}_{x\sim\pi}[f(x)]| \le \epsilon.$$

Proof. Fix any test function f and any $\epsilon > 0$. By Theorem 6.6, we know that there is some $T > 0$ such that $|\mathbb{E}[f(\overline{Y}(T))] - \mathbb{E}_{x\sim\pi}[f(x)]| \le \epsilon/2$. When $\mathrm{tr}(\Sigma(x)) \equiv 1/\|x\|_2^2$, SDE (18) with $\overline{Y}(0) = \Phi(x_0)/\|\Phi(x_0)\|_2$ has a unique global solution $\overline{Y}(t)$, and thus SDE (11) has a unique global solution given by $\overline{Y}(t)\sqrt{1 + e^{-2t}(\|Y(0)\|_2^4 - 1)}$. Hence, for this T, we apply Theorem 5.4 to get

$$|\mathbb{E}[f(\bar{x}_{\eta,\lambda}(\lfloor\tau_{\eta,\lambda}^{-1}(T)\rfloor)/R_{\eta,\lambda}(\lfloor\tau_{\eta,\lambda}^{-1}(T)\rfloor))] - \mathbb{E}[f(\overline{Y}(T))]| \le \epsilon/2.$$

Combining the above yields the desired result. □

7 Experiments

In this section, we first empirically verify our theory in a simple setting, in particular, the time scaling of the dynamics. We then show that the diffusion process exists during the training of PreResNet on CIFAR-10, and it has implicit benefit towards better generalization.

7.1 Time Scaling Verification

We firstly verify the time scaling (c.f. Equation (24)) in a simple setting. We train the following normalized linear model by ℓ_2 regression: $F_x(z) = \left\langle \frac{x}{\|x\|_2}, z \right\rangle$, where $x \in \mathbb{R}^D$ is the model parameter, and z is the input. Let $\{z_i\}_{i=1}^{\Xi}$ be the input samples, and the target labels $y_i = F_{x^*}(z_i)$ for some x^*. We set $\Xi = D - 2$ so the solution space $S = \{w : \langle w - \frac{x^*}{\|x^*\|_2}, z_i\rangle = 0, \forall i \in [\Xi]\}$ is a 2-dimensional linear space. The manifold of global minimizers Γ is then equal to $\cup_{\alpha>0}\alpha\left(S \cap \{x \mid \|x\|_2 = 1\}\right)$. As long as no z_i is in the same direction with x^*, $S \cap \{x \mid \|x\|_2 = 1\}$ is a 1-dimensional manifold

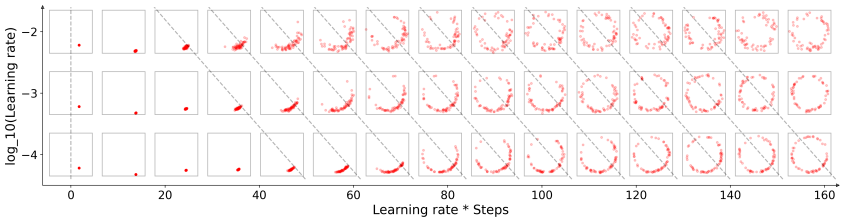


Figure 2: The dynamics of model projections on zero loss manifold with $D = 10$. Models are trained with learning rate $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and WD factor of 0.05. Each small box in the figure contains 60 models that are trained with the same learning rate and the same number of steps. The x axis indicates the product of learning rate and the number of steps of each box; the y axis indicates $\log_{10}$(learning rate) of each box. The dashed lines represent the time scaling $\frac{1}{2}\ln\left(1 + (e^{4\eta\lambda t} - 1)\frac{\eta}{2\lambda}\right) = T$ for different Ts (from 0 to 15.58). The dynamics are consistent with time scaling suggested by our theory (Theorem 5.1).

(a circle), thus satisfying Assumption 6.4. During training, we set the loss at step t be $L^{(t)}(x) = \frac{1}{\Xi}\sum_{i=1}^{\Xi}(F_x(z_i) - F_{x^*}(z_i))^2 + \left\langle \frac{x}{\|x\|_2}, \xi_t \right\rangle$, where $\xi_t \overset{iid}{\sim} N(0, \sigma^2 I_D)$ and it implies trace of noise covariance on unit sphere, thus satisfying Assumption 6.1. We can also verify $M = \text{rank}(\nabla^2 L(x)) = 8 = D - 2$ and the noise is non-degenerate, meaning Assumption 6.5 also holds.

In our experiments, we choose $D = 10$, $\sigma = 0.3$, the WD factor $\lambda = 0.05$, and learning rate $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. In Figure 2, we plot the projections of $\frac{x}{\|x\|_2}$ on the solution space for 60 different runs with identical initialization for each η. For each run, the only differences are the LR η and the noise ξ_t. The dashed lines in the figure indicates our time scaling, i.e., $\frac{1}{2}\ln\left(1 + (e^{4\eta\lambda t} - 1)\frac{\eta}{2\lambda}\right) = T$. Figure 2 shows that time scaling of $O((\ln 2\lambda/\eta + T)/(\lambda\eta))$ fits the dynamics better, compared to $O(T/(\eta\lambda))$.

7.2 Limiting diffusion on CIFAR-10

Beyond the toy example, we further observe the limit diffusion of PreResNet on CIFAR-10[20]. We train a 32-layer PreResNet[4] [10] with initial LR $\eta = 0.8$ and weight decay factor $\lambda = 5 \cdot 10^{-4}$. Unlike the normalized linear model, it is hard to visualize the model projection of PreResNet on the manifold. Instead, we choose the test accuracy of $\Phi(x_t)$ as a test function. In particular, we decay the learning rate to 10^{-3} to approximate gradient flow at different time t, and record the test accuracy after training 1000 more epochs. The results are shown in Figure 1 and Figure 3.

We observe that without LR decay, the train accuracy, test accuracy and parameter norm converge quickly after training 100 epochs; the test accuracy of $\Phi(x_t)$ converges much slower. It suggests that there exists a mixing process after the model reaches the manifold. Moreover, we observe $\Phi(x_t)$ after convergence is significantly higher than $\Phi(x_t)$ at 100th epoch. It indicates that the mixing process is beneficial for generalization. Although it is unclear whether $\Phi(x_t)$ converges at the same rate as the diffusion process, one can always use the time scaling in (24) as an upperbound to decide the proper time of LR decay.

8 Conclusion and Future Work

We provide an SDE-based characterization for the limiting dynamics of SGD+WD for a scale invariant loss as $\eta\lambda \to 0$ with $\eta = O(\lambda)$ and $\lambda = O(1)$. Under some technical assumptions, we further show that the limiting diffusion converges to a unique stationary distribution. It leaves as future work to relax the technical assumptions. Another interesting and important direction for future work is to identify the benefit on generalization induced by the limiting diffusion.

[4]We use the implementation from https://github.com/bearpaw/pytorch-classification, with slight changes (following Appendix C of [25])to ensure the model is scale-invariant.

References

[1] S. Arora, Z. Li, and K. Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[4] X. Cheng, D. Yin, P. Bartlett, and M. Jordan. Stochastic gradient and langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.

[5] Y. Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

[6] Y. Cooper. The critical locus of overparameterized neural networks. *arXiv preprint arXiv:2005.04210*, 2020.

[7] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[8] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[9] H. He, G. Huang, and Y. Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *arXiv preprint arXiv:1902.00744*, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

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

[12] E. Hoffer, R. Banner, I. Golan, and D. Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[13] K. Ichihara and H. Kunita. A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 30(3):235–254, 1974.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[15] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[16] G. S. Katzenberger. Solutions of a stochastic differential equation forced onto a manifold by a large drift. *The Annals of Probability*, pages 1587–1628, 1991.

[17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[18] W. Kliemann. Recurrence and invariant measures for degenerate diffusions. *The annals of probability*, pages 690–707, 1987.

[19] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[20] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of markov processes. *Journal of Multivariate Analysis*, 1(4):365–393, 1971.

[22] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

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

[24] Q. Li, C. 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.

[25] Z. Li and S. Arora. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*, 2019.

[26] Z. Li, K. Lyu, and S. Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

[27] Z. Li, T. Wang, and S. Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

[28] Z. Li, S. Bhojanapalli, M. Zaheer, S. J. Reddi, and S. Kumar. Robust training of neural networks using scale invariant architectures. *arXiv preprint arXiv:2202.00980*, 2022.

[29] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

[30] B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[31] S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *arXiv preprint arXiv:2106.09524*, 2021.

[32] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29: 901–909, 2016.

[33] J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics*, 8(1):933–961, 2017.

[34] J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems*, 10(2):124–151, 2020.

[35] T. Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

[36] R. Wan, Z. Zhu, X. Zhang, and J. Sun. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd. *arXiv preprint arXiv:2006.08419*, 2020.

[37] W. Whitt. Proofs of the martingale fclt. *Probability Surveys*, 4:268–302, 2007.

[38] S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part ii: Continuous time analysis. *arXiv preprint arXiv:2106.02588*, 2021.

[39] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[40] Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020.

[41] G. Zhang, C. Wang, B. Xu, and R. Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

[42] Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section X.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper studies the theoretical properties of stochastic gradient descent with weight decay.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]
(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [No]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]