Fast Mixing of Stochastic Gradient Descent with Normalization and Weight Decay

Zhiyuan Li
Princeton University
zhiyuanli@cs.princeton.edu

Tianhao Wang
Yale University
tianhao.wang@yale.edu

Dingli Yu
Princeton University
dingliy@cs.princeton.edu

Abstract

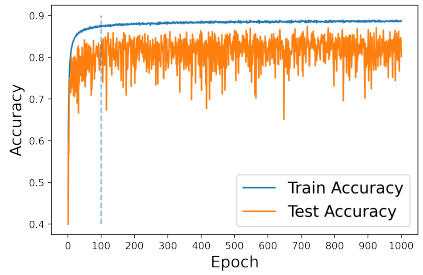
We prove the Fast Equilibrium Conjecture proposed by Li et al. [30], *i.e.*, stochastic gradient descent (SGD) on a scale-invariant loss (*e.g.*, using networks with various normalization schemes) with learning rate η and weight decay factor λ mixes in function space in $\tilde{\mathcal{O}}(\frac{1}{\lambda\eta})$ steps, under two standard assumptions: (1) the noise covariance matrix is non-degenerate and (2) the minimizers of the loss form a connected, compact and analytic manifold. The analysis uses the framework of Li et al. [31] 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 satisfying $\eta \leq O(\lambda) \leq O(1)$. Moreover, the evolution of the limiting distribution can be described by a stochastic differential equation which mixes to the same equilibrium distribution for every initialization around the manifold of minimizers as $T \to \infty$.

1 Introduction

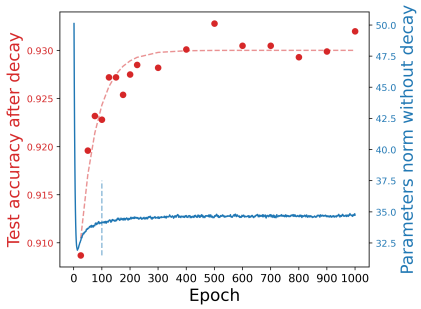
Normalization methods such as Batch Normalization [19], Group Normalization [46], Layer Normalization [4] and Weight Normalization [37] have facilitated the efficient training of modern deep neural networks, and in the extensive tool case stochastic gradient gradient (SGD) with weight decay (WD) is one 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 [29].

There is a line of research [5, 22] suggesting that small LRs (or equivalently large batch sizes) might hurt generalization since they lead to the so-called *sharp minima*. It is argued that noise can help avoid sharp minima [16, 22, 26, 20, 14], so large LRs and small batches are preferred as they increase the noise during training. However, later Li et al. [30] 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 effect of the gradient noise. To explain the effectiveness of small LRs, Li et al. [30] 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 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.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(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.

Before stating the Fast Equilibrium Conjecture, let us first specify the problem setting. Let $x \in \mathbb{R}^D \setminus \{0\}$ be the parameters and F_x be a neural network parameterized by 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 $C > 0$. In other words, the output of the network only depends on the direction of parameter, $\overline{x}$, which is a shorthand for $x/\|x\|_2$. 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 be $L(x) = \frac{1}{\Xi}\sum_{i=1}^{\Xi} L_i(x)$ where each L_i is the scale-invariant loss over the i-th sample.[1] We optimize L using SGD+WD, whose 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 LR, λ is the WD factor, $x_{\eta,\lambda}(k) \in \mathbb{R}^D \setminus \{0\}$ denotes the parameters after k iterations, $x_{\eta,\lambda}(0) = x_{\text{init}}$ 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 independently uniformly from $\{1, \ldots, \Xi\}$, so $\mathbb{E}_{\xi_k}[\sigma_{\xi_k}(x)] = 0$ for all $x \in \mathbb{R}^D \setminus \{0\}$ 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 (Fast Equilibrium Conjecture, Li et al. [30])**.** Suppose $X_{\eta,\lambda}(t)$ is a solution of (2), then for any input z, $F_{X_{\eta,\lambda}(t)}(z)$ converges to an equilibrium independent of the initial parameter x_{init} in $\widetilde{O}(1/(\lambda\eta))$ time.

We note that the above conjecture is implied by the convergence of distribution of the direction of the parameters, $\overline{X}_{\eta,\lambda}(t)$ for scale invariant networks.

This convergence rate is much faster than the standard $e^{-\Theta(1/\eta)}$ global mixing time of Langevin dynamics [38] and this is how the conjecture gets its name. Li et al. [30] 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. More recently, Wang and Wang [43] proposed a spherical SDE model to approximate SGD+WD with constant intrinsic LR. Using a novel adaption of Simon's theory, they justified the Fast Equilibrium Conjecture by showing that the SGD+WD dynamics consists of three stages: descent ($O(1/\sqrt{\lambda\eta})$ time), diffusion ($O(1/(\lambda\eta)$ time) and tunneling ($O(e^{C/(\lambda\eta)})$ time). Although Wang and Wang [43] provided an original and instructive characterization of SGD+WD, their analysis only applies to

[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.

Morse loss functions, *i.e.*, the local minimizers are all isolated, so the diffusion happens in a vanishing neighborhood of some isolated minimizer as the intrinsic LR $\eta\lambda \to 0$. As a result, the diffusion process shall bring no generalization benefit and cannot explain the improvement of final generalization if staying at training loss plateau for a longer time (see Figure 1). Except the undesired implications, the assumption that the loss function is Morse itself is probably too strong and is against the empirical evidence that the level sets of deep learning loss are connected [10]. Another weakness of existing analyses in Li et al. [30], Wang and Wang [43] is that their analyses only work for the SDE approximation (2), and do not apply to the discrete-time dynamics (1).

In this paper, we provide a proof to Conjecture 1.1 under more reasonable assumptions. That is, We allow the local minimizers form a connected manifold, which can be viewed a generalization of the Morse function assumption in Wang and Wang [43], as an isolated minimizer is just a manifold of dimension 0. The main result is stated informally below.

Theorem 1.2 (Informal version of Theorem 6.7)**.** *Suppose Γ is a connected analytic manifold consisting only of local minimizers of L and that $\Gamma_1 := \{x \in \Gamma \mid \|x\|_2 = 1\}$ is compact. Under some regularity assumptions, there is an open neighborhood U of Γ, such that for every $x_{init} \in U$ and $T > 0$, as $\eta\lambda \to 0$, both $\bar{x}_{\eta,\lambda}(\lfloor \frac{\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)}{4\eta\lambda} \rfloor)$ and $\overline{X}_{\eta,\lambda}(\frac{\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)}{4\eta\lambda})$ converge in distribution to the same distribution, denoted by $\mu_{T,x_{init}}$. Moreover, for every $x_{init} \in U$, $\mu_{T,x_{init}}$ weakly converges to the same equilibrium distribution π supported on Γ_1 as $T \to \infty$.*

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 $\lambda = O(1)$. This result is not a direct application of the framework proposed by Li et al. [31] as there is no minimizer in the usual sense, not to mention manifold of local minimizers. The key proof technique is to introduce a novel time-rescaling to make the convergence result in [21] applicable.
2. We show that SGD without WD for stochastic scale-invariant loss has the same limiting dynamics as SGD+WD, but is exponentially slower (see Theorem 4.2). This is consistent to the empirical observation [48, 30] 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 of SGD+WD 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 [30].
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 [2]. It has been shown that scale-invariance enables robust and efficient training of SGD with WD [32]. Scale invariance also brings about the interesting equivalence between the effect of WD and LR schedules studied in Hoffer et al. [17], Zhang et al. [48], Li and Arora [29]. Moreover, for SGD with WD, η being the LR and λ being the WD factor, the norm of the parameters will converge to $(\lambda/\eta)^{1/4}$ [30, 41, 42], and this induces the intrinsic LR which is equal to $\eta\lambda$ [30]. 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 [33, 27, 39, 28, 6, 40]. Many interesting properties of SGD have been discovered through this approach [24, 13, 49, 30, 47, 35, 45]. Recently, Li et al. [31] proposed a framework to study the implicit bias of SGD by showing that with a proper time-rescaling, 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

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

Recall that the training loss $L : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}$ is defined as $L = \frac{1}{\Xi}\sum_{i=1}^{\Xi} L_i$ where L_i is the loss on the i-th sample. With vast overparametrization in modern machine learning models, multiple minimizers can exist and form a manifold [8, 7]. Thus following Fehrman et al. [12], Li et al. [31], Arora et al. [3], we make the assumption below throughout this paper.

Assumption 3.1. Each loss function $L_i : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}$ is a $\mathcal{C}^4$ function. Γ is a $(D - M)$-dimensional $\mathcal{C}^2$-submanifold of $\mathbb{R}^D$ for some integer $0 \le M \le D - 1$, where each $x \in \Gamma$ is a local minimizer of L and $\mathrm{rank}(\nabla^2 L(x)) = M$ for all $x \in \Gamma$.

Note that $\nabla^2 L(x)$ must have zero eigenvalues in the tangent space of Γ at x, we are indeed assuming Hessian attains the maximal rank everywhere on mainfold.

In this paper, we are interested in the behavior of SGD+WD with each L_i being scale invariant, or equivalently, 0-homogenous. We note that the level sets of scale invariant functions are always cones, which will be used frequently in our analysis. To this end, we make Assumption 3.3.

Definition 3.2 (Homogeneous Functions)**.** We say a function $f : \mathbb{R}^D \setminus \{0\} \to \mathbb{R}^m$ is a k-*homogeneous* for some $k \in \mathbb{R}$ if and only if for all $x \in \mathbb{R}^D \setminus \{0\}$ and $\alpha > 0$, $f(\alpha x) = \alpha^k f(x)$. Specifically, we say a function f is *scale invariant* if and only if it is 0-homogeneous.

Assumption 3.3. L_i is scale invariant for each $1 \le i \le \Xi$ and Γ is a cone.

Recall that $\sigma_i(x) = \frac{1}{\sqrt{\Xi}}(\nabla L_i(x) - \nabla L(x))$, so the noise function σ is (-1)-homogeneous and thus the noise covariance $\Sigma(x) = \sigma(x)\sigma(x)^\top$ is (-2)-homogeneous. Below are two useful properties of homogeneous functions, whose proofs follow from applying direct chain rules directly.

Lemma 3.4. *For any $l \in \mathbb{N}$ and k-homogeneous function f, $\nabla^l f$ is $(k - l)$-homogeneous.*

Lemma 3.5 (Euler's Theorem for Homogeneous Functions)**.** *For any real-valued k-homogeneous function f, $\langle x, \nabla f(x)\rangle = kf(x)$.*

Next, the following limiting map of gradient flow plays a key role in our analysis.

Definition 3.6. For any $x \neq 0$, we define the *gradient flow* governed by $-\nabla L$ as the unique solution of $\phi(x, t) = x - \int_0^t \nabla L(\phi(x, s))\mathrm{d}s$ for $t \ge 0$ and its associated limiting map by $\Phi(x) = \lim_{t\to\infty} \phi(x, t)$ whenever the limit exists.

Throughout the paper, we use U to denote the attraction set of Γ under gradient flow, that is, $U = \{x \in \mathbb{R}^D \mid \Phi(x) \text{ is well-defined and } \Phi(x) \in \Gamma\}$.

Lemma 3.7 (Lemma B.15 of Arora et al. [3])**.** *U is open and Φ is $\mathcal{C}^2$ on U.*

Lemma 3.8. *Under Assumption 3.3, U is a cone and Φ is 1-homogeneous in U.*

Li et al. [31] 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 3.9 (Lemma 4.2, [31])**.** *For any $x \in \Gamma$, $\partial\Phi(x) \in \mathbb{R}^{D\times D}$ is the orthogonal projection matrix onto the tangent space $T_x(\Gamma)$. As a consequence, for any $x \in \Gamma$, Assumption 3.3 $\implies$ $x \in T_x(\Gamma) \implies \Phi(x)x = x$.*

3.1 Limiting diffusion on the manifold of local minimizers

We recap the notion of *Katzenberger processes* proposed in Li et al. [31] and the characterization of the corresponding limiting diffusion based on Katzenberger's theorems in Katzenberger [21]. In this subsection we only assumeAssumption 3.1, but not Assumption 3.3.

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 $\{Z_n(t)\}_{t\geq 0}$ be a $\mathbb{R}^{\Xi}$-valued stochastic process. In our context of SGD+WD, given loss function $L : \mathbb{R}^D \to \mathbb{R}$, noise function $\sigma : \mathbb{R}^D \to \mathbb{R}^{D\times\Xi}$ and initialization $x_{\text{init}} \in U$, we call the following stochastic process (3) a *Katzenberger process* if as $n \to \infty$,

1. A_n increases infinitely fast, i.e., $\forall \epsilon > 0, \inf_{t\geq 0}(A_n(t+\epsilon) - A_n(t)) \to \infty$;
2. $B_n(t)$ converges to $c \cdot t$ in uniform metric for some constant c.
3. Z_n converges in distribution to $(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)W$ in Skorokhod metric where W is a Ξ-dimension standard Brownian motion (See Definition of Skorokhod metric in Appendix B);

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

Theorem 3.10 (Theorem 4.6, Li et al. [31]). *Given a Katzenberger process $\{X_n(\cdot)\}_{n\in\mathbb{N}}$, if SDE (4) has a global solution Y in U with $Y(0) = \Phi(x_{init})$, then for any $t > 0$, $X_n(t)$ converges in distribution to $Y(t)$ as $n \to \infty$.*

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

We note that the global solution always exists if the manifold Γ is compact. Our notion of Katzenberger process and theorem statement is slightly more general than those in Li et al. [31] to handle the weight decay. However, our formulation is still under the original framework of Katzenberger [21] and the proof in Li et al. [31] can be easily adapted to Theorem 3.10.

4 Warm-up: Simultaneous Limit Case

As a warm-up, we first consider the setting where $\eta, \lambda \to 0$ simultaneously with $\frac{\lambda}{\eta} \equiv C$ for some constant $C \geq 0$. In this special regime, we do not need to use the scale invariance property of the loss, and we can directly apply Theorem 3.10 to obtain the limiting diffusion of SGD+WD. Nonetheless, we will see the benefit of weight decay as a source of acceleration. While for the general case of $\eta\lambda \to 0$ that will be considered in Section 5, we need to carefully design a time rescaling by calibrating with the dynamics of parameter magnitude, so that under the new scaling the dynamics can still be understood as a Katzenberger process.

Now, recall the SGD+WD updates in Equation (1), and let us fix $x_{\eta,\lambda}(0) = x_{\text{init}}$ for some $x_{\text{init}} \in U$. Define $\check{X}_{\eta,\lambda}(t) = x_{\eta,\lambda}(\lfloor t/\eta^2 \rfloor)$, which is roughly equivalent to SDE (2) with $1/\eta^2$ times acceleration, and we can rewrite the discrete-time update of $x_{\eta,\lambda}$ as

$$\check{X}_{\eta,\lambda}(t) = x_{\text{init}} + \int_0^t \sigma(\check{X}_{\eta,\lambda}(s))\mathrm{d}Z_{\eta,\lambda}(s) + \int_0^t \check{X}_{\eta,\lambda}(s)\mathrm{d}B_{\eta,\lambda}(s) - \int_0^t \nabla L(\check{X}_{\eta,\lambda}(s))\mathrm{d}A_{\eta,\lambda}(s) \tag{5}$$

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

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

Note that $A_{\eta,\lambda}(t)$ is roughly t/η which becomes very large for small η, thus the negative gradient part will drive $\check{X}_{\eta,\lambda}(t)$ rapidly towards the manifold Γ and force $\check{X}_{\eta,\lambda}(t)$ to stay close to Γ after that. On the other hand, as $\eta \to 0$, $B_{\eta,\lambda}(t)$ will converge to Ct 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.

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 $\frac{\lambda}{\eta} \equiv C$, it holds that (1) $A_{\eta,\lambda}$ increases infinitely fast, (2) $B_{\eta,\lambda}(t)$ converges to Ct in uniform metric and (3) $Z_{\eta,\lambda}$ converges in distribution to $(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)^{1/2}W$ in Skorokhod metric where $\{W(t)\}_{t\geq 0}$ is the Ξ-dimensional standard Brownian motion.*

Therefore, a direct application of Theorem 3.10 yields the limiting diffusion in this case.

Theorem 4.2. *Under Assumption 3.1, let $x_{\eta,\lambda}(0) \equiv x_{init} \in U$, $\forall \eta, \lambda > 0$ in SGD+WD (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_C in U for some $C \geq 0$ with $Y_C(0) = \Phi(x_{init})$, then the following holds:

1. *$x_{\eta,\lambda}(\lfloor t/\eta^2 \rfloor)$ converges in distribution to $Y_C(t)$ as $\lambda, \eta \to 0$ with $\frac{\lambda}{\eta} \equiv C$;*
2. *Under Assumption 3.3, for every $C' \geq 0$ SDE (7) with any $C' \geq 0$ has a global solution $Y_{C'}$ in U with $Y_{C'}(0) = \Phi(x_{init})$. Moreover, $Y_{C'}(t) \overset{d}{=} Y_0\left(\frac{e^{4C't}-1}{4C'}\right)e^{-C't}$.*

Remark 4.3. *Theorem 4.2 shows that when there is no WD, the limiting diffusion is still the same as that with WD, but exponentially slower than that with WD in the regime of LR η and WD λ going to zero with a fixed ratio.*

5 Limiting Diffusion for General Cases

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 where we drop the assumption of η/λ being fixed. Below we first explain the challenges in analysis and our solution for this general regime. Then we present the continuous-time analysis for SDE and the discrete-time analysis for SGD+WD. Our analysis applies for all the case when $\eta\lambda\ to 0$ with $\eta = O(\lambda) = O(1)$.

Challenges for the General Case. A concrete example for the challenge is when $\eta \to 0$ and λ be fixed as a constant, which is also the most natural and practical setting. We quickly find ourselves in a dilemma if we still want to apply Katzenberger's theorem [21], or its simplified version Theorem 3.10: If we view WD or ℓ_2 regularization as the 'fast' past of the dynamics, that is, a part of the loss function, then there is no minimizer for the ℓ_2 regularized scale-invariant loss and thus it doesn't satisfy the condition of Katzenberger's theorem; if we view WD as some 'slow' dynamics and formulate it as $\frac{\lambda}{\eta}X_{\eta,\lambda}\mathrm{d}B_{\eta,\lambda}(t)$ in Equation (5), then unlike the simultaneous limit case, $\frac{\lambda}{\eta}$ doesn't necessarily have a limit, and thus the condition of Katzenberger's theorem is again not met.

The above dilemma reflects two different roles of WD in early and late phase of training: in the early phase, when the norm is large, WD is more like a part of the loss function that executes the ℓ_2 regularization. In contrast, in the late phase of SGD training, especially when the norm square of parameters has stabilized to some value, *i.e.*, $\|x_{\eta,\lambda}(t)\|_2^2 \propto \sqrt{\frac{\eta}{\lambda}}$ (*e.g.* Figure 1b), WD should be viewed as the 'slow' dynamics and we can apply the analysis in the simultaneous limit case. This is because by the scale-invariance of loss L, Equation (1) can be rewritten as the following form, Equation (8), with $\tilde{\eta} = \sqrt{\eta\lambda}, \tilde{\lambda} = \sqrt{\eta\lambda}, \tilde{x}_{\tilde{\eta},\tilde{\lambda}} = (\frac{\lambda}{\eta})^{1/4}x_{\eta,\lambda}$.

$$\tilde{x}_{\tilde{\eta},\tilde{\lambda}}(k+1) = (1-\tilde{\eta}\tilde{\lambda})\tilde{x}_{\tilde{\eta},\tilde{\lambda}}(k) - \tilde{\eta}(\nabla L(x_{\tilde{\eta},\tilde{\lambda}}(k)) + \sqrt{\Xi}\sigma_{\tilde{\xi}_k}(x_{\tilde{\eta},\tilde{\lambda}}(k))) \tag{8}$$

With such a rescaling, we successfully make the norm of parameters in constant scale, that is, $\|\tilde{x}_{\tilde{\eta},\tilde{\lambda}}\|_2^2 = \sqrt{\frac{\lambda}{\eta}}\|x_{\eta,\lambda}\|_2^2 = \Theta(1)$ and thus we can apply Katzenberger's theorem. Note that we cannot do this in the early phase of SGD+WD as we start from a fixed initialization and such a rescaling will change the magnitude of the initilaization.

Our Strategy for Analysis . To overcome the above dilemma, our core strategy is to introduce a novel combination of parameter rescaling $R_{\eta,\lambda}$ (Equation (10)) and time rescaling $\tau_{\eta,\lambda}$ (Equation (11)) which interpolate the early and late regime smoothly. Because the rescalings are adaptive

to the norm of the parameter along the training trajectory, they allow us to apply Katzenberger's theorem on the rescaled dynamics $\frac{X_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(t))}{R_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(t))}$. (See formal statements in Theorem 5.1). Compared with the ordinary SGD without WD studied in Li et al. [31] where the time rescaling is set to be a fixed acceleration by $\frac{1}{\eta^2}$ times, here the design of the time rescaling is more complicated.

Since the norm has no effect on the loss value but only affects the speed, we need to consider the dynamics of the parameter direction. 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 $\|X_{\eta,\lambda}(t)\|_2$, but we can approximate it using the former special case. In specific, recall the canonical SDE approximation in (2). Li et al. [30] 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 the above further simplifies into $\mathrm{d}\|X_{\eta,\lambda}(t)\|_2^2 = -2\eta\lambda\|X_{\eta,\lambda}(t)\|_2^2\mathrm{d}t + \frac{\eta^2}{\|X_{\eta,\lambda}(t)\|_2^2}\mathrm{d}t$, which admits a closed-form solution:

$$\|X_{\eta,\lambda}(t)\|_2^4 = \frac{\eta}{2\lambda} + e^{-4\lambda\eta t}\left(\|X_{\eta,\lambda}(0)\|_2^4 - \frac{\eta}{2\lambda}\right). \tag{9}$$

This implies that the norm of the weights at the equilibrium is of order $(\eta/\lambda)^{1/4}$. Moreover, Equation (9) 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, *i.e.*, SDE approximation (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. The main result is summarized in Theorem 5.1, which shows that the limiting diffusion exists for SDE with some suitable non-linear time rescaling.

As mentioned in the previous discussion, we consider a scaling function $R_{\eta,\lambda}(t)$ inspired by the norm dynamics for the special case in Equation (9):

$$R_{\eta,\lambda}(t) = \left(\frac{\eta}{2\lambda} + e^{-4\eta\lambda t}\left(1 - \frac{\eta}{2\lambda}\right)\right)^{1/4}. \tag{10}$$

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

$$\tau_{\eta,\lambda}(t) = \int_{s=0}^{t} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s = \frac{1}{2}\ln\left(1 + (e^{4\eta\lambda t} - 1)\frac{\eta}{2\lambda}\right) \tag{11}$$

where the second equality follows from a direct calculation. It is easy to show that $\tau_{\eta,\lambda}^{-1}(t) = \frac{\ln\left(\frac{2\lambda}{\eta}(e^{2T}-1)+1\right)}{4\eta\lambda}$. Then we have the following theorem (see Appendix D for the proof), which says the rescaled version of SDE approximation Equation (2) admits the following limiting diffusion Equation (12), where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^{\Xi}$.

$$\mathrm{d}Y(t) = -\frac{1}{2}Y(t)\mathrm{d}t + \frac{1}{2}\partial^2\Phi(Y(t))[\Sigma(Y(t))]\mathrm{d}t + \partial\Phi(Y(t))\sigma(Y(t))\mathrm{d}W(t) \tag{12}$$

Theorem 5.1. *Under Assumption 3.1 and 3.3, let $X_{\eta,\lambda}(0) \equiv x_{init} \in U$ for all $\eta,\lambda > 0$ in SDE (2). Let $R_{\eta,\lambda}(t)$ and $\tau_{\eta,\lambda}(t)$ be defined in (10) and (11). If SDE (12) has a global solution Y with $Y(0) = \Phi(x_{init})$ in 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 $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c.*

Remark 5.2. *The additional constraint $\eta < 2\lambda$ when $\eta\lambda \to 0$ can be relaxed to $\eta < 2C\lambda$ for any constant $C > 0$. It suffices to note that for SDE (2), $X_{\eta,\lambda}$ with $X_{\eta,\lambda}(0) = x_{init}$ and $X_{C^{-1/2}\eta,C^{1/2}\lambda}$ with $X_{C^{-1/2}\eta,C^{1/2}\lambda}(0) = C^{-1/4}x_{init}$ have the same trajectories, up to a rescaing of $C^{-1/4}$. This is equivalent to replace $\frac{\eta}{2\lambda}$ by $\frac{\eta}{2C\lambda}$ in (10).*

5.2 Discrete-time Analysis for SGD+WD

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 D for the derivation)

$$\begin{aligned}\widehat{X}_{\eta,\lambda}(t) = &\int_{s=0}^{t} \left[-\eta\lambda \widehat{X}_{\eta,\lambda}(s) + (1-\eta\lambda)\left(\frac{R_{\eta,\lambda}(s)}{R_{\eta,\lambda}(s+1)} - 1\right)\widehat{X}_{\eta,\lambda}(s)\right] \mathrm{d}\lfloor s \rfloor \\ &- \int_{s=0}^{t} \frac{\eta}{R_{\eta,\lambda}(s)R_{\eta,\lambda}(s+1)} \nabla L(\widehat{X}_{\eta,\lambda}(s))\mathrm{d}\lfloor s \rfloor \\ &- \int_{s=0}^{t} \frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(s)R_{\eta,\lambda}(s+1)} \sigma(\widehat{X}_{\eta,\lambda}(s))\mathrm{d}\sum_{i=1}^{\lfloor s \rfloor}\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right). \end{aligned} \tag{13}$$

Next, we introduce the time rescaling through the $\tau_{\eta,\lambda}(\cdot)$ defined in Equation (11). 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 (13) as

$$\widetilde{X}_{\eta,\lambda}(\tilde{t}) = \int_{\tilde{s}=0}^{\tilde{t}} -\nabla L(\widetilde{X}_{\eta,\lambda}(\tilde{s}))\mathrm{d}A_{\eta,\lambda}(\tilde{s}) - \int_{\tilde{s}=0}^{\tilde{t}} \widetilde{X}_{\eta,\lambda}(\tilde{s})\mathrm{d}B_{\eta,\lambda}(\tilde{s}) - \int_{\tilde{s}=0}^{\tilde{t}} \sigma(\widetilde{X}(\tilde{s}))\mathrm{d}Z_{\eta,\lambda}(\tilde{s}) \tag{14}$$

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

$$A_{\eta,\lambda}(\tilde{t}) = \sum\nolimits_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \frac{\eta}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}, \tag{15}$$

$$B_{\eta,\lambda}(\tilde{t}) = \sum\nolimits_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \eta\lambda - (1-\eta\lambda)\left(\frac{R_{\eta,\lambda}(i)}{R_{\eta,\lambda}(i+1)} - 1\right), \tag{16}$$

$$Z_{\eta,\lambda}(\tilde{t}) = \sum\nolimits_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}) \rfloor} \frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right). \tag{17}$$

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 (15), (16) and (17) respectively. Then as $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c, the following holds:*

- *$A_{\eta,\lambda}$ increases infinitely fast;*
- *$B_{\eta,\lambda}(t)$ converges to $\frac{t}{2}$ in uniform metric;*
- *$Z_{\eta,\lambda}$ converges in distribution to $(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)^{1/2}W$ in Skorokhod metric where $\{W(t)\}_{t\geq 0}$ is the Ξ-dimensional standard Brownian motion.*

Therefore, $\widetilde{X}_{\eta,\lambda}(\tilde{t})$ is also a valid Katzenberger process, then applying Theorem 3.10 yields:

Theorem 5.4. *Under Assumption 3.1 and 3.3, let $x_{\eta,\lambda}(0) \equiv x_{init} \in U$ for all $\eta, \lambda > 0$ in SGD+WD (1). Let $R_{\eta,\lambda}(t)$ and $\tau_{\eta,\lambda}(t)$ be defined in (10) and (11). If SDE (12) has a global solution Y in U with $Y(0) = \Phi(x_{init})$, 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 $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ some constant c.*

6 Mixing to Equilibrium

Now we proceed to study the ergodicity of the limiting diffusion (12). Omitted proofs of this section are delayed to Appendix E.First, we can characterize the norm of the limiting dynamics in the following lemma.

Lemma 6.1. *For SDE (12), it holds that $\mathrm{d}\|Y(t)\|_2^2 = -\|Y(t)\|_2^2\mathrm{d}t + \mathrm{tr}(\Sigma(Y(t)))\mathrm{d}t$.*

To study the ergodicity of the above normalized diffusion process, we need some additional assumptions on Γ and the noise covariance. For convenience, we define $\Gamma_r := \Gamma \cap \{x \in \mathbb{R}^D : \|x\|_2 = r\}$.

To ensure the existence of stationary distribution of the limiting diffusion process, we assume Γ_1 is compact. We also need to assume that Γ_1 are connected (so is Γ) for the uniqueness of the equilibrium distribution.

Assumption 6.2. Γ_1 is compact and connected.

Assumption 6.3 (Controllability)**.** For each $x \in \Gamma$, $\text{span}(\{\partial\Phi(x)\sigma_i(x)\}_{i=1}^{\Xi}) = T_x(\Gamma_{\|x\|_2})$.

Definition 6.4. Let $V \subseteq \mathbb{R}^D$ be any open set, a function $f : V \to \mathbb{R}$ is *analytic* if for each $x \in V$, there is a neighborhood V_x of x such that the Taylor series of f expanded at x converges to f everywhere in V_x. A vector-valued function is analytic if each of its coordinate functions is analytic.

Assumption 6.5. $\text{tr}(\Sigma(\cdot))$ is an analytic function on $\mathbb{R}^D \setminus \{0\}$ and Γ is an analytic manifold.

Now we are ready to state our main result in this section, which is Theorem 6.6. It is proved in two cases respectively in Appendices E.2 and E.3, depending on whether the trace of gradient covariance $\text{tr}[\Sigma]$ is constant on Γ_1. If it is, then the diffusion process essentially is on a $(D-M-1)$-dimensional manifold (after suitable rescaling), Γ_1, just as the analysis by Wang and Wang [43]. If not, then the situation becomes more complicated and the diffusion process is on Γ, a $(D-M)$ dimensional manifold. In this case we will need the analyticity assumption, Assumption 6.5.

Theorem 6.6. *Under Assumption 3.1, 3.3, 6.2, 6.3 and 6.5, starting from any initialization $Y(0) \in \Gamma$, the distribution of $\overline{Y}(t)$ converges to a unique stationary distribution π on Γ_1 in total variation distance.*

Our main result on the fast mixing of SGD+WD and its SDE approximation follows from a direct combination between the convergence of the SGD+WD iterates proved in Theorem 5.4 and Theorem 6.6. Here we note that since Γ_1 is a compact domain, convergence in total variation implies convergence in distribution.

Theorem 6.7 (Fast Mixing)**.** *Under Assumption 3.1, 3.3, 6.2, 6.3 and 6.5, let $x_{\eta,\lambda}(0) \equiv X_{\eta,\lambda}(0) \equiv x_{init} \in U$ for all $\eta, \lambda > 0$ for SGD + WD (1) and SDE approximation (2). For any $T > 0$, as $\eta\lambda \to 0$, both $\bar{x}_{\eta,\lambda}(\lfloor \frac{\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)}{4\eta\lambda} \rfloor)$ and $\overline{X}_{\eta,\lambda}(\frac{\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)}{4\eta\lambda})$ converge in distribution to the same distribution, denoted by $\mu_{T,x_{init}}$. Moreover, for every $x_{init} \in U$, $\mu_{T,x_{init}}$ weakly converges to the same equilibrium distribution π supported on Γ_1 as $T \to \infty$.*

7 Experiments

In this section, we first empirically verify our theory in a simple setting for, 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 (11)) in a simple setting. We train the following normalized linear model by ℓ_2 regression: $F_x(z) = \langle \frac{x}{\|x\|_2}, z \rangle$, 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 given by $\Gamma = \bigcup_{\alpha>0} \alpha\,(S \cap \{x \mid \|x\|_2 = 1\})$. As long as x^* is not contained in the linear span of $\{z_i\}_{i=1}^{\Xi}$, $S \cap \{x \mid \|x\|_2 = 1\}$ is a 1-dimensional manifold (a circle), thus satisfying Assumption 6.2. During training, we set the loss at step t to 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 that the trace of noise covariance is constant on the unit sphere, thus satisfying Assumption E.2. We can also verify $M = \text{rank}(\nabla^2 L(x)) = 8 = D - 2$ and the noise is non-degenerate, meaning that Assumption 6.3 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

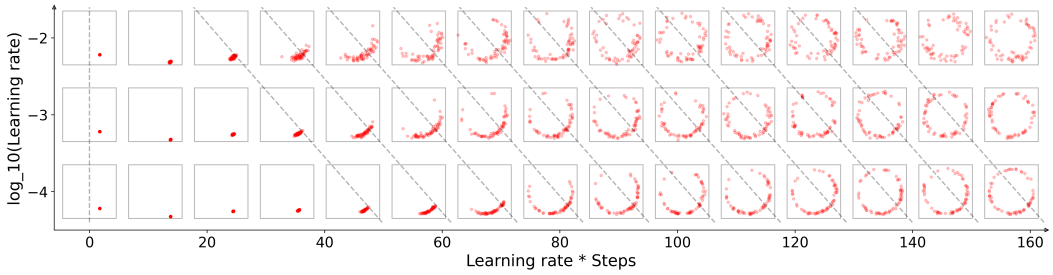


Figure 2: The dynamics of model projections on zero loss manifold with $D = 10$. Models are trained with learning rate $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and WD factor of 0.05. Each small box in the figure contains 60 models that are trained with the same learning rate and the same number of steps. The x axis indicates the product of learning rate and the number of steps of each box; the y axis indicates $\log_{10}(\text{learning rate})$ 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 T's (from 0 to 15.58). The dynamics are consistent with time scaling suggested by our theory (Theorem 5.1).

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 study the limiting diffusion of PreResNet on CIFAR-10 [25]. We train a 32-layer PreResNet[2] [15] 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 the test accuracy of $\Phi(x_t)$ after convergence is significantly higher than $\Phi(x_t)$ at 100th epoch. It indicates that the mixing process is beneficial for generalization.

Our time scaling (11) suggests that the optimal step of LR decay grows no faster than $\tilde{\Omega}(1/\eta\lambda)$ as $\eta \to 0$. Unfortunately, (11) alone is not sufficient for deciding the optimal step for decaying LR as the mixing time T for the continuous dynamics is unknown. A potential usage of time scaling (11) is to first estimate T via another SGD run with a larger learning rate, which we leave for future work.

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.

References

[1] Andrei Agrachev, Davide Barilari, and Ugo Boscain. *A comprehensive introduction to sub-Riemannian geometry*, volume 181. Cambridge University Press, 2019.

[2] Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

[3] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba

[2] We use the implementation from https://github.com/bearpaw/pytorch-classification, with slight changes (following Appendix C of [29]) to ensure the model is scale-invariant.

Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

[6] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.

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

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

[9] Giuseppe Da Prato and Jerzy Zabczyk. *Ergodicity for infinite dimensional systems*, volume 229. Cambridge university press, 1996.

[10] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

[11] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[12] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.

[13] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[14] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *arXiv preprint arXiv:1902.00744*, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[17] Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[18] Kanji Ichihara and Hiroshi 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.

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

[20] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

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

[22] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

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

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

[26] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[27] 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.

[28] 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.

[29] Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2019.

[30] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

[31] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

[32] Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *International Conference on Machine Learning*, pages 12656–12684. PMLR, 2022.

[33] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

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

[35] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *arXiv preprint arXiv:2106.09524*, 2021.

[36] David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

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

[38] Bin Shi, Weijie J Su, and Michael I Jordan. On learning rates and schr\" odinger operators. *arXiv preprint arXiv:2004.06977*, 2020.

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

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

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

[42] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd. *arXiv preprint arXiv:2006.08419*, 2020.

[43] Yi Wang and Zhiren Wang. Three-stage evolution and fast equilibrium for sgd with non-degerate critical points. In *International Conference on Machine Learning*, pages 23092–23113. PMLR, 2022.

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

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

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

[47] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020.

[48] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

[49] 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. *arXiv preprint arXiv:1803.00195*, 2018.

Checklist

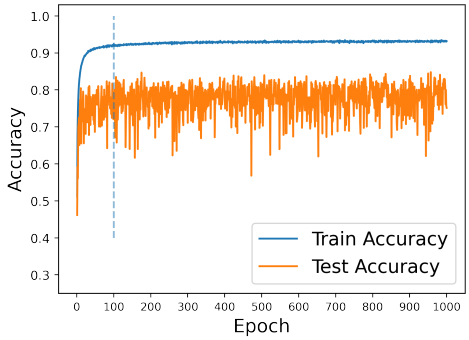
The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing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

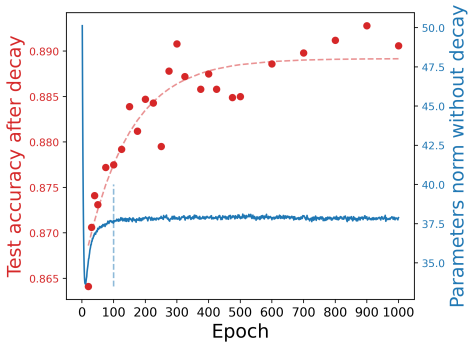
Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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)? [Yes]
   (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]

A Additional experimental results



(a) Train and test accuracy for CIFAR-10 training with $\eta = 0.8$, $\lambda = 0.0005$.



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

Figure 3: PreResNet trained without data augmentation on CIFAR-10. The train and test accuracy plateaus after parameter norm convergence within 100 epochs, but the generalization of SGD iterate after LR decay keeps improves. Figure 3a shows the train and test accuracy of scale invariant PreResNet trained with standard data augmentation on CIFAR-10. Each red dot in Figure 3b represents the test accuracy of model which decays LR to 10^{-3} at the corresponding epoch. The test accuracy is evaluated at full training accuracy after LR decay.

B Preliminaries on Stochastic Processes

In this section, we give a brief summary for the basic concepts related to stochastic processes and weak convergence that will be used in this paper.

B.1 Càdlàg Function and Metric

Definition B.1 (Càdlàg function)**.** Let $T \in [0, \infty]$. A function $g : [0, T) \to E$ is *càdlàg* if for all $t \in [0, T)$ it is right-continuous at t and its left limit $g(t-)$ exists. Let $\mathcal{D}_{\mathcal{E}}[0, T)$ be the set of all càdlàg function mapping $[0, T)$ into $\mathcal{E}$. We also use $\mathcal{D}_{\mathcal{E}}[0, T)$ to denote the set of all continuous function mapping $[0, T)$ into $\mathcal{E}$. By definition, $\mathcal{C}_{\mathcal{E}}[0, T) \subset \mathcal{D}_{\mathcal{E}}[0, T)$.

Definition B.2 (Skorokhod metric on $\mathcal{D}_{\mathcal{E}}[0, \infty)$)**.** For each finite $T > 0$ and each pair of functions $f, g \in \mathcal{D}_{\mathcal{E}}[0, \infty)$, define $d_T(f, g)$ as the infimum of all those values of δ for which there exist grids $0 \le t_0 < t_1 < \cdots < t_m$ and $0 < s_0 < s_1 < \cdots < \cdots < s_m$, with $t_k, s_k \ge T$, such that $|t_i - s_i| \le \delta$ for $i = 0, \ldots, k$, and

$$\|f(t) - g(s)\| \le \delta \qquad \text{if } (t, s) \in [t_i, t_{i+1}) \times [s_i, s_{i+1})$$

for $i = 0, \ldots, k-1$. The *Skorokhod metric* on $\mathcal{D}_{\mathcal{E}}[0, \infty)$ is defined to be

$$d(f, g) = \sum_{T=1}^{\infty} 2^{-T} \min\{1, d_T(f, g)\}.$$

B.2 Stochastic Processes and Stochastic Integral

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \ge 0}, \mathbb{P})$ be a filtered probability space.

Definition B.3 (Martingale)**.** Let $\{X(t)\}_{t \ge 0}$ be a $\{\mathcal{F}_t\}_{t \ge 0}$-adapted stochastic process. If for all $0 \le s \le t$, it holds that

$$\mathbb{E}[X(t) \mid \mathcal{F}_s] = X(s),$$

then X is called a martingale.

Lemma B.4 (Itô's Lemma)**.** *Let $\{X(t)\}_{t\geq 0}$ be defined through the following Itô drift-diffusion process:*

$$\mathrm{d}X(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t).$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion. Then for any twice differentiable function f, it holds that

$$\mathrm{d}f(t, X(t)) = \left(\frac{\partial f}{\partial t} + (\nabla_x f)^\top \mu_t + \frac{1}{2}\operatorname{tr}[\sigma^\top \nabla_x^2 f \sigma]\right)\mathrm{d}t + (\nabla_x f)^\top \sigma(t)\mathrm{d}W(t).$$

B.3 Weak Convergence for Stochastic Processes

Let $(\mathcal{D}_{\mathcal{E}}[0,\infty), \mathcal{A}, d)$ be a metric space equipped with a σ-algebra $\mathcal{A}$ and the Skorokhod metric defined in the previous subsection.

Let $\{X_n\}_{n\geq 0}$ be a sequence of stochastic processes on a sequence of probability spaces $\{(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)\}_{n\geq 0}$ such that each X_n has sample paths in $\mathcal{D}_{\mathcal{E}}[0,\infty)$. Also, let X be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ with sample paths on $\mathcal{D}_{\mathcal{E}}[0,\infty)$.

Definition B.5 (Weak convergence)**.** A sequence of stochastic processes $\{X_n\}_{n\geq 0}$ is said to *converge in distribution* or *weakly converge* to X (written as $X_n \Rightarrow X$) if and only if for all $\mathcal{A}$-measurable, bounded, and continuous function $f : \mathcal{D}_{\mathcal{E}}[0,\infty) \to \mathbb{R}$, it holds that

$$\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right]. \tag{18}$$

The weak convergence of random variables can be defined in a very similar way. Note though we only define weak convergence for a countable sequence of stochastic processes/ random variables, in this paper we extend the notion of convergence with real number as indices, *e.g.*, $\lim_{\eta,\lambda\to 0}$ where $\eta \leq 2\lambda \leq C$ for some constant C. Such convergence should be interpreted as $\lim_{n\to\infty}$ for any sequence of $\{\eta_n\}, \{\lambda_n\}$, such that $\eta_n\lambda_n \to 0$, with $\eta_n \leq 2\lambda_n \leq C$.

The main convergence result in Katzenberger [21] are in the sense of Skorokhod metric in Definition B.2, which is harder to understand and use compared to the more common *uniform metric* (Definition B.6). However, convergence in Skorokhod metric and uniform metric indeed coincide with each other when the limit is in $\mathcal{C}_{\mathbb{R}^D}[0,\infty)$, i.e., the continuous functions.

Definition B.6 (Uniform metric on $\mathcal{D}_{\mathcal{E}}[0,\infty)$)**.** For each finite $T > 0$ and each pair of functions $f, g \in \mathcal{D}_{\mathcal{E}}[0,T)$, the *uniform metric* is defined to be

$$d_U(f, g; T) = \sup_{t\in[0,T)} \|f(t) - g(t)\|.$$

The *uniform metric* on $\mathcal{D}_{\mathcal{E}}[0,\infty)$ is defined to be

$$d_U(f, g) = \sum_{T=1}^{\infty} 2^{-T} \min\{1, d_U(f, g; T)\}.$$

Lemma B.7 (Problem 7, Section 5, Pollard [36])**.** *If $X_n \Rightarrow X$ in the Skorokhod sense, and X has sample paths in $\mathcal{C}_{\mathbb{R}^D}[0,\infty)$, then $X_n \Rightarrow X$ in the uniform metric.*

Remark B.8. *We shall note convergence under the uniform metric defined above is weaker than convergence under the following metric $\sup_{t\in[0,\infty)} \|f(t) - g(t)\|$. Convergence in the uniform metric on $[0,\infty]$ defined in Definition B.6 is equivalent to convergence in the uniform metric on each compact set $[0,T]$ for $T \in \mathbb{N}^+$. The same holds for the Skorokhod topology.*

B.4 Omitted proofs for properties induced by scale-invariance

We first prove that the attraction set U is an open neighborhood of Γ and also a cone.

Proof of Lemma 3.7. The first claim is due to Lemma B.15 of Arora et al. [3]. Next, further assume that the loss L is scale invariant, then for any $x \in U$, we have $\Phi(x) = \lim_{t\to\infty} \phi(x,t) \in \Gamma$. Define $x(t) = \phi(x,t)$, and for any $\alpha > 0$, let $\tilde{x}(t) = \alpha \cdot x(t/\alpha^2)$, then we have $\tilde{x}(0) = \alpha x$ and $\frac{\mathrm{d}\tilde{x}(t)}{\mathrm{d}t} = -\alpha^{-1}\nabla L(x(t/\alpha^2)) = -\nabla L(\tilde{x}(t))$ where the second equality follows from the scale invariance of L. Therefore, by the definition of $\phi(x,t)$, we see that $\phi(\alpha x, t) = \tilde{x}(t) = \alpha \cdot x(t/\alpha^2) = \alpha \cdot \phi(x, t/\alpha^2)$. Let $t \to \infty$, and we get $\Phi(\alpha x) = \alpha \cdot \Phi(x)$. Since Γ itself is a cone, it follows that $\Phi(\alpha x) \in \Gamma$, and thus $\alpha x \in U$. Hence, we conclude that U is also a cone. □

C Omitted Proofs for the simultaneous limit case in Section 4

Now we present the proofs for the results in Section 4 where $\lambda, \eta \to 0$ simultaneously, that is, $\frac{\eta}{\lambda} \equiv C$ where C is some positive constant. This case can be regarded as a direct application of the framework in Li et al. [31].

Proof of Lemma 4.1. The proof can be directly adapted from that of Lemma 4.2 in Li et al. [31]. □

Proof of Theorem 4.2. The first claim is a direct consequence of Lemma 4.1 and Theorem 3.10.

For the second claim, we first show that if Y_0 exists, then for any $C > 0$, Y_C also exists and they are equal in distribution after the desired rescaling. By Lemma F.1, there is some Brownian motion $W'(t)$ such that $W(\frac{e^{4Ct}-1}{4C}) = \int_{\tau=0}^{t} e^{2Ct} \mathrm{d}W'(t)$, $\forall t \geq 0$ almost surely. For convenience, we denote $Y_0(\frac{e^{4Ct}-1}{4C})e^{-Ct}$ by $\widetilde{Y}_C(t)$. Thus we have

$$
\begin{aligned}
&\mathrm{d}\widetilde{Y}_C(t) + C\widetilde{Y}_C(t)\mathrm{d}t \\
=&e^{-Ct}\left(\frac{1}{2}\partial^2\Phi(\widetilde{Y}_C e^{Ct})[\Sigma(\widetilde{Y}_C e^{Ct})]\mathrm{d}\Big(\frac{e^{4Ct}-1}{4C}\Big) - \partial\Phi(\widetilde{Y}_C e^{Ct})\sigma(\widetilde{Y}_C e^{Ct})\mathrm{d}W\Big(\frac{e^{4Ct}-1}{4C}\Big)\right) \\
=&e^{3Ct}\frac{1}{2}\partial^2\Phi(\widetilde{Y}_C e^{Ct})[\Sigma(\widetilde{Y}_C e^{Ct})]\mathrm{d}t - e^{Ct}\partial\Phi(\widetilde{Y}_C e^{Ct})\sigma(\widetilde{Y}_C e^{Ct})\mathrm{d}W'(t) \\
=&\frac{1}{2}\partial^2\Phi(\widetilde{Y}_C)[\Sigma(\widetilde{Y}_C)]\mathrm{d}t - \partial\Phi(\widetilde{Y}_C)\sigma(\widetilde{Y}_C)\mathrm{d}W'(t),
\end{aligned}
$$

where the last equality uses the fact that $\partial^k\Phi$ is $(1-k)$-homogeneous, σ is (-1)-homogeneous and Σ is (-2)-homogeneous by Lemma 3.4. Thus we construct a global solution for (7) with hyperparameter C. Since U is a cone and $Y_0(t) \in U$ for all $t \geq 0$, it also holds that $\widetilde{Y}_C(t) \in U$.

By applying the above argument in the other direction, we can show that if Y_C exists then Y_0 exists and they are equivalent up to some rescaling. This finishes the proof. □

D Omitted proofs for general case in Section 5

Here we provide the proofs for the limiting diffusion for continuous-time SDE and discrete SGD+WD respectively, in a general case of $\eta\lambda \to 0$ where $\eta \leq 2\lambda \leq c$ and c is any constant. Note the analysis here can be easily generalized to the case $\eta\lambda \to 0$ where $\eta = \mathcal{O}(\lambda) = \mathcal{O}(1)$, that is, exists constants C_1, C_2, such that $\eta \leq C_1\lambda \leq C_2$. Such a limit is more interesting than the simultaneous limit as the it is more common to tune LR without fixing the ratio between WD and LR.

Below we will show how these rescalings work first in the continuous case in Appendix D.1, and then generalize the same proof idea to the discrete case, SGD+WD, in Appendix D.2.

D.1 Proof for the limiting diffusion of SDE

Proof of Theorem 5.1. Recall the scaling function $R_{\eta,\lambda}(t)$ defined in Equation (10):

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

which is exactly Equation (9) except $\|X(0)\|_2$ is replaced with 1. A simple calculation gives

$$
\mathrm{d}\ln R_{\eta,\lambda}(t) = -\frac{\eta\lambda e^{-4\eta\lambda t}(1-\eta/(2\lambda))}{\eta/(2\lambda) + e^{-4\eta\lambda t}(1-\eta/(2\lambda))}. \tag{19}
$$

Now, define $\widehat{X}_{\eta,\lambda}(t) = X_{\eta,\lambda}(t)/R_{\eta,\lambda}(t)$, a normalized version of $X_{\eta,\lambda}(t)$. Its dynamics is

$$\begin{aligned}
\mathrm{d}\widehat{X}_{\eta,\lambda}(t) &= -\frac{\eta}{R_{\eta,\lambda}(t)}\nabla L(X_{\eta,\lambda}(t))\mathrm{d}t - \frac{\eta}{R_{\eta,\lambda}(t)}\sigma(X_{\eta,\lambda}(t))\mathrm{d}W(t) \\
&\quad - \frac{\eta\lambda X_{\eta,\lambda}(t)}{R_{\eta,\lambda}(t)}\mathrm{d}t - \frac{X_{\eta,\lambda}(t)}{R_{\eta,\lambda}(t)}\mathrm{d}\ln R_{\eta,\lambda}(t) \\
&= -\frac{\eta}{R_{\eta,\lambda}(t)^2}\nabla L(\widehat{X}_{\eta,\lambda}(t))\mathrm{d}t - \frac{\eta}{R_{\eta,\lambda}(t)^2}\sigma(\widehat{X}_{\eta,\lambda}(t))\mathrm{d}W(t) \\
&\quad - \widehat{X}_{\eta,\lambda}(t)(\eta\lambda\mathrm{d}t + \mathrm{d}\ln R_{\eta,\lambda}(t)). \qquad (20)
\end{aligned}$$

Next, to rescale the time, recall the function $\tau_{\eta,\lambda}$ defined in Equation (11):

$$\tau_{\eta,\lambda}(t) = \int_{s=0}^{t} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s,$$

and consider $\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)) = \widehat{X}_{\eta,\lambda}(t)$. Note that $\mathrm{d}R_{\eta,\lambda}(t)^4 = -4\eta\lambda R_{\eta,\lambda}(t)^4 + 2\eta^2\mathrm{d}t$, so we have

$$\mathrm{d}\ln R_{\eta,\lambda}(t)^4 = \frac{\mathrm{d}R_{\eta,\lambda}(t)^4}{R_{\eta,\lambda}(t)^4} = -4\eta\lambda\mathrm{d}t + 2\frac{\eta^2}{R_{\eta,\lambda}(t)^4}\mathrm{d}t, \qquad (21)$$

which implies that

$$\tau_{\eta,\lambda}(t) = \frac{1}{2}\int_{s=0}^{t} 4\eta\lambda\mathrm{d}s + \mathrm{d}\ln R_{\eta,\lambda}(s)^4 = 2\eta\lambda t + 2\ln R_{\eta,\lambda}(t) = \frac{1}{2}\ln\left(1 + (e^{4\eta\lambda t} - 1)\frac{\eta}{2\lambda}\right). \qquad (22)$$

Based on the above closed-form expression of $\tau_{\eta,\lambda}(t)$, it is straightforward to verify that

$$\tau_{\eta,\lambda}^{-1}(T) = \frac{\ln\left(\frac{2\lambda}{\eta}(e^{2T} - 1) + 1\right)}{4\eta\lambda}. \qquad (23)$$

Applying Lemma F.1, we have

$$\begin{aligned}
\mathrm{d}\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)) &= -\frac{R_{\eta,\lambda}(t)^2}{\eta}\nabla L(\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)))\mathrm{d}\tau_{\eta,\lambda}(t) - \sigma(\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)))\mathrm{d}W(\tau_{\eta,\lambda}(t)) \\
&\quad - \widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t))(\eta\lambda\mathrm{d}t + \mathrm{d}\ln R_{\eta,\lambda}(t)). \qquad (24)
\end{aligned}$$

By the property of $R_{\eta,\lambda}(t)$ in Equation (19) and the definition of $\tau_{\eta,\lambda}(t)$ in Equation (11), we have

$$\eta\lambda\mathrm{d}t + \mathrm{d}\ln R_{\eta,\lambda}(t) = \frac{\eta^2/2}{\eta/(2\lambda) + e^{-4\eta\lambda t}(1 - \eta/(2\lambda))}\mathrm{d}t = \frac{\eta^2}{2R_{\eta,\lambda}(t)^4}\mathrm{d}t = \frac{1}{2}\mathrm{d}\tau_{\eta,\lambda}(t)$$

where the second equality follows from the definition of $R_{\eta,\lambda}(t)$ in Equation (10). Then combining the above identity and Equation (24) yields

$$\begin{aligned}
\mathrm{d}\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)) &= -\frac{R_{\eta,\lambda}(t)^2}{\eta}\nabla L(\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)))\mathrm{d}\tau_{\eta,\lambda}(t) - \frac{\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t))}{2}\mathrm{d}\tau_{\eta,\lambda}(t) \\
&\quad - \sigma(\widetilde{X}_{\eta,\lambda}(\tau_{\eta,\lambda}(t)))\mathrm{d}W(\tau_{\eta,\lambda}(t)).
\end{aligned}$$

Note that $\tau_{\eta,\lambda} : [0,\infty) \to [0,\infty)$ is a bijection, so we can rewrite the above in $\tilde{t} := \tau_{\eta,\lambda}(t)$ to get

$$\mathrm{d}\widetilde{X}_{\eta,\lambda}(\tilde{t}) = -\frac{R(\tau_{\eta,\lambda}^{-1}(\tilde{t}))^2}{\eta}\nabla L(\widetilde{X}_{\eta,\lambda}(\tilde{t}))\mathrm{d}\tilde{t} + \sigma(\widetilde{X}_{\eta,\lambda}(\tilde{t}))\mathrm{d}W(\tilde{t}) - \frac{\widetilde{X}_{\eta,\lambda}(\tilde{t})}{2}\mathrm{d}\tilde{t} \qquad (25)$$

where by writing $\tau_{\eta,\lambda}^{-1}(\cdot)$ we mean the functional inverse of $\tau_{\eta,\lambda}(\cdot)$ defined in Equation (11).

By assumption $\eta < 2\lambda$, can see from Equation (10) that $R_{\eta,\lambda}(t)^2$ is decreasing on $[0,\infty)$ and is bounded from below by $\sqrt{\eta/(2\lambda)}$. Therefore, for any $t \geq 0$, $R_{\eta,\lambda}(t)^2/\eta \geq 1/\sqrt{2\lambda\eta}$, so we can again apply Theorem 3.10 to get the desired limiting diffusion. This finishes the proof. □

D.2 Proof for the limiting diffusion of SGD+WD

Recall the update of SGD+WD in Equation (1) and the scaling function $R_{\eta,\lambda}$ in Equation (10), and the normalized version $x_{\eta,\lambda}(k)/R_{\eta,\lambda}(k)$ satisfies the following:

$$\begin{aligned}\frac{x_{\eta,\lambda}(k+1)}{R_{\eta,\lambda}(k+1)} &= (1-\eta\lambda)\frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k+1)} - \frac{\eta}{R_{\eta,\lambda}(k+1)}(\nabla L(x_{\eta,\lambda}(k)) + \sqrt{\Xi}\sigma_{\xi_k}(x_{\eta,\lambda}(k))) \\ &= \frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k)} - \frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k)} \cdot \frac{\eta\lambda R_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k+1)} + \left(\frac{R_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k+1)} - 1\right)\frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k)} \\ &\quad - \frac{\eta}{R_{\eta,\lambda}(k)R_{\eta,\lambda}(k+1)}\left(\nabla L\left(\frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k)}\right) + \sqrt{\Xi}\sigma_{\xi_k}\left(\frac{x_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k)}\right)\right).\end{aligned}$$

Then $\hat{x}_{\eta,\lambda}(k) = x_{\eta,\lambda}(k)/R_{\eta,\lambda}(k)$ satisfies

$$\begin{aligned}\hat{x}_{\eta,\lambda}(k+1) = \hat{x}_{\eta,\lambda}(k) &- \frac{\eta\lambda R_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k+1)}\hat{x}_{\eta,\lambda}(k) + \left(\frac{R_{\eta,\lambda}(k)}{R_{\eta,\lambda}(k+1)} - 1\right)\hat{x}_{\eta,\lambda}(k) \\ &- \frac{\eta}{R_{\eta,\lambda}(k)R_{\eta,\lambda}(k+1)}(\nabla L(\hat{x}_{\eta,\lambda}(k)) + \sqrt{\Xi}\sigma_{\xi_k}(\hat{x}_{\eta,\lambda}(k)))\end{aligned}$$

Define $\widehat{X}_{\eta,\lambda}(t) = \hat{x}_{\eta,\lambda}(\lfloor t \rfloor)$, and rewriting the above equation yields Equation (13):

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

For clarity, we break Lemma 5.3 into the following series of lemmas and prove them respectively.

Lemma D.1. *Let $\{A_{\eta,\lambda}\}_{\eta,\lambda>0}$ be defined as in Equation (15). Then $A_{\eta,\lambda}$ increases infinitely fast as $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c, that is, for any $\epsilon > 0$,*

$$\lim_{\eta\to 0}\inf_{\tilde{t}\geq 0}(A_{\eta,\lambda}(\tilde{t}+\epsilon) - A_{\eta,\lambda}(\tilde{t})) = \infty.$$

Proof of Lemma D.1. Fix any $\epsilon > 0$ and $\tilde{t} \geq 0$, and let η be sufficiently small. By definition, we have

$$\begin{aligned}A_{\eta,\lambda}(\tilde{t}+\epsilon) - A_{\eta,\lambda}(\tilde{t}) &= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor}\frac{\eta}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)} - \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)} \\ &= \sum_{i=\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor+1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor}\frac{\eta}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)} \\ &\geq \sum_{i=\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor}\frac{\eta}{R_{\eta,\lambda}(i)^2} \geq \int_{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor-1}\frac{\eta}{R_{\eta,\lambda}(s)^2}\mathrm{d}s\end{aligned}$$

where the two inequalities follow from the fact that $R_{\eta,\lambda}(t)$ is monotonically decreasing. Moreover, since $R_{\eta,\lambda}(t) \in ((\frac{\eta}{2\lambda})^{1/4}, 1]$, it holds that $\eta/R_{\eta,\lambda}(t)^2 \in [\eta, \sqrt{2\eta\lambda})$, for all $t \geq 0$. We then have

$$\begin{aligned}
A_{\eta,\lambda}(\tilde{t}+\epsilon) - A_{\eta,\lambda}(\tilde{t}) &\geq \int_{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor - 1} \frac{\eta^2}{R_{\eta,\lambda}(s)^4} \cdot \frac{R_{\eta,\lambda}(s)^2}{\eta} \mathrm{d}s \\
&\geq \frac{1}{\sqrt{2\eta\lambda}} \int_{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)\rfloor - 1} \frac{\eta^2}{R_{\eta,\lambda}(s)^4} \mathrm{d}s \\
&\geq \frac{1}{\sqrt{2\eta\lambda}} \int_{\tau_{\eta,\lambda}^{-1}(\tilde{t})}^{\tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)} \frac{\eta^2}{R_{\eta,\lambda}(s)^4} \mathrm{d}s - 3\sqrt{2\eta\lambda}.
\end{aligned}$$

Recall the definition of $\tau_{\eta,\lambda}(t)$ in Equation (11), we further have

$$A_{\eta,\lambda}(\tilde{t}+\epsilon) - A_{\eta,\lambda}(\tilde{t}) \geq \frac{\tau_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(\tilde{t}+\epsilon)) - \tau_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(\tilde{t}))}{\sqrt{2\eta\lambda}} - 3\sqrt{2\eta\lambda} = \frac{\epsilon}{\sqrt{2\eta\lambda}} - 3\sqrt{2\eta\lambda},$$

which holds for all $\tilde{t} \geq 0$ simultaneously. Hence, we obtain the desired result:

$$\lim_{\eta\to 0} \inf_{\tilde{t}\geq 0} A_{\eta,\lambda}(\tilde{t}+\epsilon) - A_{\eta,\lambda}(\tilde{t}) \geq \lim_{\eta\to 0} \frac{\epsilon}{\sqrt{2\eta\lambda}} - 3\sqrt{2\eta\lambda} = \infty.$$

□

Lemma D.2. *Let $\{B_{\eta,\lambda}\}_{\eta>0}$ be defined as in Equation (16). Then as $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c, $B_{\eta,\lambda}(\tilde{t})$ converges to $\tilde{t}/2$ in uniform metric.*

Proof of Lemma D.2. Since $\tilde{t} = \tau_{\eta,\lambda}(t) = \int_0^t \tau'_{\eta,\lambda}(s)\mathrm{d}s$, the idea is to show that the above sum can be seen as an approximation of the Riemann sum, which then yields the approximation of the integral. To see this, recall the definition of $R_{\eta,\lambda}(t)$ in Equation (10), and note that there exists some $t_i \in [i, i+1]$ such that

$$\frac{R_{\eta,\lambda}(i)}{R_{\eta,\lambda}(i+1)} - 1 = \frac{-R'_{\eta,\lambda}(t_i)}{R_{\eta,\lambda}(i+1)} = \frac{\eta\lambda e^{-4\eta\lambda t_i}(1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(i+1)R_{\eta,\lambda}(t_i)^3}. \tag{26}$$

Then since $R_{\eta,\lambda}(t)$ is positive and monotonically decreasing, it follows that

$$\begin{aligned}
B_{\eta,\lambda}(\tilde{t}) &= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \eta\lambda - (1-\eta\lambda)\frac{\eta\lambda e^{-4\eta\lambda t_i}(1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(i+1)R_{\eta,\lambda}(t_i)^3} \\
&\leq \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \eta\lambda - (1-\eta\lambda)\frac{\eta\lambda e^{-4\eta\lambda t_i}(1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(t_i)^4} \\
&= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2/2 + \eta^2\lambda^2 e^{-4\eta\lambda t_i}(1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(t_i)^4} \\
&\leq \frac{1}{2}\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} + \left(1-\frac{\eta}{2\lambda}\right)\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2\lambda^2 e^{-4\eta\lambda i}}{R_{\eta,\lambda}(i+1)^4}
\end{aligned}$$

where the last inequality again follows from the monotonicity of $R(t)$.

Note that when $i \le \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}$, we have $R_{\eta,\lambda}(i)^4 \ge \sqrt{\eta}(1-\frac{\eta}{2\lambda})$, and when $i > \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}$, we have $e^{-4\eta\lambda i} \le \sqrt{\eta}$. Thus we further have

$$
\begin{aligned}
B_{\eta,\lambda}(\tilde{t}) &\le \frac{1}{2}\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} + \left(1-\frac{\eta}{2\lambda}\right)\sum_{i=1}^{\lfloor \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}\rfloor-1}\frac{\eta^2\lambda^2 e^{-4\eta\lambda i}}{R_{\eta,\lambda}(i+1)^4} \\
&\quad + \left(1-\frac{\eta}{2\lambda}\right)\sum_{i=\lfloor \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2\lambda^2 e^{-4\eta\lambda i}}{R_{\eta,\lambda}(i+1)^4} \\
&\le \frac{1}{2}\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} + \sum_{i=1}^{\lfloor \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}\rfloor-1}\eta^{3/2}\lambda^2 e^{-4\eta\lambda i} \\
&\quad + \sqrt{\eta}\left(\lambda^2-\frac{\eta\lambda}{2}\right)\sum_{i=\lfloor \frac{1}{8\eta\lambda}\ln\frac{1}{\eta}\rfloor}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} \\
&\le \left(\frac{1}{2}+\sqrt{\eta}\lambda^2\right)\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} + \frac{1}{8}\sqrt{\eta}\lambda\ln\frac{1}{\eta}.
\end{aligned}
$$

Then since $R_{\eta,\lambda}(t)$ is monotonically decreasing, it follows that

$$
\begin{aligned}
B_{\eta,\lambda}(\tilde{t}) &\le \left(\frac{1}{2}+\sqrt{\eta}\lambda^2\right)\int_3^{\tau_{\eta,\lambda}^{-1}(\tilde{t})+2}\frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s + \frac{1}{8}\sqrt{\eta}\lambda\ln\frac{1}{\eta} \\
&\le \left(\frac{1}{2}+\sqrt{\eta}\lambda^2\right)\tau_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(\tilde{t})) + 2\eta\lambda + 4\eta^{3/2}\lambda^3 + \frac{1}{8}\sqrt{\eta\lambda}\ln\frac{1}{\eta} \\
&= \left(\frac{1}{2}+\sqrt{\eta}\lambda^2\right)\tilde{t} + 2\eta\lambda + 4\eta^{3/2}\lambda^3 + \frac{1}{8}\sqrt{\eta\lambda}\ln\frac{1}{\eta} \qquad (27)
\end{aligned}
$$

where the second inequality follows from the fact that $\eta^2/R_{\eta,\lambda}(t)^4 \in [\eta^2, 2\eta\lambda)$ for all $t \ge 0$.

On the other hand, we similarly establish a lower bound for $B_{\eta,\lambda}(\tilde{t})$. It follows from Equation (26) that

$$
\begin{aligned}
B_{\eta,\lambda}(\tilde{t}) &\ge \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\eta\lambda - (1-\eta\lambda)\frac{\eta\lambda e^{-4\eta\lambda t_i}(1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(i+1)^4} \\
&= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2/2 + [\eta\lambda e^{-4\eta\lambda(i+1)} - (1-\eta\lambda)\eta\lambda e^{-4\eta\lambda t_i}](1-\frac{\eta}{2\lambda})}{R_{\eta,\lambda}(i+1)^4} \\
&\ge \frac{1}{2}\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} + \left(1-\frac{\eta}{2\lambda}\right)\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta\lambda(e^{-4\eta\lambda}-1)e^{-4\eta\lambda i}}{R_{\eta,\lambda}(i+1)^4} \\
&\ge \frac{1}{2}\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2}{R_{\eta,\lambda}(i+1)^4} - \left(4-\frac{2\eta}{\lambda}\right)\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor}\frac{\eta^2\lambda^2 e^{-4\eta\lambda i}}{R_{\eta,\lambda}(i+1)^4}
\end{aligned}
$$

where the last inequality is due to the fact that $1-e^{-x} \le x$ for $x \ge 0$. Then applying the similar argument as we have used for the upper bound, we get

$$
B_{\eta,\lambda}(\tilde{t}) \ge \left(\frac{1}{2}-4\sqrt{\eta}\lambda^2\right)\tilde{t} - C\sqrt{\eta}\ln\frac{1}{\eta} \qquad (28)
$$

for some universal constant $C > 0$.

Now, combining the upper bound in Equation (27) and the lower bound in Equation (28), we obtain

$$
\left|B_{\eta,\lambda}(\tilde{t}) - \frac{\tilde{t}}{2}\right| \le 4\sqrt{\eta}\lambda^2\tilde{t} + C\sqrt{\eta}\ln\frac{1}{\eta}
$$

for all $\tilde{t} \geq 0$, where $C > 0$ is some universal constant. Recall the definition of the uniform metric, and we then have

$$d_U(B_{\eta,\lambda}(\tilde{t}), \tilde{t}/2) \leq \sum_{T=1}^{\infty} 2^{-T} \min\left\{1, \sup_{\tilde{t}\in[0,T)} \left|B_{\eta,\lambda}(\tilde{t}) - \frac{\tilde{t}}{2}\right|\right\}$$
$$\leq \sum_{T=1}^{\infty} 2^{-T} \min\left\{1, 4\sqrt{\eta}\lambda^2 T + C\sqrt{\eta}\ln\frac{1}{\eta}\right\}$$

where the right-hand-side converges to 0 by a standard ϵ-δ argument as $\eta \to 0$. This completes the proof. □

Lemma D.3. *Let $R_{\eta,\lambda}(t)$ be defined in Equation (10), and $\tau_{\eta,\lambda}(t)$ in Equation (11). Then it holds that*

$$\lim_{\eta\to 0, \eta\lambda\to 0} \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2 R_{\eta,\lambda}(i+1)^2} = \tilde{t}$$

uniformly for all $\tilde{t} \geq 0$.

Proof of Lemma D.3. First, since $R_{\eta,\lambda}(t)$ is positive and monotonically decreasing, we have

$$\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^4} \leq \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2 R_{\eta,\lambda}(i+1)^2} \leq \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i+1)^4}$$

which further implies that

$$\int_1^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s \leq \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2 R_{\eta,\lambda}(i+1)^2} \leq \int_2^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor+2} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s.$$

Then since $\eta^2/R_{\eta,\lambda}(t)^4 \leq \max\{2\eta\lambda, \eta^2\}$ for all $t \geq 0$, we have

$$\int_0^{\tau_{\eta,\lambda}^{-1}(\tilde{t})} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s - 2\max\{2\eta\lambda, \eta^2\} \leq \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2 R_{\eta,\lambda}(i+1)^2}$$
$$\leq \int_0^{\tau_{\eta,\lambda}^{-1}(\tilde{t})} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s + 2\max\{2\eta\lambda, \eta^2\}.$$

Recall the definition of $\tau_{\eta,\lambda}(t)$ in Equation (11), so $\int_0^{\tau_{\eta,\lambda}^{-1}(\tilde{t})} \frac{\eta^2}{R_{\eta,\lambda}(s)^4}\mathrm{d}s = \tau_{\eta,\lambda}(\tau_{\eta,\lambda}^{-1}(\tilde{t})) = \tilde{t}$. Therefore, we get

$$\left|\sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2 R_{\eta,\lambda}(i+1)^2} - \tilde{t}\right| \leq 2\max\{2\eta\lambda, \eta^2\}$$

for all $\tilde{t} \geq 0$. This implies the uniform convergence for all $\tilde{t} \geq 0$ and thus completes the proof. □

Lemma D.4. *Let $\{Z_{\eta,\lambda}\}_{\eta>0}$ be defined as in Equation (17). Then as $\eta\lambda \to 0$ with $\eta < 2\lambda < c$ for some constant c, $Z_{\eta,\lambda}$ weakly converges to $(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)^{1/2}W$ in the Skorokhod metric, where W is a Ξ-dimensional standard Brownian motion.*

Proof of Lemma D.4. For any $\eta > 0$, note that $Z_{\eta,\lambda}$ is a stochastic process with independent increments, so $Z_{\eta,\lambda}$ is indeed a Ξ-dimensional martingale. Thus we can apply the multidimensional martingale functional central limit theorem (Theorem F.2) to get the desired convergence.

To do so, we need to verify two facts: 1) the expected value of the maximum jump in $Z_{\eta,\lambda}$ is asymptotically negligible; 2) for each fixed $\tilde{t} \geq 0$, $\lim_{\eta\to 0} \mathbb{E}[Z_{\eta,\lambda}(\tilde{t})Z_{\eta,\lambda}(\tilde{t})^\top] = \tilde{t}(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)$. These two facts correspond to conditions (37) and (38) respectively.

For the first fact, note that the norm of the i-th jump in $Z_{\eta,\lambda}$ is bounded by $\frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}\|\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\|_2$. Since $R_{\eta,\lambda}(t) \geq \min\{(\frac{\eta}{2\lambda})^{1/4}, 1\}$ for all $t \geq 0$, we see that $\frac{\eta\sqrt{\Xi}}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(i+1)}\|\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\|_2 \leq (1+\sqrt{\Xi})\max\{\sqrt{2\eta\lambda}, \eta\}$, which implies that any jump in $Z_{\eta,\lambda}$ is asymptotically negligible as $\eta \to 0$ and $\eta\lambda \to 0$.

For the second fact, fixing any $\tilde{t} \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[Z_{\eta,\lambda}(\tilde{t})Z_{\eta,\lambda}(\tilde{t})^\top] &= \sum_{i,j=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2\Xi}{R_{\eta,\lambda}(i)R_{\eta,\lambda}(j)R_{\eta,\lambda}(i+1)R_{\eta,\lambda}(j+1)}\mathbb{E}\left[\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right)\left(\mathbb{1}_{\xi_j} - \frac{1}{\Xi}\mathbb{1}\right)^\top\right] \\
&= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2\Xi}{R_{\eta,\lambda}(i)^2R_{\eta,\lambda}(i+1)^2}\mathbb{E}\left[\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right)\left(\mathbb{1}_{\xi_i} - \frac{1}{\Xi}\mathbb{1}\right)^\top\right] \\
&= \sum_{i=1}^{\lfloor \tau_{\eta,\lambda}^{-1}(\tilde{t})\rfloor} \frac{\eta^2}{R_{\eta,\lambda}(i)^2R_{\eta,\lambda}(i+1)^2}\left(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top\right)
\end{aligned}$$

where the second equality follows from the independence between ξ_i and ξ_j for $i \neq j$. Now, applying Lemma D.3 yields $\lim_{\eta\to 0}\mathbb{E}[Z_{\eta,\lambda}(\tilde{t})Z_{\eta,\lambda}(\tilde{t})^\top] = \tilde{t}(I_\Xi - \frac{1}{\Xi}\mathbb{1}\mathbb{1}^\top)$.

Combining the above two facts, applying Theorem F.2 completes the proof. □

E Omitted Proofs in in Section 6

Here we provide the proof of the fast mixing of SGD+WD. By Theorem 5.4, we know that the iterates of SGD+WD track the solution to the limiting diffusion in Equation (12) at time T after roughly $\frac{1}{4\eta\lambda}\ln(\frac{2\lambda}{\eta}(e^{2T}-1)+1)$ steps. Here T is the time index for the limiting diffusion, which can be shown to mix to the equilibrium and the mixing speed is independent of η or λ since Equation (12) does not rely on these parameters. Therefore, it suffices to establish the ergodicity of the limiting diffusion, and the mixing of SGD+WD immediately follows. Then the total number of steps of SGD+WD is on the order of $O(\frac{1}{\eta\lambda}\ln\frac{\lambda}{\eta})$, yielding the fast mixing of SGD+WD.

E.1 Proof for the Norm Dynamics

We first show the dynamics of the squared ℓ_2 norm of the limiting dynamics. We need the following intermediate lemma.

Lemma E.1. *For any $x \in \Gamma$ and $A \in \mathbb{R}^{D\times D}$, it holds that $\langle x, \partial^2\Phi(x)[A]\rangle = \mathrm{tr}(A) - \mathrm{tr}(\partial\Phi(x)A\partial\Phi(x)^\top)$.*

Proof of Lemma E.1. Since $\langle x, L(x)\rangle = 0$ for all x, we have $\|\Phi(x)\|_2^2 = \|x\|_2^2$, and thus $\partial^2\|\Phi(x)\|_2^2 = 2I_D$, that is, $\partial^2\|\Phi(x)\|_2^2[A] = 2I_D[A] = 2\,\mathrm{tr}(A)$ for any $u, v \in \mathbb{R}^D$. On the other hand, we have

$$\partial_j\|\Phi(x)\|_2^2 = 2\sum_{i=1}^D \partial_j\Phi_i(x)\Phi_i(x)$$

for any $j \in [D]$, and then

$$\partial_{jk}\|\Phi(x)\|_2^2 = 2\sum_{i=1}^D \partial_{jk}\Phi_i(x)\Phi_i(x) + 2\sum_{i=1}^D \partial_j\Phi_i(x)\partial_k\Phi_i(x)$$

for any $j, k \in [D]$. Therefore, for any $A \in \mathbb{R}^{D\times D}$, we have

$$\begin{aligned}\partial^2\|\Phi(x)\|_2^2[A] &= \sum_{j,k=1}^{D} \partial_{jk}\|\Phi(x)\|_2^2 A_{jk} \\ &= 2\sum_{i,j,k=1}^{D} \partial_{jk}\Phi_i(x)\Phi_i(x)A_{jk} + 2\sum_{i,j,k=1}^{D} \partial_j\Phi_i(x)\partial_k\Phi_i(x)A_{jk} \\ &= 2\langle \Phi(x), \partial^2\Phi(x)[A]\rangle + 2\,\mathrm{tr}(\partial\Phi(x)A\partial\Phi(x)^\top),\end{aligned}$$

which further implies that

$$\mathrm{tr}(A) = \langle \Phi(x), \partial^2\Phi(x)[A]\rangle + \mathrm{tr}(\partial\Phi(x)A\partial\Phi(x)^\top).$$

Since $x \in \Gamma$, we have $\Phi(x) = x$, so rearranging the above equation yields

$$\langle x, \partial^2\Phi(x)[A]\rangle = \mathrm{tr}(A) - \mathrm{tr}(\partial\Phi(x)A\partial\Phi(x)^\top),$$

which completes the proof. □

Proof of Lemma 6.1. Consider the dynamics of $\|Y(t)\|_2^2$, and it follows from Ito's lemma that

$$\begin{aligned}\mathrm{d}\|Y(t)\|_2^2 = &-\langle Y(t), \partial\Phi(Y(t))Y(t)\rangle\mathrm{d}t + \langle Y(t), \partial^2\Phi(Y(t))[\Sigma(Y(t))]\rangle\mathrm{d}t \\ &+ 2\langle Y(t), \partial\Phi(Y(t))\sigma(Y(t))\mathrm{d}W(t)\rangle + \mathrm{tr}\left[\partial\Phi(Y(t))\Sigma(Y(t))\partial\Phi(Y(t))^\top\right]\mathrm{d}t.\end{aligned}$$

Note that since $\partial\Phi(Y(t))Y(t) = Y(t)$, we have

$$\begin{aligned}\langle Y(t), \partial\Phi(Y(t))\sigma(Y(t))\mathrm{d}W(t)\rangle &= \langle \partial\Phi(Y(t))Y(t), \sigma(Y(t))\mathrm{d}W(t)\rangle \\ &= \langle Y(t), \sigma(Y(t))\mathrm{d}W(t)\rangle = 0.\end{aligned}$$

Since $Y(t)$ stays on Γ, it follows from Lemma E.1 that $\langle Y(t), \partial^2\Phi(Y(t))[\Sigma(Y(t))] = \mathrm{tr}(\Sigma(Y(t))) - \mathrm{tr}[\partial\Phi(Y(t))\Sigma(Y(t))\partial\Phi(Y(t))^\top]$. Combining the above yields the dynamics of $\|Y(t)\|_2^2$. □

E.2 Proof for the Mixing When Trace is a Constant

In this subsection, we first prove the mixing result when the $\mathrm{tr}[\Sigma]$ is a constant on Γ_1, that is, under Assumption E.2.

Assumption E.2. For any $x \in \Gamma$ such that $\|x\|_2 = 1$, $\mathrm{tr}(\Sigma(x)) = 1$.[3]

Note that, by Lemma 6.1, Assumption E.2 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:

$$\|Y(t)\|_2^2 = \sqrt{1 + e^{-2t}(\|Y(0)\|_2^4 - 1)}. \tag{29}$$

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 E.3. *Under Assumption E.2, for SDE (12), define $\overline{Y}(t) = Y(t)/\|Y(t)\|_2$. Then*

$$\mathrm{d}\overline{Y}(t) = \frac{\partial^2\Phi(\overline{Y}(t))[\Sigma(\overline{Y}(t))] - \overline{Y}(t)}{2 + 2e^{-2t}(\|Y(0)\|_2^4 - 1)}\mathrm{d}t + \frac{\partial\Phi(\overline{Y}(t))\sigma(\overline{Y}(t))}{\sqrt{1 + e^{-2t}(\|Y(0)\|_2^4 - 1)}}\mathrm{d}W(t). \tag{30}$$

Based on Assumption E.2 and Lemma 6.1, we then can obtain the dynamics of direction of $Y(t)$, i.e., $Y(t)/\|Y(t)\|_2$.

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

Proof of Lemma E.3. By Lemma 6.1, we have

$$\begin{aligned}
\mathrm{d}\overline{Y}(t) &= \frac{1}{\|Y(t)\|_2}\mathrm{d}Y(t) - \frac{Y(t)}{2\|Y(t)\|_2^3}\mathrm{d}\|Y(t)\|_2^2 \\
&= -\frac{\partial\Phi(Y(t))Y(t)}{2\|Y(t)\|_2}\mathrm{d}t + \frac{\partial^2\Phi(Y(t))[\Sigma(Y(t))]}{2\|Y(t)\|_2}\mathrm{d}t + \frac{\partial\Phi(Y(t))\sigma(Y(t))}{\|Y(t)\|_2}\mathrm{d}W(t) \\
&\quad + \frac{Y(t)}{2\|Y(t)\|_2}\mathrm{d}t - \frac{\mathrm{tr}(\Sigma(Y(t)))Y(t)}{2\|Y(t)\|_2^3}\mathrm{d}t \\
&= \frac{\partial^2\Phi(\overline{Y}(t))[\Sigma(\overline{Y}(t))] - \overline{Y}(t)}{2\|Y(t)\|_2^4}\mathrm{d}t + \frac{\partial\Phi(\overline{Y}(t))\sigma(\overline{Y}(t))}{\|Y(t)\|_2^2}\mathrm{d}W(t) \qquad (31)
\end{aligned}$$

where the third equality follows from Lemma 3.9 and the assumption that $\mathrm{tr}(\Sigma(x)) \equiv 1/\|x\|_2^2$. Finally plugging-in the expression for $\|Y(t)\|_2^2$ in Equation (29) finishes the proof. □

Now according to Lemma E.3, $\{\overline{Y}(t)\}_{\tilde{t}\geq 0}$ is a diffusion process on Γ_1. We denote the associated class of Markov transition kernel by $\{\overline{\mathcal{P}}_t\}_{t\geq 0}$ where each $\overline{\mathcal{P}}_t : \Gamma_1 \times \mathcal{B}(\Gamma_1) \to \mathbb{R}_+$ and $\overline{\mathcal{P}}_t(x, A)$ is equal to the probability that starting from x the process is in the set A at time t.

To show the ergodicity of $\{\overline{\mathcal{P}}_t\}_{t\geq 0}$, we follow the approach introduced in Ichihara and Kunita [18], Kliemann [23] that proves ergodicity of a diffusion process by relating to its associated control system. Define $f_0(x) = \frac{1}{2}(\partial^2\Phi(x)[\Sigma(x)]-x)$ and $f_i(x) = \partial\Phi(x)\sigma_i(x)$ for each $i \in [\Xi]$. We rescale the time by defining $\tau(t) = \int_0^t \frac{1}{1+e^{-2s}(\|Y(0)\|_2^4-1)}\mathrm{d}s$. Let $\widetilde{Y}(\tau(t)) = \overline{Y}(t)$, identifying $\tilde{t} \equiv \tau(t)$, then $\widetilde{Y}$ admits $\mathrm{d}\widetilde{Y}(\tilde{t}) = f_0(\widetilde{Y}(\tilde{t}))\mathrm{d}\tilde{t} + \sum_{i=1}^{\Xi} f_i(x)\mathrm{d}W_i(\tilde{t})$. For $\widetilde{Y}$, we denote the associated class of Markov transition kernel by $\{\widetilde{\mathcal{P}}_{\tilde{t}}\}_{\tilde{t}\geq 0}$. Since there is a bijection between t and $\tilde{t}$ and $\lim_{t\to\infty}\tilde{t}/t = 1$, the mixing of $\widetilde{\mathcal{P}}_{\tilde{t}}$ implies that of $\overline{\mathcal{P}}_t$. Next, wrting the dynamics of $\widetilde{Y}$ as a Stratonovich SDE, we get

$$\mathrm{d}\widetilde{Y}(\tilde{t}) = f_0(\widetilde{Y}(\tilde{t}))\mathrm{d}\tilde{t} - \frac{1}{2}\sum_{i=1}^{\Xi}\partial f_i(\widetilde{Y}(\tilde{t}))f_i(\widetilde{Y}(\tilde{t}))\mathrm{d}\tilde{t} + \sum_{i=1}^{\Xi} f_i(\widetilde{Y}(\tilde{t}))\circ\mathrm{d}W_i(\tilde{t}).$$

Then we define its associated deterministic control system:

$$\frac{\mathrm{d}\psi(t)}{\mathrm{d}t} = f_0(\psi(t)) - \frac{1}{2}\sum_{i=1}^{\Xi}\partial f_i(\psi(t))f_i(\psi(t)) + \sum_{i=1}^{\Xi} f_i(\psi(t))u_i(t) \qquad (32)$$

where u is any piecewise continuous $\mathbb{R}^{\Xi}$-valued functions. We denote by $\psi(t, x; u)$ the solution of Equation (32) at time t with initialization x under the control u. Let $\mathcal{U}$ be the set of all piecewise continuous function from $\mathbb{R}_+$ to $\mathbb{R}^{\Xi}$. For any $x \in \Gamma$, define the positive orbit $O^+(x) = \{\psi(t, x; u) : t \geq 0, u \in \mathcal{U}\}$, that is, all points reachable from x via the control system (32). For any $S \subseteq \Gamma$, we say S is an *invariant control set* under the control system (32) if $\overline{O^+(x)} = \bar{S}$ for all $x \in S$.

The following lemma establishes the key property of the control system in Equation (32).

Lemma E.4. *Under Assumption 6.2 and 6.3, Γ_1 itself is the unique invariant control set contained in Γ_1 for the control system (32).*

Proof of Lemma E.4. It suffices to show that for any $x, y \in \Gamma_1$, $y \in \overline{O^+(x)}$. By Assumption 6.3, $\{f_i\}_{i=1}^{\Xi}$ satisfies the Lie algebra rank condition, so the following driftless control system

$$\frac{\mathrm{d}\hat{\psi}(t)}{\mathrm{d}t} = \sum_{i=1}^{\Xi} f_i(\hat{\psi}(t))u_i(t) \qquad (33)$$

is globally controllable by Rashevski-Chow theorem (see, e.g., [1]). Therefore, there exists some $T > 0$ and a control $u : \mathbb{R}_+ \to \mathbb{R}^{\Xi}$ such that $y = \hat{\psi}(T, x; u)$, where $\hat{\psi}(T, x; u)$ is the solution of Equation (33) at time T with initialization x under the control u. Next, we can use the global controllability of the driftless control system to show that $y \in \overline{O^+(x)}$ for the original control system (32).

For any $\delta > 0$, define a control $u_\delta : \mathbb{R}_+ \to \mathbb{R}^\Xi$ as $u_\delta(t) = u(t/\delta)$, then we have $y = \hat{\psi}(T, x; u) = \hat{\psi}(T\delta, x; u_\delta)$. Now using u_δ for the original control system (32), it follows that $\lim_{\delta\to 0} \psi(T\delta, x; u_\delta) = y$ as the drift is dominated by the controlled terms. Thus we conclude that $y \in \overline{O^+}(x)$, and this completes the proof. □

Now we are ready to prove Theorem 6.6.

Proof of Theorem 6.6 when trace is constant. Since $\widetilde{\mathcal{P}}_{\tilde{t}} = \overline{\mathcal{P}}_t$ and $\lim_{t\to\infty} \tilde{t}/t = 1$ by L'Hospital's rule, it suffices to show the ergodicity of $\widetilde{\mathcal{P}}_{\tilde{t}} = \overline{\mathcal{P}}_t$. Let $\mathcal{L}(f_1, \ldots, f_\Xi)$ be the Lie algebra generated by $\{f_1, \ldots, f_\Xi\}$. Assumption 6.3 implies that $\dim \mathcal{L}(f_1, \ldots, f_\Xi) = D - M - 1$, which is equal to the dimension of Γ_1. Therefore, by Lemma 5.1 in Ichihara and Kunita [18], $\widetilde{\mathcal{P}}_{\tilde{t}}$ satisfies the strong Feller property. Since Γ_1 is compact by Assumption 6.2, we see that there exists at least one invariant measure by Theorem F.3, which further implies that there exists at least on extremal invariant measure by applying Krein-Milman theorem. Moreover, combining the result of Lemma E.4 and Lemma 4.1 in Kliemann [23], we see that $\{\widetilde{\mathcal{P}}_{\tilde{t}}\}_{\tilde{t}\geq 0}$ has a unique invariant measure, which we denote by π. Finally, by Proposition 5.1 in Ichihara and Kunita [18], $\widetilde{\mathcal{P}}_{\tilde{t}}$ is ergodic in the sense that

$$\lim_{\tilde{t}\to\infty} \|\widetilde{\mathcal{P}}_{\tilde{t}}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} = 0, \quad \forall x \in \mathrm{supp}(\pi)$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance and $\mathrm{supp}(\pi) = \Gamma_1$ again by Lemma 4.1 in Kliemann [23]. This completes the proof. □

E.3 Proof for the Mixing When Trace is not Constant

We continue to use the notation $f_i(x) = \partial\Phi(x)\sigma_i(x)$ for each $i \in [\Xi]$ but we change the definition of $f_0(x)$ to $\frac{1}{2}\left(-x + \partial^2\Phi(x)[\Sigma(x)] - \frac{1}{2}\sum_{i=1}^{\Xi} \partial f_i(x) f_i(x)\right)$. Writing the dynamics of Y in (12) as a Stratonovich SDE, we get

$$dY(t) = f_0(Y(t)\mathrm{d}t + \sum_{i=1}^{\Xi} f_i(Y(t)) \circ \mathrm{d}W_i(t).$$

Then we define its associated deterministic control system:

$$\frac{\mathrm{d}\psi(t)}{\mathrm{d}t} = f_0(\psi(t)) + \sum_{i=1}^{\Xi} f_i(\psi(t))u_i(t) \tag{34}$$

Similar to Lemma E.4, we have the following lemma.

Lemma E.5. *Under Assumption 6.2 and 6.3, Γ_1 itself is the unique invariant control set contained in Γ_1 for the control system (34).*

Note that $\mathrm{tr}[\Sigma]$ is continuous and Γ_1 bounded, we know that both $\max_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)]$ and $\min_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)]$ can be attained in Γ_1. By Assumption 6.3, we know that $\min_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)] > 0$. By Lemma 6.1, we know if $\|Y(t)\|_2 \geq (\max_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)])^{0.25}$ then $\frac{\mathrm{d}\|Y(t)\|_2^2}{\mathrm{d}t} \leq 0$ and that if $\|Y(t)\|_2 \leq (\max_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)])^{0.25}$ then $\frac{\mathrm{d}\|Y(t)\|_2^2}{\mathrm{d}t} \geq 0$. Thus $\Gamma_{a,b} := \{x \in \Gamma \mid a \leq b\}$ is an invariant set for both stochastic process (12) and control system (34) if $a \leq (\min_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)])^{0.25}$ and $b \geq (\max_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)])^{0.25}$. We also denote $\{x \in \Gamma \mid \min_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)] < \|x\|_2^4 < \max_{x\in\Gamma_1} \mathrm{tr}[\Sigma(x)]\}$ by Γ_*. Without loss of generality, we can pick some large enough b and small enough a such that $x_{\text{init}} \in \Gamma_{a,b}$ and use $\Gamma_{a,b}$ as our domain. In this way, our domain is compact and we denote it by $\widehat{\Gamma}$.

Lemma E.6. *Under Assumption 6.2 and 6.3, Γ_* is the unique invariant control set contained in Γ for the control system (34).*

Proof of Lemma E.6. Note that $\overline{\Gamma_*}$ is an invariant set, it suffices to show that for every $x \in \Gamma$, $\overline{\Gamma_*} \subset \overline{O^+(x)}$. Below we will show that for any $y \in \Gamma_*$, $x \in \Gamma$, $y \in \overline{O^+(x)}$.

By Lemma E.5, we know that for any $x \in \Gamma$, we have that $\Gamma_{\|x\|_2} \subset \overline{O^+(x)}$. Therefore, it suffices to show there exist $u \in \Gamma_{\|x\|_2}$ and $v \in \Gamma_{\|y\|_2}$ such that $v \in \overline{O^+(u)}$. If this is true, then we know that $y \in \overline{O^+(v)} \subset \overline{O^+(u)} \subset \overline{O^+(x)}$.

Without loss of generality, we assume that $\|x\|_2 \geq \|y\|_2$. Our strategy is to pick any $z \in \operatorname{argmin}_{z \in \Gamma_1} \operatorname{tr}[\Sigma(z)]$ and let $v = \|y\|_2 z$ and $u = \|x\|_2 z$. Then we can split the control $u_\delta : \mathbb{R}_+ \to \mathbb{R}^\Xi$ into many small piece-wise constant intervals, such that in each interval, $\sum_{i=1}^{\Xi} f_i(\psi(t)) u_i(t)$ cancels the tangent component of $f_0(\psi(t))$ (in the tangent space of $\Gamma_{\|\psi(t)\|_2}$ at $\psi(t)$). Thus through the whole trajectory, the direction of $\psi(t)$ can be arbitrarily close to z. That is, we claim there exists a sequence of control $\{u^n\}_{n \in \mathbb{N}}$, such that the corresponding $\psi^n(t) \to \psi^\infty(t)$ pointwise, where ψ^∞ is defined as the solution of the following ODE.

$$\frac{\mathrm{d}\psi^\infty(t)}{\mathrm{d}t} = (I - \frac{\psi(t)}{\|\psi(t)\|_2} \frac{\psi(t)^\top}{\|\psi(t)\|_2}) f_0(\psi(t)) \quad \text{with} \phi(0) = \|x\|_2 z.$$

By definition of z, we know $\|y\|_2 \geq \|z\|_2$. Thus there exists some finite time t such that $\psi^\infty(t) = v$. That is, $v \in \overline{O^+(u)}$. This completes the proof. □

Lemma E.7 (Lie Algebra Rank Condition)**.** *For any $k \geq 1, k \in \mathbb{N}$, define $A^k := \{[[[f_{i_1}, f_{i_2}], f_{i_3}], \ldots, f_{i_k}] \mid i_1 \in [\Xi], i_2, \ldots, i_k \in \{0\} \cup [\Xi]\}$ and $A = \cup_{k=1}^\infty A_k$. For any $x \in \Gamma$, $\operatorname{rank}(\operatorname{span}(\{f(x) \mid f \in A\})) = D - M$.*

Proof of Lemma E.7. Note the $\operatorname{rank}(T_x(\Gamma_{\|x\|_2})) = D - M - 1$ for any $x \in \Gamma$, by Assumption 6.3, it suffices to show for each $x \in \Gamma$, there exists some $f \in A$ and $\langle x, f(x) \rangle \neq 0$. Since $\operatorname{tr}[\Sigma]$ is an non-constant analytic function and Γ_1 is an analytic manifold, for every $x \in \Gamma_1$, there must exists some $k^* \geq 1$ and some linear combination of f_i, $i \in [\Xi]$, denoted by f^*, such that

$$\langle \cdots \langle \langle \operatorname{tr}[\Sigma], f^* \rangle, f^* \rangle \cdots, f^* \rangle (x) \neq 0. \tag{35}$$

By Lemma 3.5 and the fact that $\operatorname{tr}[\Sigma]$ is (-2)-homogeneous, we know that $\operatorname{tr}[\Sigma](x) = -\frac{1}{2} \langle x, f_0(x) + \frac{x}{2} \rangle$. Then we can apply Lemma E.8 iteratively on Equation (35) and get that

$$[[[f_0(x) + \frac{x}{2}, f^*], f^*], \ldots, x] \neq 0. \tag{36}$$

The proof is completed by noting that $[[[\frac{x}{,} f^*], f^*], \ldots, x] = 0$ by apply the second claim of Lemma E.8 iteratively. □

Lemma E.8. *Let f, g be any two vector field on Γ, where $\langle f(x), x \rangle = 0$ for every $x \in \Gamma$. It holds that $\langle x, [f, g] \rangle = \langle \nabla \langle x, g \rangle, f \rangle$. In specific, $\langle x, [x, f] \rangle = 0$.*

Proof of Theorem 6.6 when trace is not constant. Let $\{\mathcal{P}_t\}_{t \geq 0}$ be the Markov semigroup associated with the limiting diffusion. Since the manifold $\widehat{\Gamma}$ is compact, the space of the probability measures on $\widehat{\Gamma}$ is also compact, thus for any probability measure on $\widehat{\Gamma}$, the sequence $\{\mathcal{P}_t \mu_0\}_{t \geq 0} \subseteq \mathcal{P}(\widehat{\Gamma})$ is tight. Therefore, by Theorem F.3, there exists at least one invariant probability measure for $\{\mathcal{P}_t\}_{t \geq 0}$, which further implies that there exists at least on extremal invariant measure by applying Krein-Milman theorem. Moreover, combining the result of Lemma E.6 and Lemma 4.1 in Kliemann [23], we see that $\{\mathcal{P}_t\}_{t \geq 0}$ has a unique invariant measure, which we denote by π. Finally, by Proposition 5.1 in Ichihara and Kunita [18], $\mathcal{P}_t$ is ergodic in the sense that

$$\lim_{t \to \infty} \|\mathcal{P}_{\tilde{t}}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} = 0, \quad \forall x \in \operatorname{supp}(\pi)$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance and $\operatorname{supp}(\pi) = \Gamma_1$ again by Lemma 4.1 in Kliemann [23]. This completes the proof.

□

F Auxiliary Results

The following lemma is a classic result. See, e.g., Corollary 8.5.5 in Oksendal [34].

Lemma F.1 (Time-changed Brownian motion)**.** *Let $g : [0, \infty) \to [0, \infty)$ be a differentiable function with nonnegative first-order derivative. Then it holds that*

$$W(g(t)) \stackrel{d}{=} \int_{s=0}^{t} \sqrt{g'(t)} \mathrm{d}W(s).$$

The multidimensional martingale functional central limit theorem can be found in, e.g., Theorem 7.1 in Ethier and Kurtz [11] or Theorem 2.1 in Whitt [44]. We use a version of Theorem 2.1 in Whitt [44] given in Theorem F.2.

Theorem F.2 (Multidimensional martingale FCLT, [44])**.** *For $n \geq 1$, let $M_n = (M_{n,1}, \ldots, M_{n,k})$ be a martingale satisfying $M_n(0) = (0, \ldots, 0)$. Let $C \in \mathbb{R}^{k \times k}$ be a semi-positive definite matrix. Suppose the following holds:*

- *The expected value of the maximum jump in M_n is asymptotically negligible, i.e., for each $T > 0$,*

$$\lim_{n \to \infty} \mathbb{E}[J(M_n, T)] = 0 \tag{37}$$

  where $J(x, T) = \sup\{|x(t) - x(t-)| : 0 < t \leq T\}$ for any function $x : [0, \infty) \to \mathbb{R}$.
- *For each pair $(i, j) \in [k] \times [k]$ and each $t > 0$,*

$$[M_{n,i}, M_{n,j}](t) \Rightarrow C_{ij} t \text{ as } n \to \infty. \tag{38}$$

Then $M_n \Rightarrow \sqrt{C} W$ in Skorokhod metric where W is the k-dimensional standard Brownian motion.

Theorem F.3 (Krylov-Bogolioubov Theorem, [9])**.** *Let $\{\mathcal{P}_t\}_{t \geq 0}$ be a Feller Markov semigroup over a Polish space $\mathcal{X}$. Assume that there exists $\mu_0 \in \mathcal{M}_1(\mathcal{X})$ such that the sequence $\{\mathcal{P}_t \mu_0\}_{t \geq 0}$ is tight. Then there exists at least one invariant probability measure for $\{\mathcal{P}_t\}_{t \geq 0}$.*