OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Why (and When) does Local SGD Generalize Better than SGD?

Xinran Gu[*] GXR21@MAILS.TSINGHUA.EDU.CN
IIIS, Tsinghua University, China
Kaifeng Lyu[*] KLYU@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, USA
Longbo Huang[†] LONGBOHUANG@TSINGHUA.EDU.CN
IIIS, Tsinghua University, China
Sanjeev Arora [†] ARORA@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, USA

Abstract

Local SGD is a communication-efficient variant of SGD for large-scale training, where multiple GPUs perform SGD independently and average the model parameters periodically. It has been recently observed that Local SGD can not only achieve the design goal of reducing the communication overhead but also lead to higher test accuracy than the corresponding SGD baseline [39], though the training regimes for this to happen are still in debate [47]. This paper aims to understand why (and when) Local SGD generalizes better based on Stochastic Differential Equation (SDE) approximation. The main contributions of this paper include (i) the derivation of an SDE that captures the long-term behavior of Local SGD with a small learning rate, after approaching the manifold of minima, (ii) a comparison between the SDEs of Local SGD and SGD, showing that Local SGD induces a stronger drift term that can result in a stronger effect of regularization, e.g., a faster reduction of sharpness, and (iii) empirical evidence validating that having small learning rate and long enough training time enables the generalization improvement over SGD but removing either of the two conditions leads to no improvement.

1. Introduction

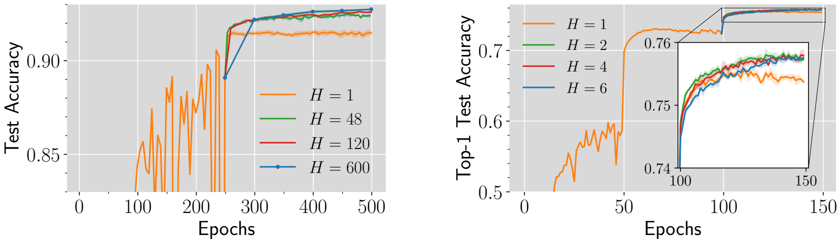
Recent advances suggest that the ultimate performance of deep learning on test sets can be drastically improved by scaling up the dataset and increasing the model size, but this requires more computation. In response, recent works [15, 24, 65] seek to speed up standard training methods by exploiting data parallelism in a distributed computing setting and most works focus on improving Stochastic Gradient Descent (SGD).

SGD tries to solve problems of the form $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{\xi\sim\tilde{\mathcal{D}}}[\ell(\boldsymbol{\theta};\xi)]$, where $\boldsymbol{\theta}\in\mathbb{R}^d$ is the parameter vector of the model, $\ell(\boldsymbol{\theta};\xi)$ is the loss function for a data sample ξ drawn from the training distribution $\tilde{\mathcal{D}}$, e.g., the uniform distribution over the training set. SGD with learning rate η and batch size B does the following update at each step, using a batch of B independent $\xi_{t,1},\dots,\xi_{t,B}\sim\tilde{\mathcal{D}}$:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \boldsymbol{g}_t, \quad \text{where} \quad \boldsymbol{g}_t = \frac{1}{B}\sum_{i=1}^{B}\nabla\ell(\boldsymbol{\theta}_t;\xi_{t,i}). \tag{1}$$

Parallel SGD tries to improve efficiency by distributing the gradient computation to $K \geq 2$ workers, each of whom focuses on a local batch of $B_{\mathrm{loc}} := B/K$ samples and computes the average gradient over the local batch. Finally, $\boldsymbol{g}_t$ is obtained by averaging the local gradients over the K workers.

[*] Equal contribution
[†] Corresponding author



(*a*) CIFAR-10, $B = 4096$, ResNet-56. (*b*) ImageNet, $B = 8192$, ResNet-50.

Figure 1: Post-Local SGD ($H > 1$) generalizes better than SGD ($H = 1$). We switch to Local SGD at the first learning rate decay (epoch 250) for CIFAR-10 and at the second learning rate decay (epoch 100) for ImageNet. See Appendix L.1 for training details.

In the ideal case where the batch size B is large enough, the speedup of parallel SGD over single-machine SGD (in terms of wall-clock time) can be linear in the number of workers K. However, large-batch training leads to a significant test accuracy drop compared to a small-batch training baseline with the same number of training steps or epochs [23, 27, 53, 55], which probably stems from a low level of gradient noise (see Appendix A for discussion).

A variant of SGD, called *Local SGD* [61, 67, 69] has been recently observed to help resolve the generalization degradation in certain regimes of large-batch training. Perhaps surprisingly, Local SGD is not designed for better generalization, but for reducing the high communication cost, another important issue that bottlenecks large-batch training [5, 50, 52, 58].

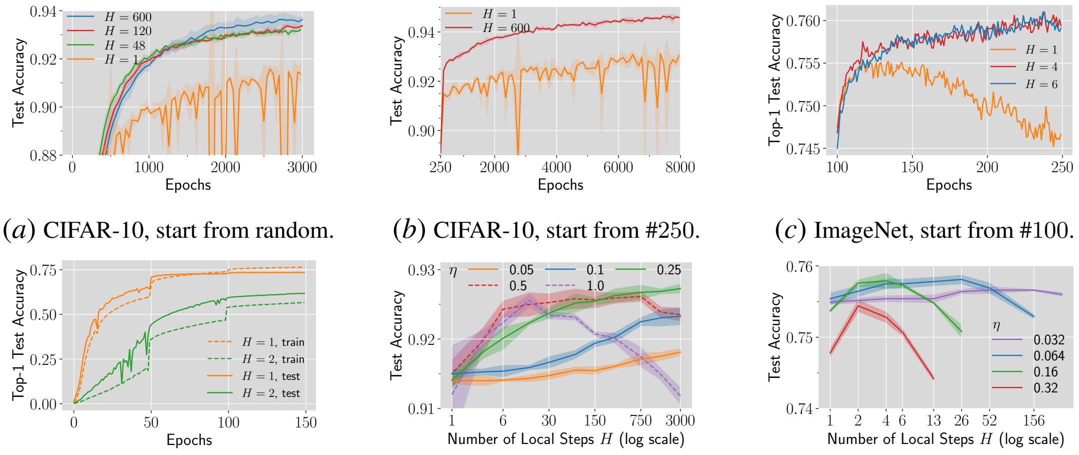
Instead of averaging the local gradients per step as in parallel SGD, Local SGD averages the local *model parameters* on the K workers whenever they finish H local steps of SGD (reducing the communication by a factor H). Formally, Local SGD proceeds in multiple rounds of model averaging, where each round produces a global iterate $\bar{\boldsymbol{\theta}}^{(s)}$. In the $(s+1)$-th round, every worker $k \in [K]$ starts with its local copy of the global iterate $\boldsymbol{\theta}^{(s)}_{k,0} \leftarrow \bar{\boldsymbol{\theta}}^{(s)}$ and does H steps of SGD with local batches. The k-th worker at its t-th local step draws a local batch of $B_{\mathrm{loc}} := B/K$ independent samples $\xi^{(s)}_{k,t,1}, \dots, \xi^{(s)}_{k,t,B_{\mathrm{loc}}}$ from a shared training distribution $\tilde{\mathcal{D}}$ and updates as follows:

$$\boldsymbol{\theta}^{(s)}_{k,t+1} \leftarrow \boldsymbol{\theta}^{(s)}_{k,t} - \eta \boldsymbol{g}^{(s)}_{k,t}, \quad \text{where} \quad \boldsymbol{g}^{(s)}_{k,t} = \frac{1}{B_{\mathrm{loc}}} \sum_{i=1}^{B_{\mathrm{loc}}} \nabla \ell(\boldsymbol{\theta}_t; \xi^{(s)}_{k,t,i}), \quad t = 0, \dots, H-1. \tag{2}$$

After finishing the H local steps, the workers aggregate the resulting local iterates $\boldsymbol{\theta}^{(s)}_{k,H}$ and assign the average to the next global iterate: $\bar{\boldsymbol{\theta}}^{(s+1)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\theta}^{(s)}_{k,H}$.

Lin et al. [39] discovered that Local SGD can be used as a strong component to improve generalization in large-batch training. They proposed *Post-local SGD*, a hybrid method that starts with parallel SGD (equivalent to Local SGD with $H = 1$ in math) and switches to Local SGD with $H > 1$ after a fixed number of steps t_0. Following a standard training procedure with momentum and multiple learning rate decays, they showed through extensive CIFAR-10 [31] experiments that Post-local SGD significantly outperforms parallel SGD in test accuracy, where t_0 is carefully chosen to be the time of the first learning rate decay. Ortiz et al. [47] conducted a large-scale empirical study on ImageNet [51] but found that switching at the first learning rate decay can hurt generalization; instead, switching at a later time may improve validation accuracy. Our experiments on CIFAR-10 and ImageNet (Figure 1) reproduce the generalization improvement of Post-local SGD.

The success of Post-local SGD suggests that Local SGD induces a generalization benefit if the training starts from a model pre-trained by (parallel) SGD, though Local SGD is designed for a



(a) CIFAR-10, start from random.
(b) CIFAR-10, start from #250.
(c) ImageNet, start from #100.
(d) ImageNet, $\eta_1 = 3.2$.
(e) CIFAR-10, test acc v.s. H.
(f) ImageNet, test acc v.s. H.

Figure 2: Effect of η, H and training time. See Appendix L.2 for training details.

different purpose. The current paper tries to understand: *Why does Local SGD generalize better? Under what general conditions does this generalization benefit arise?*

Previous theoretical research on Local SGD is mainly restricted to the convergence rate for minimizing a convex or non-convex objective (see Appendix A for a survey), and the generalization aspect of Local SGD is still unclear.

Our Contributions. In this paper, we provide the first theoretical understanding on why (and when) switching from parallel SGD to Local SGD improves generalization.

1. We conduct ablation studies on CIFAR-10 and ImageNet and identify that **small learning rate** and **sufficient training time** are two important factors that contribute to the generalization improvement of Local SGD over the corresponding (parallel) SGD baseline.
2. Inspired by a previous analysis of the long-term generalization benefit of SGD [37], we derive an SDE for Local SGD that can track its long-term behavior with a novel proof.
3. We explain the generalization improvement of Local SGD over SGD through comparison with the corresponding SDEs: increasing the number of local steps H strengthens the drift term of SDE. We then connect the stronger drift term to a stronger implicit regularization effect.

2. When does Local SGD Generalize Better?

It is still in debate under what general conditions this generalization benefit arises; see Appendix B.1 for a review. In response, we focus on a simple yet insightful setup: the learning rate η is constant with time, and we train SGD and Local SGD without additional tricks (e.g., without momentum). We observed the generalization improvement of Local SGD over SGD even in this setting (Figure 1). We further conduct ablation studies and identify that **small learning rate** and **sufficient training time** are key factors that enable the improved generalization of Local SGD over its SGD counterpart.

Finding 2.1 *Given a sufficiently small learning rate η and sufficiently long training time, Local SGD exhibits better generalization than (parallel) SGD (with the same η, B_{loc}, K), if the number of local steps H is tuned properly according to η. This holds for both training from random initialization and from SGD pre-trained models.*

Our finding can help to settle the debate to a large extent; see Appendix B.3 for discussion. Now we go through each point of our finding. **(1). Pretraining is not necessary.** In contrast to previous works claiming the benefits of Post-local SGD over Local SGD [39, 47], we observe that

Local SGD with random initialization also generalizes significantly better than SGD, as long as η is small and the training time is sufficiently long (Figure 2(*a*)). **(2). The learning rate should be small.** The learning rate is 0.32 for Figures 2(*a*) and 2(*b*) and is 0.16 for Figure 2(*c*). As shown in Figure 2(*d*), Local SGD encounters optimization difficulty in the first phase when η is large (e.g., $\eta = 3.2$), resulting in inferior final test accuracy. Even for training from a pre-trained model, the generalization improvement of Local SGD disappears for large learning rates (e.g., $\eta = 1.6$ in Figure 3(*d*)). In contrast, Figure 3(*c*) shows that Local SGD achieves a comparable test accuracy with a much smaller learning rate $\eta = 0.064$ (with H and the training budget set properly). **(3). Training time should be long enough.** In Figures 2(*b*) and 2(*c*), we extend the training budget for the Post-local SGD experiments in Section 1 and observe that a longer training time leads to greater generalization improvement upon SGD. On the other hand, Local SGD generalizes worse than SGD in the first few epochs of Figures 2(*a*) and 2(*c*); see Figures 3(*a*) and 3(*b*) for an enlarged view. **(4). The number of local steps H should be tuned carefully.** The number of local steps H has a complex interplay with the learning rate η, but generally speaking, the test accuracy first rises as H increases, then it begins to fall when H is too large. A smaller η needs a higher H to achieve consistent generalization improvement. See Figures 2(*e*) and 2(*f*).

3. Theoretical Analysis of Local SGD: The Slow SDE

In this section, we use an SDE-based approach to establish the generalization benefit of Local SGD. Below, we first identify the difficulty of adapting the conventional SDE framework to Local SGD. Then, we present our new SDE and explain the generalization benefit.

Notations. We follow the notations in Section 1. We also define $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{\xi\sim\tilde{\mathcal{D}}}[\ell(\boldsymbol{\theta};\xi)]$ as the expected loss, $\boldsymbol{\Sigma}(\boldsymbol{\theta}) := \mathrm{Cov}_{\xi\sim\tilde{\mathcal{D}}}[\nabla\ell(\boldsymbol{\theta};\xi)]$ as the noise covariance of gradients at $\boldsymbol{\theta}$. Let $\{\boldsymbol{W}_t\}_{t\geq 0}$ denote the standard Wiener process. For a mapping $F : \mathbb{R}^d \to \mathbb{R}^d$, denote by $\partial F(\boldsymbol{\theta})$ the Jacobian at $\boldsymbol{\theta}$ and $\partial^2 F(\boldsymbol{\theta})$ the second order derivative at $\boldsymbol{\theta}$. Furthermore, for any matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, $\partial^2 F(\boldsymbol{\theta})[\boldsymbol{M}] = \sum_{i\in[d]}\langle \frac{\partial^2 F_i}{\partial \boldsymbol{\theta}^2}, \boldsymbol{M}\rangle \boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the i-th vector of the standard basis. We write $\partial^2(\nabla\mathcal{L})(\boldsymbol{\theta})[\boldsymbol{M}]$ as $\nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{M}]$ for short.

3.1. Difficulty of Adapting the SDE Framework to Local SGD

A widely-adopted approach to understanding the dynamics of SGD is to approximate it with the following SDE (3), which we call the *conventional SDE approximation*. Below, we discuss why it cannot be directly adopted to characterize the behavior of Local SGD.

$$\mathrm{d}\boldsymbol{X}(t) = -\nabla\mathcal{L}(\boldsymbol{X})\mathrm{d}t + \sqrt{\tfrac{\eta}{B}}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X})\mathrm{d}\boldsymbol{W}_t. \tag{3}$$

It is proved by Li et al. [34] that this SDE is a first-order approximation to SGD, where each discrete step corresponds to a continuous time interval of η. By Finding 2.1, it is tempting to consider the limit $\eta \to 0$ and see if Local SGD can also be modeled via a variant of the conventional SDE. In this case the typical time length that guarantees a good SDE approximation error is $\mathcal{O}(\eta^{-1})$ discrete steps [34, 36]. However, this time scaling is too short for the difference to appear between Local SGD and SGD. We show in Theorem 9 that they closely track each other for $\mathcal{O}(\eta^{-1})$ steps. See also Appendix D for Lin et al. [39]'s attempt to model Local SGD with multiple conventional SDEs and for our discussion on why it does not give much insight.

3.2. SDE Approximation near the Minimizer Manifold

Inspired by a recent paper [37], our strategy to overcome the shortcomings of the conventional SDE is to design a new SDE that can guarantee a good approximation for $\mathcal{O}(\eta^{-2})$ discrete steps.

Following their setting, we assume the existence of a manifold Γ consisting only of local minimizers and track the global iterate $\bar{\boldsymbol{\theta}}^{(s)}$ around Γ after it takes $\tilde{\mathcal{O}}(\eta^{-1})$ steps to approach Γ.

Assumption 3.1 *The loss function $\mathcal{L}(\cdot)$ and the matrix square root of the covariance $\boldsymbol{\Sigma}^{1/2}(\cdot)$ are $\mathcal{C}^{\infty}$-smooth. Besides, we assume that $\|\nabla \ell(\boldsymbol{\theta};\xi)\|_2$ is bounded by a constant for all $\boldsymbol{\theta}$ and ξ.*

Assumption 3.2 *Γ is a $\mathcal{C}^{\infty}$-smooth, $(d-m)$-dimensional submanifold of $\mathbb{R}^d$, where any $\boldsymbol{\zeta} \in \Gamma$ is a local minimizer of $\mathcal{L}$. For all $\boldsymbol{\zeta} \in \Gamma$, $\operatorname{rank}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) = m$. Additionally, there exists an open neighborhood of Γ, denoted as U, such that $\Gamma = \arg\min_{\boldsymbol{\theta}\in U} \mathcal{L}(\boldsymbol{\theta})$.*

Assumption 3.3 *Γ is a compact manifold.*

The existence of a minimizer manifold with $\operatorname{rank}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) = m$ has also been made as a key assumption in Fehrman et al. [11], Li et al. [37], Lyu et al. [40]. $\operatorname{rank}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) = m$ ensures that the Hessian is maximally non-degenerate on the manifold and implies that the tangent space at $\boldsymbol{\zeta} \in \Gamma$ equals the null space of $\nabla^2\mathcal{L}(\boldsymbol{\zeta})$.

Our SDE for Local SGD characterizes the training dynamics near Γ. For ease of presentation, we define the following projection operators Φ, $P_{\boldsymbol{\zeta}}$ for points and differential forms.

Definition 1 (Gradient Flow Projection) *Fix a point $\boldsymbol{\theta}_{\text{null}} \notin \Gamma$. For $\boldsymbol{x} \in \mathbb{R}^d$, consider the gradient flow $\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\nabla\mathcal{L}(\boldsymbol{x}(t))$ with $\boldsymbol{x}(0) = \boldsymbol{x}$. We denote the gradient flow projection of $\boldsymbol{x}$ as $\Phi(\boldsymbol{x})$. $\Phi(\boldsymbol{x}) := \lim_{t\to+\infty} \boldsymbol{x}(t)$ if the limit exists and belongs to Γ; otherwise, $\Phi(\boldsymbol{x}) = \boldsymbol{\theta}_{\text{null}}$.*

Definition 2 *For any $\boldsymbol{\zeta} \in \Gamma$ and any differential form $\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t$ in Itô calculus, where $\boldsymbol{A}$ is a matrix and $\boldsymbol{b}$ is a vector, we use $P_{\boldsymbol{\zeta}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t)$ as a shorthand for the differential form $\partial\Phi(\boldsymbol{\zeta})\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \left(\partial\Phi(\boldsymbol{\zeta})\boldsymbol{b} + \frac{1}{2}\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{A}\boldsymbol{A}^\top]\right)\mathrm{d}t$.*

Here $P_{\boldsymbol{\zeta}}$ equals $\Phi(\boldsymbol{\zeta} + \boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t) - \Phi(\boldsymbol{\zeta})$ by Itô calculus, which means $\boldsymbol{\zeta} + P_{\boldsymbol{\zeta}}(\boldsymbol{A}\mathrm{d}\boldsymbol{W}_t + \boldsymbol{b}\mathrm{d}t)$ does not leave Γ. It can be shown that $\partial\Phi(\boldsymbol{\zeta})$ equals the projection matrix onto the tangent space of Γ at $\boldsymbol{\zeta}$. We decompose the noise covariance $\boldsymbol{\Sigma}(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \in \Gamma$ into the tangent space part $\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta}) := \partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}(\boldsymbol{\zeta})\partial\Phi(\boldsymbol{\zeta})$ and the rest part $\boldsymbol{\Sigma}_{\Diamond}(\boldsymbol{\zeta}) := \boldsymbol{\Sigma}(\boldsymbol{\zeta}) - \boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta})$. Now we are ready to state our SDE.

Definition 3 (Slow SDE for Local SGD) *Given $\eta, H > 0$ and $\boldsymbol{\zeta}_0 \in \Gamma$, define $\boldsymbol{\zeta}(t)$ as the solution of the following SDE with initial condition $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0$:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t}_{\text{(a) diffusion}}\underbrace{-\tfrac{1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{(b) drift-I}}\underbrace{-\tfrac{K-1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{(c) drift-II}}\Big). \tag{4}$$

Here $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})$, $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta}) \in \mathbb{R}^{d\times d}$ are defined as

$$\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta}) := \textstyle\sum_{i,j:(\lambda_i\neq 0)\vee(\lambda_j\neq 0)} \frac{1}{\lambda_i+\lambda_j}\left\langle \boldsymbol{\Sigma}_{\Diamond}(\boldsymbol{\zeta}), \boldsymbol{v}_i\boldsymbol{v}_j^\top\right\rangle \boldsymbol{v}_i\boldsymbol{v}_j^\top, \tag{5}$$

$$\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta}) := \textstyle\sum_{i,j:(\lambda_i\neq 0)\vee(\lambda_j\neq 0)} \frac{\psi(\eta H\cdot(\lambda_i+\lambda_j))}{\lambda_i+\lambda_j}\left\langle \boldsymbol{\Sigma}_{\Diamond}(\boldsymbol{\zeta}), \boldsymbol{v}_i\boldsymbol{v}_j^\top\right\rangle \boldsymbol{v}_i\boldsymbol{v}_j^\top, \tag{6}$$

where $\{\boldsymbol{v}_i\}_{i=1}^d$ is a set of eigenvectors of $\nabla^2\mathcal{L}(\boldsymbol{\zeta})$ that forms an orthonormal eigenbasis, $\lambda_1, \dots, \lambda_d$ are the corresponding eigenvalues. Additionally, $\psi(0) = 0$ and $\psi(x) := \frac{e^{-x}-1+x}{x}$ for $x \neq 0$.

The use of $P_{\boldsymbol{\zeta}}$ keeps $\boldsymbol{\zeta}(t)$ on Γ through projection. $\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})$ introduces a diffusion term to the SDE in the tangent space. The two drift terms involve $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\cdot)$ and $\widehat{\boldsymbol{\Psi}}(\cdot)$, which can be intuitively understood as rescaling the entries of the noise covariance in the eigenbasis of Hessian. In the special case where $\nabla^2\mathcal{L} = \operatorname{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d\times d}$, we have $\widehat{\Sigma}_{\Diamond,i,j} = \frac{1}{\lambda_i+\lambda_j}\Sigma_{0,i,j}$. $\widehat{\Psi}_{i,j} = \frac{\psi(\eta H(\lambda_i+\lambda_j))}{\lambda_i+\lambda_j}\Sigma_{0,i,j}$. $\psi(x)$ increases from 0 to 1 as x goes from 0 to infinity (see Figure 7)

We name this SDE as the *Slow SDE for Local SGD* because we will show that each discrete step of Local SGD corresponds to a continuous time interval of η^2 instead of an interval of η in the conventional SDE. This Slow SDE is inspired by Li et al. [37]. Under nearly the same set of assumptions, they proved that SGD can be tracked by an SDE that is essentially equivalent to (4) with $K = 1$, namely, without the drift-II term:

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}(t)}_{\text{(a) diffusion}} \underbrace{-\tfrac{1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{(b) drift-I}}\Big). \tag{7}$$

We refer to (7) as the *Slow SDE for SGD*. We remark that the drfit-II term in (4) is novel and is the key to separate the generalization behaviors of Local SGD and SGD in theory. We will discuss this point later in Section 3.3. Now we present our SDE approximation theorem for Local SGD.

Theorem 4 *Let Assumptions 3.1 to 3.3 hold. Let $T > 0$ be a constant and $\boldsymbol{\zeta}(t)$ be the solution to (4) with the initial condition $\boldsymbol{\zeta}(0) = \Phi(\bar{\boldsymbol{\theta}}^{(0)}) \in \Gamma$. If H is set to $\frac{\alpha}{\eta}$ for some constant $\alpha > 0$, then for any $\mathcal{C}^3$-smooth function $g(\boldsymbol{\theta})$, $\max_{0 \le s \le \frac{T}{H\eta^2}} \left|\mathbb{E}[g(\Phi(\bar{\boldsymbol{\theta}}^{(s)})] - \mathbb{E}[g(\boldsymbol{\zeta}(sH\eta^2)]\right| = \tilde{\mathcal{O}}(\eta^{0.25})$, where $\tilde{\mathcal{O}}(\cdot)$ hides log factors and constants that are independent of η but can depend on $g(\boldsymbol{\theta})$.*

Theorem 5 *For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1 - \delta$, it holds for all $\mathcal{O}(\frac{1}{\alpha}\log\frac{1}{\eta}) \le s \le \frac{T}{\alpha\eta}$ that $\Phi(\bar{\boldsymbol{\theta}}^{(s)}) \in \Gamma$ and $\|\bar{\boldsymbol{\theta}}^{(s)} - \Phi(\bar{\boldsymbol{\theta}}^{(s)})\|_2 = \mathcal{O}(\sqrt{\alpha\eta\log\frac{\alpha}{\eta\delta}})$, where $\mathcal{O}(\cdot)$ hides constants independent of η, α and δ.*

Theorem 4 suggests that the trajectories of the manifold projection and the solution to the Slow SDE (4) are close to each other in the weak approximation sense. Theorem 5 further states that the iterate $\bar{\boldsymbol{\theta}}^{(s)}$ keeps close to its manifold projection after the first few rounds.

Remark 6 *To connect to Finding 2.1, we remark that our theorems (1) do not require the model to be pre-trained (as long as the gradient flow starting with $\boldsymbol{\theta}^{(0)}$ can converge to Γ); (2) give better bounds for smaller η; (3) characterize a long training horizon $\sim \eta^{-2}$. The need for tuning H will be discussed in Remark 8.*

3.3. Interpretation of the Slow SDEs

In this subsection, we compare the Slow SDEs for SGD and Local SGD and provide an important insight into why Local SGD generalizes better than SGD: Local SGD strengthens the drift term in the Slow SDE which makes the implicit regularization of stochastic gradient noise more effective.

Interpretation of the Slow SDE for SGD. The Slow SDE for SGD (7) consists of the diffusion and drift-I terms. The former injects noise into the dynamics; the latter one drives the dynamics to move along the negative gradient of $\frac{1}{2B}\langle\nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\rangle$ projected onto the tangent space, but ignoring the dependency of $\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})$ on $\boldsymbol{\zeta}$. This can be connected to the class of semi-gradient methods which only computes a part of the gradient [4, 45, 60]. In this view, the long-term behavior of SGD is similar to a stochastic semi-gradient method minimizing the implicit regularizer $\frac{1}{2B}\langle\nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\rangle$ on Γ. This argument reveals that SGD has a deterministic trend toward the region with a smaller magnitude of Hessian, which is commonly believed to correlate with better generalization [19, 25, 27, 46]. In contrast, the diffusion term can be regarded as a random perturbation to this trend, which can impede optimization when the drift-I term is not strong enough. Based on this view, we conjecture that **strengthening the drift term** of the Slow SDE can help SGD to better regularize the model, yielding a better generalization performance. More specifically, we propose the following hypothesis, which compares the generalization performances of the following generalized Slow SDEs. Note that $(\frac{1}{B}, \frac{1}{2B})$-Slow SDE corresponds to the Slow SDE for SGD (7).

Definition 7 *For $\kappa_1, \kappa_2 \geq 0$, define (κ_1, κ_2)-Slow SDE to be the following:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\sqrt{\kappa_1}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}(t) - \kappa_2 \nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t\Big). \tag{8}$$

Hypothesis 3.1 *Starting at a minimizer $\boldsymbol{\zeta}_0 \in \Gamma$, run (κ_1, κ_2)-Slow SDE and (κ_1, κ_2')-Slow SDE respectively for the same amount of time $T > 0$ and obtain $\boldsymbol{\zeta}(T), \boldsymbol{\zeta}'(T)$. If $\kappa_2 > \kappa_2'$, then the expected test accuracy at $\boldsymbol{\zeta}(T)$ is better than that at $\boldsymbol{\zeta}'(T)$.*

One motivating example is SGD with label noise regularization. In this case, the Slow SDE for SGD turns out as a simple gradient flow on Γ aimed at minimizing $\mathrm{tr}(\nabla^2\mathcal{L})$, and larger drift term means flatter minima; see Appendix E. Due to the No Free Lunch Theorem, we do not claim that our hypothesis is always true, but we do believe that the hypothesis holds when training usual neural networks (e.g., ResNets, VGGNets) on standard benchmarks (e.g., CIFAR-10, ImageNet).

Local SGD Strengthens the Drift Term in Slow SDE. Our hypothesis can help us understand why Local SGD generalizes better than SGD. The Slow SDE for Local SGD (4) has an additional drfit-II term. Similarly, this term has the effect of driving the dynamics to move along the negative semi-gradient of $\frac{K-1}{2B}\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})\rangle$ (with the dependency of $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})$ on $\boldsymbol{\zeta}$ ignored). Combining it with the implicit regularizer induced by the drift-I term, we can see that the long-term behavior of Local SGD is similar to a stochastic semi-gradient method minimizing the implicit regularizer $\frac{1}{2B}\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\rangle + \frac{K-1}{2B}\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})\rangle$ on the minimizer manifold of the original loss $\mathcal{L}$.

$\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})$ is basically a rescaling of the entries of $\widehat{\boldsymbol{\Sigma}}_{\diamond}(\boldsymbol{\zeta})$ in the eigenbasis of Hessian, where the rescaling factor $\psi(\eta H \cdot (\lambda_i + \lambda_j))$ is between 0 and 1 (see Figure 7 for the plot of ψ). When ηH is small, the rescaling factors should be close to 0, then $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta}) \approx \mathbf{0}$, leading to almost no additional regularization. On the other hand, when ηH is large, the rescaling factors should be close to 1, then $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta}) \approx \widehat{\boldsymbol{\Sigma}}_{\diamond}(\boldsymbol{\zeta})$. We can then merge the two implicit regularizers to be $\frac{K}{2B}\langle \nabla^2\mathcal{L}(\boldsymbol{\zeta}), \widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})\rangle$. The corresponding Slow SDE is approximately the $(\frac{1}{B}, \frac{K}{2B})$-Slow SDE, which is restated below:

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}(t) - \tfrac{K}{2B}\nabla^3 \mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t\Big). \tag{9}$$

Comparing the $(\frac{1}{B}, \frac{1}{2B})$-Slow SDE for SGD (7) and the $(\frac{1}{B}, \frac{K}{2B})$-Slow SDE above (9), the difference is that the drift term is amplified by K times. According to our hypothesis, we can then attribute the generalization improvement of Local SGD to the amplified drift term.

Since the Slow SDE of Local SGD is better approximated by the $(\frac{1}{B}, \frac{K}{2B})$-Slow SDE above (9) when $H\eta$ increases, our hypothesis also implies that we should increase H when decreasing η to experience the similar generalization benefit, which is consistent with our empirical finding in Section 2 that the optimal H for generalization increases as η decreases (Figures 2(*e*) and 2(*f*)).

Remark 8 *When η is small but finite, tuning H offers a trade-off between regularization strength and SDE approximation quality. Larger $\alpha := \eta H$ makes the regularization stronger in the SDE, but the SDE itself may lose track of Local SGD, which can be seen from the error bound in Theorem 5. This matches our finding that tuning H is important for better generalization (Finding 2.1).*

Prediction: Increasing the number of workers helps generalization. In addition to strengthening the drift term of the Slow SDE for SGD, another way to help the corresponding semi-gradient method to optimize the implicit regularizer is to **reduce the diffusion term**. We conduct experiments where we keep H and η fixed and gradually increase the number of workers K to reduce the diffusion term. As shown in Figure 6, a higher test accuracy is achieved for larger K.

4. Conclusions

In this paper, we provide a theoretical analysis for Local SGD that captures its long-term generalization benefit in the small learning rate regime. We derive the Slow SDE for Local SGD as a generalization of the Slow SDE for SGD [37], and attribute the generalization improvement over SGD to the larger drift term in the SDE for Local SGD. Our empirical validation show that Local SGD indeed induces generalization benefits with small learning rate and long enough training time. The main limitation of our work is that our analysis does not imply any direct theoretical separation between SGD and Local SGD in terms of test accuracy, which requires a much deeper understanding of loss landscape and the Slow SDEs and is left for future work. Another future work direction is to design a distributed training method that provably generalizes better than SGD based on the insights from Slow SDEs.

References

[1] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[2] Yoshua Bengio. *Practical Recommendations for Gradient-Based Training of Deep Architectures*, pages 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_26.

[3] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 483–513. PMLR, 09–12 Jul 2020.

[4] David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear TD learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[5] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous sgd. *arXiv preprint arXiv:1604.00981*, 2016.

[6] Kai Chen and Qiang Huo. Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5880–5884, 2016. doi: 10.1109/ICASSP.2016.7472805.

[7] Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[8] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017.

[9] Aijun Du and Jinqiao Duan. Invariant manifold reduction for stochastic dynamical systems. *arXiv preprint math/0607366*, 2006.

[10] KJ Falconer. Differentiation of the limit mapping in a dynamical system. *Journal of the London Mathematical Society*, 2(2):356–372, 1983.

[11] 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*, 21:136, 2020.

[12] Damir Filipović. Invariant manifolds for weak solutions to stochastic equations. *Probability theory and related fields*, 118(3):323–341, 2000.

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[14] Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

[15] 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.

[16] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

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

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

[20] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[21] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[22] Hikaru Ibayashi and Masaaki Imaizumi. Exponential escape efficiency of SGD from sharp minima in non-stationary regime. *arXiv preprint arXiv:2111.04004*, 2021.

[23] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[24] Xianyan Jia, Shutao Song, Wei He, Yangzihao Wang, Haidong Rong, Feihu Zhou, Liqiang Xie, Zhenyu Guo, Yuanzhou Yang, Liwei Yu, et al. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. *arXiv preprint arXiv:1807.11205*, 2018.

[25] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[26] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[27] 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. In *International Conference on Learning Representations*, 2017.

[28] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[29] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018.

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

[31] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[32] Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, and Aleksander Madry. ffcv. https://github.com/libffcv/ffcv/, 2022. commit xxxxxxx.

[33] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_3.

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

[35] 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*, 33:14544–14555, 2020.

[36] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021.

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

[38] Tao Lin, Lingjing Kong, Sebastian Stich, and Martin Jaggi. Extrapolation for large-batch training in deep learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6094–6104. PMLR, 13–18 Jul 2020.

[39] Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. In *International Conference on Learning Representations*, 2020.

[40] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction, 2022.

[41] Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16805–16817. Curran Associates, Inc., 2021.

[42] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. *arXiv preprint arXiv:2205.10287*, 2022.

[43] Gideon Mann, Ryan T. McDonald, Mehryar Mohri, Nathan Silberman, and Dan Walker. Efficient large-scale distributed training of conditional maximum entropy models. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1231–1239. Curran Associates, Inc., 2009.

[44] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[45] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[46] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[47] Jose Javier Gonzalez Ortiz, Jonathan Frankle, Mike Rabbat, Ari Morcos, and Nicolas Ballas. Trade-offs of Local SGD at scale: An empirical study. *arXiv preprint arXiv:2110.08133*, 2021.

[48] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455*, 2014.

[49] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[50] Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 693–701, 2011.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[52] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors, *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 1058–1062. ISCA, 2014. URL http://www.isca-speech.org/archive/interspeech_2014/i14_1058.html.

[53] Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

[54] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015.

[55] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9058–9067. PMLR, 13–18 Jul 2020.

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

[57] Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

[58] Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1488–1492. ISCA, 2015.

[59] Hang Su and Haoyu Chen. Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*, 2015.

[60] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 978-0-262-19398-6.

[61] Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd. *Proceedings of Machine Learning and Systems*, 1:212–229, 2019.

[62] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *Journal of Machine Learning Research*, 22(213): 1–50, 2021.

[63] Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[64] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

[65] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. Imagenet training in minutes. In *Proceedings of the 47th International Conference on Parallel Processing*, pages 1–10, 2018.

[66] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020.

[67] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[68] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 215–219, 2014. doi: 10.1109/ICASSP.2014.6853589.

[69] Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3219–3227. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/447. URL https://doi.org/10.24963/ijcai.2018/447.

[70] 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.

[71] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Contents

Appendix A. Additional Related Works

Optimization Aspect of Local SGD. Local SGD is a communication-efficient variant of parallel SGD, where multiple workers perform SGD independently and average the model parameters periodically. Dating back to Mann et al. [43] and Zinkevich et al. [71], this strategy has been widely adopted to reduce the communication cost and speed up training in both scenarios of data center distributed training [6, 48, 59, 68] and Federated Learning [26, 44]. To further accelerate training, Wang and Joshi [61] and Haddadpour et al. [16] proposed adaptive schemes for the averaging frequency, and Basu et al. [1] combined Local SGD with gradient compression. Motivated to theoretically understand the empirical success of Local SGD, a lot of researchers [14, 28, 57, 62, 67] analyzed the convergence rate of Local SGD under various settings, e.g., convex/non-convex objective functions. The error bound of Local SGD obtained by these works is typically inferior to that of SGD for fixed number of iterations/epochs and becomes worse as the number of local steps increases, revealing a trade-off between less communication and better optimization. In this paper, we are interested in the generalization aspect of Local SGD, assuming the training loss can be optimized to a small value.

Gradient noise and generalization. The effect of stochastic gradient noise on generalization has been studied from different aspects, e.g., changing the order of learning different patterns Li et al. [34], inducing an implicit regularizer in the second-order SDE approximation Li et al. [34], Smith et al. [56]. Our work follows a line of works studying the effect of noise in the lens of sharpness, which is long believed to be related to generalization Hochreiter and Schmidhuber [19], Neyshabur et al. [46]. Keskar et al. [27] empirically observed that large-batch training leads to worse generalization and sharper minima than small-batch training. Hu et al. [21], Ma and Ying [41], Wu et al. [63] showed that gradient noise destabilizes the training around sharp minima, and Ibayashi and Imaizumi [22], Kleinberg et al. [29], Xie et al. [64], Zhu et al. [70] quantitatively characterized how SGD escapes sharp minima. The most related papers are Blanc et al. [3], Damian et al. [7], Li et al. [37], which focus on the training dynamics near a manifold of minima and study the effect of noise on sharpness (see also Section 3.2). Though the mathematical definition of sharpness may be vulnerable to the various symmetries in deep neural nets [8], sharpness still appears to be one of the most promising tools for predicting generalization [13, 25].

Improving generalization in large-batch training. The generalization issue of the large-batch (or full-batch) training has been observed as early as [2, 33]. As mentioned in Section 1, the generalization issue of large-batch training could be due to the lack of a sufficient amount of stochastic noise. To make up the noise in large-batch training, Goyal et al. [15], Krizhevsky [30] empirically discovered the *Linear Scaling Rule* for SGD, which suggests enlarging the learning rate proportionally to the batch size. Jastrzębski et al. [23] adopted an SDE-based analysis to justify that this scaling rule indeed retains the same amount of noise as small-batch training (see also Section 3.1). However, the SDE approximation may fail if the learning rate is too large [36], especially in the early phase of training before the first learning rate decay [55]. Shallue et al. [53] demonstrated that generalization gap between small- and large-batch training can also depend on many other training hyperparameters. Besides enlarging learning rate, other approaches have also been proposed to reduce the gap, including training longer [20], learning rate warmup [15], LARS [65], LAMB [66]. In this paper, we focus on using Local SGD to improve generalization, but adding local steps is a generic training trick that can also be combined with others, e.g., Local LARS [39], Local Extrap-SGD [38].

Appendix B. Supplementary for Section 2

B.1. The Debate on Local SGD

In this section, we summarize a debate in the literature regarding *when* to switch the training mode in Post-local SGD. As Post-local SGD can be viewed as running Local SGD from an SGD-pretrained model, the discussion around the time point for switching can reveal some information about the conditions for Local SGD to generalize better.

Local SGD generalizes better than SGD on CIFAR-10. Lin et al. [39] empirically observed that Post-local SGD exhibits a better generalization performance than SGD. Most of their experiments are conducted on CIFAR-10 and CIFAR-100 with multiple learning rate decay, and the algorithm switches from (parallel) SGD to Local SGD right after the first learning rate decay. We refer to this particular choice of the switching time point as the *first-decay switching strategy* for short. To justify this strategy, they empirically showed that the generalization improvement can be less significant if starting Local SGD from the beginning or right after the second learning rate decay. It has also been observed by Wang and Joshi [62] that running Local SGD from the beginning improves generalization, but the test accuracy improvement may not be large enough. A subsequent work by Lin et al. [38] showed that adding local steps to Extrap-SGD, a variant of SGD proposed therein, after the first learning rate decay also improves generalization, suggesting that the first-decay switching strategy can also be applied to the post-local variant of other optimizers.

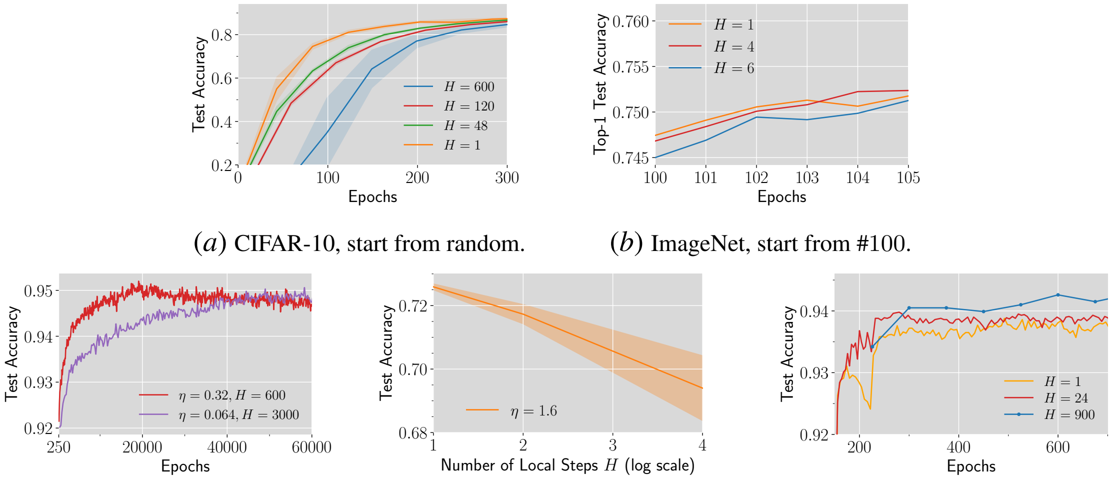
Does Local SGD exhibit the same generalization benefit on large-scale datasets? Going beyond CIFAR-10, Lin et al. [39] conducted a few ImageNet experiments and showed that Post-local SGD with first-decay switching strategy still leads to better generalization than SGD. However, the improvement is sometimes marginal, e.g., 0.1% for batch size 8192. For the general case, Lin et al. [39] suggested that the time of switching should be tuned aiming at "capturing the time when trajectory starts to get into the influence basin of a local minimum" in a footnote, but no further discussion or experiments are provided to justify this guideline. Ortiz et al. [47] conducted a more extensive evaluation on ImageNet (with a different set of hyperparameters) and concluded with the opposite: the first-decay switching strategy can hurt the validation accuracy. Instead, switching at a later time, such as the second learning rate decay, leads to a better validation accuracy than SGD.[1] To explain this phenomenon, they conjecture that switching to Local SGD has a regularization effect that is beneficial only in the short-term, so it is always better to switch as late as possible. They further conjecture that this discrepancy between CIFAR-10 and ImageNet is mainly due to the task scale. On TinyImageNet, which is a spatially downscaled subset of ImageNet, the first-decay switching strategy indeed leads to better validation accuracy.

B.2. Additional Experimental Results for Section 2

In Figure 3, we present additional experimental results for Section 2 to further verify our finding. Specifically, in Figure 3, (a) and (b) are enlarged views for Figure 2 (a) and (c) respectively, showing that Local SGD can generalize worse than SGD in the first few epochs. (c) shows that Local SGD can still chieve comparable test accuracy when we use a much smaller learning rate (e.g., $\eta = 0.064$) on the condition that H and the training budget are set properly. (d) presents the case where, with a large learning rate, the generalization improvement of Local SGD disappears even starting from

[1] This generalization improvement is not mentioned explicitly in [47] but can be clearly seen from Figures 7 and 8 in their paper.

a pre-trained model. In (e), the generalization benefit of Local SGD with $H = 24$ becomes less significant after the learning rate decay at epoch 226, which is consistent with the observation by Ortiz et al. [47] that the generalization benefit of Local SGD usually disappears after the learning rate decay. But we can preserve the improvement by increasing H to 900. Here, we use Local SGD with momentum. We refer the readers to Appendix L.2 for training details.



(a) CIFAR-10, start from random.

(b) ImageNet, start from #100.

(c) CIFAR-10, start from #250.

(d) ImageNet, start from #100.

(e) CIFAR-10, start from #150.

Figure 3: Additional experimental results about the effect of learning rate, training time and the number of local steps.

B.3. Reconciling previous works

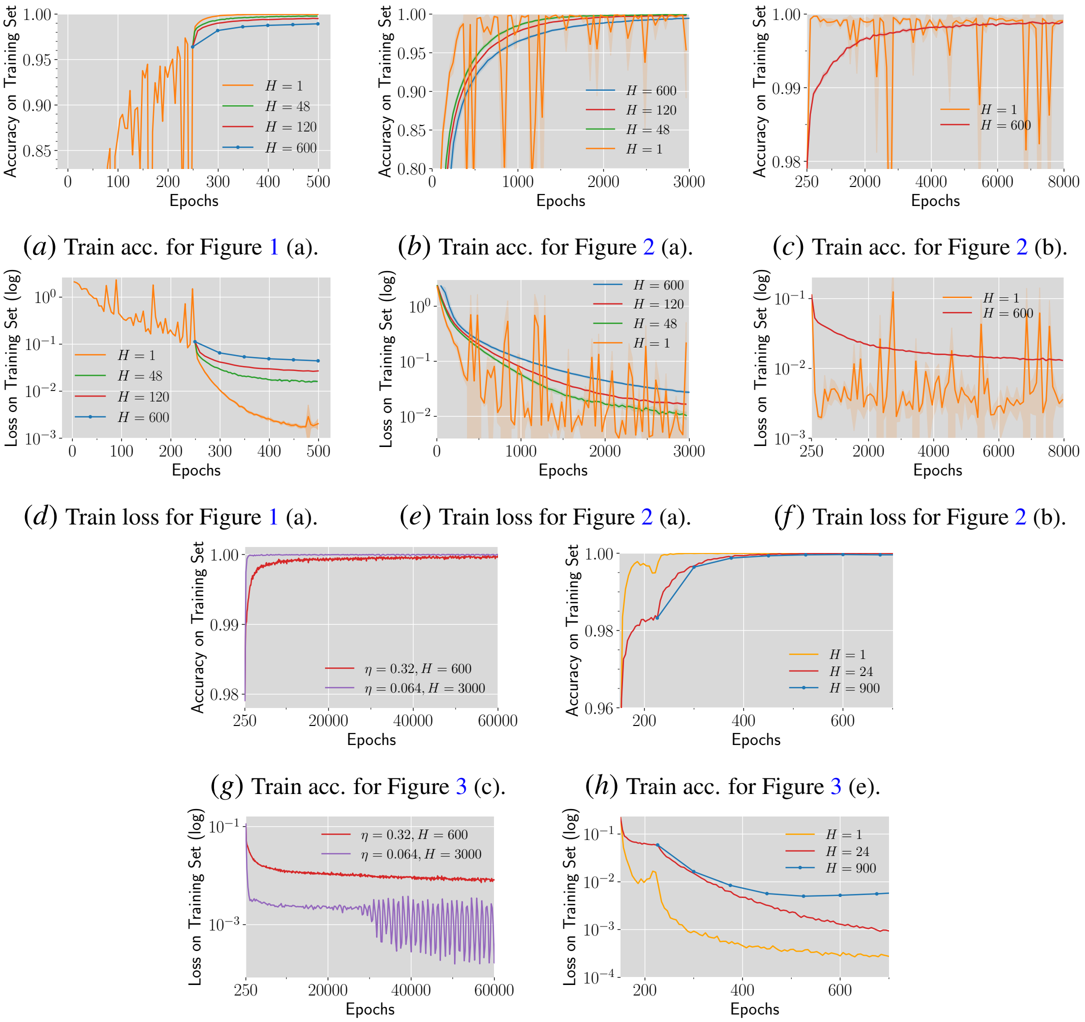
Our finding can help to settle the debate presented in Appendix B.1 to a large extent. Simultaneously requiring a small learning rate and sufficient training time poses a trade-off when learning rate decay is used with a limited training budget: switching to Local SGD earlier may lead to a large learning rate, while switching later may result in insufficient training time. It is thus unsurprising that first-decay switching strategy is not always the best when the dataset and learning rate schedule change.

The need for sufficient training time does not contradict with Ortiz et al. [47]'s conjecture that Local SGD only has a "short-term" generalization benefit. In their experiments, the generalization improvement usually disappears right after the next learning rate decay (instead of after a fixed amount of time). We suspect that the real reason why the improvement vanishes is that the number of local steps H was kept as a constant, but our finding suggests to tune H after η changes. In Figure 3 (e), we reproduce this phenomenon and show that increasing H after learning rate decay retains the improvement.

Appendix C. Accuracy and Loss on the Training Set

This section visualizes the accuracy and loss on the training set for CIFAR-10 experiments in Figures 1, 2 and 3. The accuracy and loss on the training set are computed on the averaged parameters.

In each evaluation, we first randomly sample 100 batches of training data (batch size = 128) without replacement and go through them to estimate the running mean and variance. Then we compute accuracy and loss on the whole training set. We omit the plots for training accuracy and loss for ImageNet experiments since it is computationally expensive to go through the whole training set of ImageNet.



(a) Train acc. for Figure 1 (a).

(b) Train acc. for Figure 2 (a).

(c) Train acc. for Figure 2 (b).

(d) Train loss for Figure 1 (a).

(e) Train loss for Figure 2 (a).

(f) Train loss for Figure 2 (b).

(g) Train acc. for Figure 3 (c).

(h) Train acc. for Figure 3 (e).

(i) Train loss for Figure 3 (c).

(j) Train loss for Figure 3 (e).

Figure 4: Accuracy and loss on the training set for CIFAR-10 experiments.

Appendix D. Modeling Local SGD with Multiple Conventional SDEs

Several previous works adopt the conventional SDE approximation (3) and connect good generalization to having a large diffusion term $\sqrt{\frac{\eta}{B}}\boldsymbol{\Sigma}^{1/2}\mathrm{d}\boldsymbol{W}_t$ in the SDE [23, 55], because a suitable amount of noise can be necessary for large-batch training to generalize well (see also Appendix A).

Lin et al. [39] tried to informally explain the success of Local SGD by adopting this argument. Basically, they attempted to write multiple SDEs, each of which describes the H-step local training process of each worker in each round (from $\boldsymbol{\theta}_{k,0}^{(s)}$ to $\boldsymbol{\theta}_{k,H}^{(s)}$). The key difference between each of these SDEs and the SDE for SGD (3) is that the former one has a larger diffusion term because the workers use batch size B_{loc} instead of B:

$$\mathrm{d}\boldsymbol{X}(t) = -\nabla\mathcal{L}(\boldsymbol{X})\mathrm{d}t + \sqrt{\frac{\eta}{B_{\mathrm{loc}}}}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X})\mathrm{d}\boldsymbol{W}_t. \tag{10}$$

Lin et al. [39] then argue that the total amount of "noise" in the training dynamics of Local SGD is larger than that of SGD. However, it is hard to see whether it is indeed larger, since the model averaging step at the end of each round can reduce the variance in training and may cancel the effect of having larger diffusion terms.

More formally, a complete modeling of Local SGD following this idea should view the sequence of global iterates $\{\bar{\boldsymbol{\theta}}^{(s)}\}$ as a Markov process $\{\boldsymbol{X}^{(s)}\}$. Let $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}, B, t)$ the distribution of $\boldsymbol{X}(t)$ in (3) with initial condition $\boldsymbol{X}(0) = \boldsymbol{x}$. Then the Markov transition should be $\boldsymbol{X}^{(s+1)} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{X}_{k,H}^{(s)}$, where $\boldsymbol{X}_{1,H}^{(s)}, \dots, \boldsymbol{X}_{K,H}^{(s)}$ are K independent samples from $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B_{\mathrm{loc}}, H\eta)$, i.e., sampling from (10).

Consider one round of model averaging. It is true that $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B_{\mathrm{loc}}, H\eta)$ may have a larger variance than the corresponding SGD baseline $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B, H\eta)$ because the former one has a smaller batch size. However, it is unclear whether $\boldsymbol{X}^{(s+1)}$ also has a larger variance than $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B, H\eta)$. This is because $\boldsymbol{X}^{(s+1)}$ is the average of K samples, which means we have to compare $\frac{1}{K}$ times the variance of $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B_{\mathrm{loc}}, H\eta)$ with the variance of $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B, H\eta)$. Then it is unclear which one is larger.

In the special case where $H\eta$ is small, $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{X}^{(s)}, B_{\mathrm{loc}}, H\eta)$ is approximately equal to the following Gaussian distribution:

$$\mathcal{N}\left(\boldsymbol{X}^{(s)} - \eta H\nabla\mathcal{L}(\boldsymbol{X}^{(s)}), \frac{\eta^2 H}{B_{\mathrm{loc}}}\boldsymbol{\Sigma}(\boldsymbol{X}^{(s)})\right) \tag{11}$$

Then averaging over K samples gives

$$\mathcal{N}\left(\boldsymbol{X}^{(s)} - \eta H\nabla\mathcal{L}(\boldsymbol{X}^{(s)}), \frac{\eta^2 H}{B}\boldsymbol{\Sigma}(\boldsymbol{X}^{(s)})\right), \tag{12}$$

which is exactly the same as the Gaussian approximation of the SGD baseline. This means there do exist certain cases where Lin et al. [39]'s argument does not give a good separation between Local SGD and SGD.

Moreover, we do not gain any further insights from this modeling since it is hard to see how model averaging interacts with the SDEs.

Appendix E. The Slow SDE for Training with Label Noise Regularization

In this section, we study the case of training over-parameterized neural nets with label noise regularization to exemplify the generalization benefit of having a larger drift term [3, 7, 37].

For a C-class classification task, the label noise regularization is as follows: every time we draw a sample from the training set, we make the true label as it is with probability $1 - p$, and replace it with any other label with equal probability $\frac{p}{C-1}$. When we use cross-entropy loss, the Slow SDE for

SGD turns out to be a simple deterministic gradient flow on Γ (instead of a semi-gradient method) for minimizing the trace of Hessian:

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{1}{4B}\nabla_{\Gamma}\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta}))\mathrm{d}t, \tag{13}$$

where $\nabla_{\Gamma} f$ stands for the gradient of the function f projected to the tangent space of Γ. Checking the validity of Hypothesis 3.1 reduces to the following question: *Is minimizing the trace of Hessian beneficial to generalization?* Many previous works provide positive answers, including the line of works we just mentioned. Blanc et al. [3] and Li et al. [37] connect minimizing the trace of Hessian to finding sparse or low-rank solutions for training two-layer linear nets. Damian et al. [7] empirically showed that good generalization correlates with a smaller trace of Hessian in training ResNets with label noise. Besides, Ma and Ying [41] connects the trace of Hessian to the smoothness of the function represented by a deep neural net. We defer all the proofs to Appendix K.

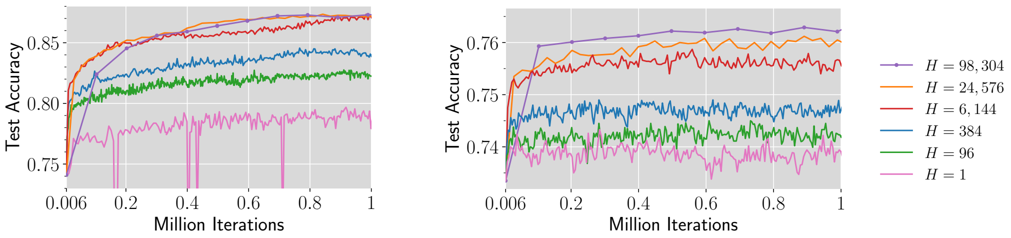
As for Local SGD, the Slow SDE can be simplified as:

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{1}{4B}\nabla_{\Gamma}\left(\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) + (K-1)\cdot\frac{\mathrm{tr}(F(2H\eta\nabla^2\mathcal{L}(\boldsymbol{\zeta})))}{2H\eta}\right)\mathrm{d}t, \tag{14}$$

where $F(x) := \int_0^x \psi(y)\mathrm{d}y$ and is interpreted as a matrix function in (14). Note that the magnitude of the RHS in (14) becomes larger as H increases. When H gets large enough, the RHS in (14) is approximately K times of the RHS in (13):

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{K}{4B}\nabla_{\Gamma}\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta}))\mathrm{d}t. \tag{15}$$

Comparing (13) and (15), we can conclude that Local SGD accelerates the process of sharpness reduction, thus leading to better generalization. Furthermore, the regularization effect gets stronger for larger H. We also conduct experiments on non-augmented CIFAR-10 with label noise regularization to verify our conclusion. As shown in Figure 5, adding local steps indeed gives better generalization performance.



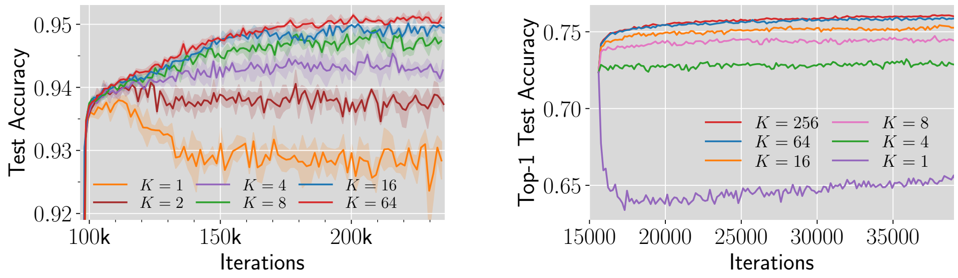
(*a*) ResNet-56 + GroupNorm.

(*b*) VGG-16 w/o normalization.

Figure 5: Local SGD with label noise regularization on CIFAR-10 without data augmentation. A larger number of local steps indeed enables higher test accuracy. For both architectures, we replace ReLU with Swish. See Appendix L.4 for training details.

Appendix F. Experimental Results on Reducing the Diffusion Term

We conduct experiments on CIFAR-10 with varying K and fixed η, H to justify that reducing the diffusion term is beneficial to generalization. As shown in Figure 6, we can achieve higher test accuracy for larger K where the diffusion term is smaller.



(*a*) CIFAR-10, $H = 600$ for $K > 1$.

(*b*) ImageNet, $H = 78$ for $K > 1$.

Figure 6: A smaller diffusion term leads to better generalization. Test accuracy improves as we increase K with fixed η and H to reduce the diffusion term while keeping the generalization part untouched. See Appendix L.3 for training details.

Appendix G. Discussion

Connection to the conventional wisdom that the diffusion term matters more. As mentioned in Section 3.1, it is believed in the literature is that a large diffusion term in the conventional SDE leads to good generalization. One may think that the diffusion term in the Slow SDE corresponds to that in the conventional SDE, and thus enlarging the diffusion term rather than the drift term should lead to better generalization. However, we note that both the diffusion and drift terms in the Slow SDEs are associated with the diffusion term in the conventional SDE (Slow SDEs become stationary if $\boldsymbol{\Sigma} = \mathbf{0}$). Our view identifies the effect of each different component of the noise covariance on generalization, and therefore, goes one step further on the conventional wisdom.

Can Local SGD close the generalization gap between small- and large-batch training? We remark that the mechanism causing the generalization gap is different from the mechanism of Local SGD for improving generalization, so it is unclear what a general claim can be made. As noted by Smith et al. [55], the generalization gap occurs early in the training when the learning rate is large and SDE does not give a good approximation. In contrast, as shown in our paper, Local SGD yields generalization benefits mainly for small learning rate in later phases of training.

Appendix H. Local SGD stays close to SGD for $\mathcal{O}(\eta^{-1})$ steps

The following theorem states that Local SGD and SGD closely track each other for $\mathcal{O}(\eta^{-1})$ steps. It suggests that this time horizon is too short for the difference between the two algorithm to appear, necessitating an analysis that lasts for a longer time.

Theorem 9 *Assume that the loss function $\mathcal{L}$ is $\mathcal{C}^3$-smooth with bounded second and third order derivatives and that $\nabla \ell(\boldsymbol{\theta}; \xi)$ is bounded. Let $T > 0$ be a constant, $\bar{\boldsymbol{\theta}}^{(s)}$ be the s-th global iterate of Local SGD and $\boldsymbol{w}_t$ be the t-th iterate of SGD with the same initialization $\boldsymbol{w}_0 = \bar{\boldsymbol{\theta}}^{(0)}$ and same $\eta, B_{\mathrm{loc}}, K$. Then for any $H \leq \frac{T}{\eta}$ and $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, it holds with probability at least $1 - \delta$ that for all $s \leq \frac{T}{\eta H}$, $\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{w}_{sH}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta\delta}})$.*

H.1. Proof of Theorem 9

This subsection presents the proof for Theorem 9. First, we define some notations that will be used throughout this section. For the sequence of Local SGD iterates $\{\boldsymbol{\theta}_{k,t}^{(s)} : k \in [K], 0 \leq t \leq H, s \geq 0\}$, we introduce an auxiliary sequence $\{\hat{\boldsymbol{u}}_t\}_{t\in\mathbb{N}}$ which starts from $\bar{\boldsymbol{\theta}}^{(0)}$ and follows. Define $\hat{\boldsymbol{\Delta}}_{k,t}^{(s)} := \boldsymbol{\theta}_{k,t}^{(s)} - \hat{\boldsymbol{u}}_t^{(s)}$ and $\bar{\boldsymbol{\Delta}}^{(s)} := \bar{\boldsymbol{\theta}}^{(s)} - \hat{\boldsymbol{u}}_0^{(s)}$, which stand for the difference between the Local SGD iterate and GD iterate. For convenience, let $\hat{\boldsymbol{u}}_t^{(s)} = \hat{\boldsymbol{u}}_{sH+t}$ and $\boldsymbol{z}_{k,sH+t} = \boldsymbol{z}_{k,t}^{(s)}$, which will be used interchangeably. Assume that $\mathcal{L}$ is $\mathcal{C}^3$-smooth with bounded second and third order derivatives. Let $\nu_2 := \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\nabla^2\mathcal{L}(\boldsymbol{\theta})\|_2$ and $\nu_3 := \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\nabla^3\mathcal{L}(\boldsymbol{\theta})\|_2$. Since $\nabla\ell(\boldsymbol{\theta};\boldsymbol{\zeta})$ is bounded, the gradient noise $\boldsymbol{z}_{k,t}^{(s)}$ is also bounded and we denote by $\sigma_{\max}$ the upper bound such that $\|\boldsymbol{z}_{k,t}^{(s)}\|_2 \leq \sigma_{\max}, \forall s, k, t$.

For each client, define the following sequence $\{\hat{\boldsymbol{Z}}_{k,t} : t \geq 0\}$:

$$\hat{\boldsymbol{Z}}_{k,t} = \sum_{\tau=0}^{t-1}\left[\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_l))\right]\boldsymbol{z}_{k,\tau}, \qquad \hat{\boldsymbol{Z}}_{k,0} = \boldsymbol{0}, \forall k \in [K].$$

To prove Theorem 9, we will show that both Local SGD iterates $\bar{\boldsymbol{\theta}}^{(s)}$ and SGD iterates $\boldsymbol{w}_{sH}$ will closely track GD iterates $\hat{\boldsymbol{u}}_{sH}$ with high probability. The following lemma establishes the concentration property of $\hat{\boldsymbol{Z}}_{k,t}$.

Lemma 10 (Concentration property of $\{\hat{\boldsymbol{Z}}_{k,t}\}$) *With probability at least $1-\delta$, the following holds simultaneously for all $k \in [K]$, $0 \leq t < \lfloor \frac{T}{\eta} \rfloor$:*

$$\|\hat{\boldsymbol{Z}}_{k,t}\|_2 \leq \hat{C}_1\sigma_{\max}\sqrt{\frac{2T}{\eta}\log\frac{2TK}{\delta\eta}},$$

where $\hat{C}_1 := \exp(T\nu_2)$.

Proof For each $\hat{\boldsymbol{Z}}_{k,t}$, construct a sequence $\{\hat{\boldsymbol{Z}}_{k,t,t'}\}_{t'=0}^{t}$:

$$\hat{\boldsymbol{Z}}_{k,t,t'} := \sum_{\tau=0}^{t'-1}\left(\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_l))\right)\boldsymbol{z}_{k,\tau}^{(s)}, \qquad \tilde{\boldsymbol{Z}}_{k,t,0}^{(s)} = \boldsymbol{0}.$$

Since $\|\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_l)\|_2 \leq \nu_2$ for all $l \geq 0$, the following holds for all $0 \leq \tau < t-1$ and $0 < t < \lfloor \frac{T}{\eta} \rfloor$:

$$\|\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_l))\|_2 \leq (1+\rho_2\eta)^t \leq \exp(T\nu_2) = \hat{C}_1.$$

Notice that $\{\hat{\boldsymbol{Z}}_{k,t,t'}\}_{t'=0}^{t}$ is a martingale with $\|\hat{\boldsymbol{Z}}_{k,t,t'} - \hat{\boldsymbol{Z}}_{k,t,t'-1}\|_2 \leq \hat{C}_1\sigma_{\max}$. Since $\hat{\boldsymbol{Z}}_{k,t} = \hat{\boldsymbol{Z}}_{k,t,t}$, by Azuma-Hoeffding's inequality,

$$\mathbb{P}(\|\hat{\boldsymbol{Z}}_{k,t}\|_2 \geq \epsilon') \leq 2\exp\left(\frac{-\epsilon'^2}{2t\left(\hat{C}_1\sigma_{\max}\right)^2}\right).$$

Taking union bound on all $k \in [K]$ and $0 \le t \le \lfloor \frac{T}{\eta} \rfloor$, we can conclude that with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{Z}}_{k,t}\|_2 \le \hat{C}_1 \sigma_{\max} \sqrt{\frac{2T}{\eta} \log \frac{2TK}{\delta\eta}}, \qquad \forall 0 \le t < \lfloor \frac{T}{\eta} \rfloor, k \in [K].$$

■

The following lemma states that, with high probability, Local SGD iterates $\boldsymbol{\theta}_{k,t}^{(s)}$ and $\bar{\boldsymbol{\theta}}^{(s)}$ closely track the gradient descent iterates $\hat{\boldsymbol{u}}_{sH}$ for $\lfloor \frac{T}{H\eta} \rfloor$ rounds.

Lemma 11 *For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, the following inequalities hold with probability at least $1-\delta$:*

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \hat{\boldsymbol{u}}_{sH+t}\|_2 \le \hat{C}_3 \sqrt{\eta \log \frac{1}{\eta\delta}}, \qquad \forall k \in [K], 0 \le s < \lfloor \frac{T}{H\eta} \rfloor, 0 \le t \le H,$$

and

$$\|\bar{\boldsymbol{\theta}}^{(s)} - \hat{\boldsymbol{u}}_{sH}\|_2 \le \hat{C}_3 \sqrt{\eta \log \frac{1}{\eta\delta}}, \qquad \forall 0 \le s \le \lfloor \frac{T}{H\eta} \rfloor,$$

where $\hat{C}_3$ is a constant independent of η and H.

Proof According to the update rule for $\boldsymbol{\theta}_{k,t}^{(s)}$ and $\hat{\boldsymbol{u}}_t^{(s)}$,

$$\boldsymbol{\theta}_{k,t+1}^{(s)} = \boldsymbol{\theta}_{k,t}^{(s)} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \eta \boldsymbol{z}_{k,t}^{(s)} \tag{16}$$

$$\hat{\boldsymbol{u}}_{t+1}^{(s)} = \hat{\boldsymbol{u}}_t^{(s)} - \eta \nabla \mathcal{L}(\hat{\boldsymbol{u}}_t^{(s)}). \tag{17}$$

Subtracting (16) from (17) gives

$$\begin{aligned} \hat{\boldsymbol{\Delta}}_{k,t+1}^{(s)} &= \hat{\boldsymbol{\Delta}}_{k,t}^{(s)} - \eta(\nabla \mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \nabla \mathcal{L}(\hat{\boldsymbol{u}}_t^{(s)})) - \eta \boldsymbol{z}_{k,t}^{(s)} \\ &= (\boldsymbol{I} - \eta \nabla^2 \mathcal{L}(\hat{\boldsymbol{u}}_t^{(s)}))\hat{\boldsymbol{\Delta}}_{k,t}^{(s)} - \eta \boldsymbol{z}_{k,t}^{(s)} + \eta \hat{\boldsymbol{v}}_{k,t}^{(s)}, \end{aligned} \tag{18}$$

where $\|\hat{\boldsymbol{v}}_{k,t}^{(s)}\|_2 \le \frac{\rho_3}{2} \|\hat{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2^2$. Applying (18) t times, we have

$$\begin{aligned} \hat{\boldsymbol{\Delta}}_{k,t}^{(s)} = &\left[\prod_{\tau=0}^{t-1} (\boldsymbol{I} - \eta \nabla^2 \mathcal{L}(\hat{\boldsymbol{u}}_\tau^{(s)}))\right] \hat{\boldsymbol{\Delta}}_{k,0}^{(s)} - \eta \underbrace{\sum_{\tau=0}^{t-1} \left[\prod_{l=\tau+1}^{t-1} (\boldsymbol{I} - \eta \nabla^2 \mathcal{L}(\hat{\boldsymbol{u}}_l^{(s)}))\right] \boldsymbol{z}_{k,\tau}^{(s)}}_{\mathcal{T}} \\ &+ \eta \sum_{\tau=0}^{t-1} \prod_{l=\tau+1}^{t-1} (\boldsymbol{I} - \eta \nabla^2 \mathcal{L}(\hat{\boldsymbol{u}}_l^{(s)}))\hat{\boldsymbol{v}}_{k,\tau}^{(s)}. \end{aligned} \tag{19}$$

Here, $\mathcal{T}$ can expressed in the following form:

$$\mathcal{T} = \hat{\boldsymbol{Z}}_{k,sH+t} - \left[\prod_{l=sH+1}^{sH+t-1} (\boldsymbol{I} - \eta \nabla^2 \mathcal{L}(\hat{\boldsymbol{u}}_l))\right] \hat{\boldsymbol{Z}}_{k,sH}.$$

Substituting in $t = H$ and taking the average, we derive the following recursion:

$$\begin{aligned}
\bar{\boldsymbol{\Delta}}^{(s+1)} &= \frac{1}{K}\sum_{k\in[K]} \hat{\boldsymbol{\Delta}}^{(s)}_{k,H} \\
&= \left[\prod_{\tau=0}^{H-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}^{(s)}_{\tau}))\right]\bar{\boldsymbol{\Delta}}^{(s)} - \frac{\eta}{K}\sum_{k\in[K]}\sum_{\tau=0}^{H-1}\left[\prod_{l=\tau+1}^{H-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}^{(s)}_{l}))\right]\boldsymbol{z}^{(s)}_{k,\tau} \\
&\quad + \frac{\eta}{K}\sum_{k\in[K]}\sum_{\tau=0}^{H-1}\prod_{l=\tau+1}^{H-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}^{(s)}_{l}))\hat{\boldsymbol{v}}^{(s)}_{k,\tau}. \qquad (20)
\end{aligned}$$

Applying (20) s times yields

$$\bar{\boldsymbol{\Delta}}^{(s)} = -\frac{\eta}{K}\sum_{k\in[K]}\hat{\boldsymbol{Z}}_{k,sH} + \frac{\eta}{K}\sum_{r=0}^{s-1}\sum_{\tau=0}^{H-1}\sum_{k\in[K]}\left[\prod_{l=rH+\tau+1}^{sH}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_{l}))\right]\hat{\boldsymbol{v}}^{(r)}_{k,\tau}. \qquad (21)$$

Substitute (21) into (19) and we have

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}^{(s)}_{k,t} = &-\frac{\eta}{K}\sum_{k'\in[K]}\hat{\boldsymbol{Z}}_{k',sH} - \eta\hat{\boldsymbol{Z}}_{k,sH+t} + \eta\left[\prod_{l=sH+1}^{sH+t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_{l}))\right]\hat{\boldsymbol{Z}}_{k,sH} \\
&+ \frac{\eta}{K}\sum_{r=0}^{s-1}\sum_{\tau=0}^{H-1}\sum_{k'\in[K]}\left[\prod_{l=rH+\tau+1}^{sH+t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_{l}))\right]\hat{\boldsymbol{v}}^{(r)}_{k',\tau} \\
&+ \eta\sum_{\tau=0}^{t-1}\left[\prod_{l=sH+\tau+1}^{sH+t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\hat{\boldsymbol{u}}_{l}))\right]\hat{\boldsymbol{v}}^{(s)}_{k,\tau}.
\end{aligned}$$

By Cauchy-Schwartz inequality and triangle inequality, we have

$$\begin{aligned}
\|\hat{\boldsymbol{\Delta}}^{(s)}_{k,t}\|_2 \leq \frac{\eta}{K}&\left(\sum_{k'\in[K]}\|\hat{\boldsymbol{Z}}_{k',sH}\|_2\right) + \eta\|\hat{\boldsymbol{Z}}_{k,sH+t}\|_2 + \eta\hat{C}_1\|\hat{\boldsymbol{Z}}_{k,sH}\|_2 \\
&+ \frac{\eta\hat{C}_1\nu_3}{2K}\sum_{r=0}^{s-1}\sum_{\tau=0}^{H-1}\sum_{k'\in[K]}\|\hat{\boldsymbol{\Delta}}^{(r)}_{k',\tau}\|_2^2 + \frac{\eta\hat{C}_1\nu_3}{2}\sum_{\tau=0}^{t-1}\|\hat{\boldsymbol{\Delta}}^{(r)}_{k,\tau}\|_2^2,
\end{aligned} \qquad (22)$$

where $\hat{C}_1 = \exp(\nu_2 T)$. Below we prove by induction that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, if

$$\|\hat{\boldsymbol{Z}}_{k,t}\|_2 \leq \hat{C}_1\sigma_{\max}\sqrt{\frac{2T}{\eta}\log\frac{2TK}{\eta\delta}}, \quad \forall 0 \leq t < \lfloor\frac{T}{\eta}\rfloor, k \in [K], \qquad (23)$$

then there exists a constant $\hat{C}_2$ such that for all $k \in [K], 0 \leq s < \lfloor\frac{T}{\eta H}\rfloor$ and $0 \leq t \leq H$,

$$\|\hat{\boldsymbol{\Delta}}^{(s)}_{k,t}\|_2 \leq \hat{C}_2\sqrt{\eta\log\frac{2TK}{\eta\delta}}. \qquad (24)$$

First, for all $k \in [K]$, $\|\hat{\mathbf{\Delta}}^{(0)}_{k,0}\|_2 = 0$ and hence (24) holds. Assuming that (24) holds for all $\hat{\mathbf{\Delta}}^{(r)}_{k',\tau}$ where $k' \in [K], 0 \leq r < s$, $0 \leq \tau \leq H$ and $r = s$, $0 \leq \tau < t$, then by (22), then for all $k \in [K]$, the following holds:

$$\|\hat{\mathbf{\Delta}}^{(s)}_{k,t}\|_2 \leq 3\tilde{C}_1^2 \sigma_{\max} \sqrt{2T\eta \log \frac{2TK}{\eta\delta}} + \tilde{C}_1 \hat{C}_2^2 T \eta \nu_3 \log \frac{2TK}{\eta\delta}.$$

Let $\hat{C}_2 \geq 6\tilde{C}_1^2 \sigma_{\max}\sqrt{2T}$. Then for sufficiently small η, (24) holds. By Theorem 10, (23) holds with probability at least $1-\delta$. Furthermore, notice that $\bar{\boldsymbol{\theta}}^{(s)} - \hat{\boldsymbol{u}}_{sH} = \frac{1}{K}\sum_{k\in[K]} \hat{\mathbf{\Delta}}^{(s-1)}_{k,H}$. Hence we have the lemma. ■

The iterates of standard SGD can be viewed as the local iterates on a single client with the number of local steps $\lfloor \frac{T}{\eta} \rfloor$. Therefore, we can directly apply Theorem 11 and obtain the following lemma about the SGD iterates $\boldsymbol{w}_t$.

Corollary 12 *For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, the following holds with probability at least $1-\delta$:*

$$\|\boldsymbol{w}_{sH} - \hat{\boldsymbol{u}}_{sH}\|_2 \leq \hat{C}_3 \sqrt{\eta \log \frac{1}{\eta\delta}}, \qquad \forall 0 \leq s \leq \frac{T}{H\eta},$$

where $\hat{C}_3$ is the same constant as in Theorem 11.

Applying Theorem 11 and Theorem 12 and taking union bound, we have Theorem 9.

Appendix I. Proof Outline of Main Theorems

In this section, we provide the proof outline of Theorem 4. The proof details are deferred to Appendix J. We first introduce additional notations that will be used throughout Appendix I and Appendix J.

I.1. Additional Notations and Definitions

We first introduce the notion of μ-PL.

Definition 13 (Polyak-Łojasiewicz Condition) *For $\mu > 0$, we say a function $\mathcal{L}(\cdot)$ satisfies μ-Polyak-Łojasiewicz condition (abbreviated as μ-PL) on set U if*

$$\frac{1}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \geq \mu(\mathcal{L}(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in U} \mathcal{L}(\boldsymbol{\theta}')).$$

We will later show that there exists a neighborhood of Γ where $\mathcal{L}$ satisfies μ-PL.

We then introduce the definitions of the ϵ-ball at a point and the ϵ-neighborhood of a set. For $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\epsilon > 0$, $B^{\epsilon}(\boldsymbol{\theta}) := \{\boldsymbol{\theta}' \mid \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2 < \epsilon\}$ is the open ϵ-ball centered at $\boldsymbol{\theta}$. For a set $\mathcal{Z} \subseteq \mathbb{R}^d$, $\mathcal{M}^{\epsilon} := \bigcup_{\boldsymbol{\theta}\in\mathcal{M}} B^{\epsilon}(\boldsymbol{\theta})$ is the ϵ-neighborhood of $\mathcal{M}$.

I.2. Construction of working zones

We construct four nested working zones $(\Gamma^{\epsilon_0}, \Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3})$ in the neighborhood of Γ. Later we will show that the local iterates $\boldsymbol{\theta}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}$ and the global iterates $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$ with high probability after $\mathcal{O}(\log \frac{1}{\eta})$ rounds. The following lemma illustrates the properties the working zones should satisfy and the proof is deferred to Appendix J.4.

Lemma 14 (Working zone lemma) *There exists constants $\epsilon_0 < \epsilon_1 < \epsilon_2 < \epsilon_3$ such that $(\Gamma^{\epsilon_0}, \Gamma^{\epsilon_1}, \Gamma^{\epsilon_2}, \Gamma^{\epsilon_3})$ satisfy the following properties:*

1. *$\mathcal{L}$ satisfies μ-PL in Γ^{ϵ_3} for some $\mu > 0$.*
2. *Any gradient flow starting in Γ^{ϵ_2} converges to some point in Γ. Then, by [10], $\Phi(\cdot)$ is $\mathcal{C}^\infty$ in Γ^{ϵ_2}.*
3. *Any $\boldsymbol{\theta} \in \Gamma^{\epsilon_1}$ has an ϵ_1 neighborhood $B^{\epsilon_1}(\boldsymbol{\theta})$ such that $B^{\epsilon_1}(\boldsymbol{\theta}) \subseteq \Gamma^{\epsilon_2}$.*
4. *Any gradient descent starting in Γ^{ϵ_0} with sufficiently small learning rate will stay in Γ^{ϵ_1}.*

I.3. Proof Outline

We are now in a position to state the proof outline. The general idea is to adopt the framework proposed by Li et al. [34] to bound the closeness of the manifold projection $\{\Phi(\bar{\boldsymbol{\theta}}^{(s)})\}_{s=0}^{\lfloor T/(H\eta^2)\rfloor}$ and the solution to SDE (4), $\{\boldsymbol{\zeta}(t) : t \in [0, T]\}$. A key component of this framework is to estimate the moments of change over a fixed time interval. To obtain the estimation of moments for $\boldsymbol{\zeta}(t)$, we can directly apply the results in Li et al. [34]. However, the estimation of the moments for $\Phi(\bar{\boldsymbol{\theta}}^{(s)})$ requires a careful analysis of the limiting dynamics of $\{\bar{\boldsymbol{\theta}}^{(s)}\}_{s=0}^{\lfloor T/(H\eta^2)\rfloor}$. The dynamics of $\{\bar{\boldsymbol{\theta}}^{(s)}\}_{s=0}^{\lfloor T/(H\eta^2)\rfloor}$ are divided into two phases: the approaching phase and the drift phase.

For the approaching phase (Phase 1), we show that after $\mathcal{O}(\log \frac{1}{\eta})$ rounds, the iterate will reach within $\tilde{\mathcal{O}}(\sqrt{\eta})$ from Γ (see Appendix J.5).

For the drift phase (Phase 2), we first prove that, with high probability, both $\bar{\boldsymbol{\theta}}^{(s)}$ and $\boldsymbol{\theta}_{k,t}^{(s)}$ stay close to Γ with a distance of only $\tilde{\mathcal{O}}(\sqrt{\eta})$ for all $\mathcal{O}(\log \frac{1}{\eta}) < s < \lfloor \frac{T}{H\eta^2} \rfloor$. We also provide a high probability bound on the movement of the manifold projection (see Appendix J.6). Based on these high probability bounds, we group $R_{\mathrm{grp}} := \lfloor \frac{1}{\alpha\eta^\beta} \rfloor$ rounds together and compute the first and second moments of $\Phi(\bar{\boldsymbol{\theta}}^{(s+R_{\mathrm{grp}})}) - \Phi(\bar{\boldsymbol{\theta}}^{(s)})$, which is the movement of manifold projection over R_{grp} rounds (see Appendix J.9). Here, β is a constant between 0 and 0.5 and will be specified later.

Finally, utilizing the estimation of moments in Appendix J.9, we prove that $\{\boldsymbol{\zeta}(t) : t \in [0, T]\}$ following the SDE (4) are weak approximations of each other following Li et al. [34] in Appendix J.10.

Appendix J. Proof Details of Main Theorems

The detailed proof is organized as follows. In Appendix J.1, we introduce the notations that will be used throughout the proof. To establish preliminary knowledge, Appendix J.2 provides explicit expression for the projection operator $\Phi(\cdot)$ and Appendix J.3 presents lemmas about gradient descent (GD) and gradient flow (GF). Based on the preliminary knowledge, we prove the working zone lemma in Appendix J.4. Appendices J.5 to J.10 make up the main body of the proof. Specifically, Appendices J.5 and J.6 derive high probability bounds for phase 1 and 2 respectively. Then, we provide a summary of these high probability bounds in Appendix J.7 and the proof of Theorem 5 in Appendix J.8. Utilizing the high probability bounds, we derive the estimation for the first and second moments of the one step update $\Phi(\bar{\boldsymbol{\theta}}^{(s+R_{\mathrm{grp}})}) - \Phi(\bar{\boldsymbol{\theta}}^{(s)})$ in Appendix J.9. Finally, we prove the approximation theorem 4 in Appendix J.10.

J.1. Additional Notations

Let $R_{\mathrm{tot}} := \lfloor \frac{T}{H\eta^2} \rfloor$ be the total number of rounds. Denote by $\boldsymbol{\phi}^{(s)}$ the manifold projection of the global iterate at the beginning of round s. Let $\boldsymbol{x}_{k,t}^{(s)} := \boldsymbol{\theta}_{k,t}^{(s)} - \boldsymbol{\phi}^{(s)}$ be the difference between the local iterate and the manifold projection of the global iterate. Also define $\bar{\boldsymbol{x}}_H^{(s)} := \frac{1}{K}\sum_{k\in[K]} \boldsymbol{x}_{k,H}^{(s)}$ and $\bar{\boldsymbol{x}}_0^{(s)} := \frac{1}{K}\sum_{k\in[K]} \boldsymbol{x}_{k,0}^{(s)}$ which is the average of $\boldsymbol{x}_{k,t}^{(s)}$ among K workers at step 0 and H. Then for all $k \in [K]$, $\boldsymbol{x}_{k,0}^{(s)} = \bar{\boldsymbol{x}}_0^{(s)} = \bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}$. Finally, Since $\nabla \ell(\boldsymbol{\theta};\boldsymbol{\zeta})$ is bounded, the gradient noise $\boldsymbol{z}_{k,t}^{(s)}$ is also bounded and we denote by $\sigma_{\max}$ the upper bound such that $\|\boldsymbol{z}_{k,t}^{(s)}\|_2 \le \sigma_{\max}, \forall s, k, t$.

J.2. Computing the Derivatives of the Limiting Mapping

In subsection, we present lemmas that relate the derivatives of the limiting mapping $\Phi(\cdot)$ to the derivatives of the loss function $\mathcal{L}(\cdot)$. We first introduce the operator $\mathcal{V}_{\boldsymbol{H}}$.

Definition 15 *For a semi-definite symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$, let λ_j, $\boldsymbol{v}_j$ be the j-th eigenvalue and eigenvector and $\boldsymbol{v}_j$'s form an orthonormal basis of $\mathbb{R}^d$. Then, define the operator $\mathcal{V}_{\boldsymbol{H}} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ as*

$$\mathcal{V}_{\boldsymbol{H}}(\boldsymbol{M}) := \sum_{i,j:\lambda_i\neq 0 \vee \lambda_j \neq 0} \frac{1}{\lambda_i \lambda_j} \left\langle \boldsymbol{M}, \boldsymbol{v}_i \boldsymbol{v}_j^\top \right\rangle \boldsymbol{v}_i \boldsymbol{v}_j^\top, \forall \boldsymbol{M} \in \mathbb{R}^{d\times d}.$$

Intuitively, this operator projects $\boldsymbol{M}$ to the base matrix $\boldsymbol{v}_i\boldsymbol{v}_j^\top$ and sums up the projections with weights $\frac{1}{\lambda_i+\lambda_j}$.

Additionally, for $\boldsymbol{\theta} \in \Gamma$, denote by $T_{\boldsymbol{\theta}}$ and $T_{\boldsymbol{\theta}}^{\perp}$ the tangent and normal space of Γ at $\boldsymbol{\theta}$ respectively. Lemmas 16 to 19 are from Li et al. [37]. We include them to make the paper self-contained.

Lemma 16 (Lemma C.1 of Li et al. [37]) *For any $\boldsymbol{\theta} \in \Gamma$ and any $\boldsymbol{v} \in T_{\boldsymbol{\theta}}(\Gamma)$, it holds that $\nabla^2 \mathcal{L}(\boldsymbol{\theta})\boldsymbol{v} = \mathbf{0}$.*

Lemma 17 (Lemma 4.3 of Li et al. [37]) *For any $\boldsymbol{\theta} \in \Gamma$, $\partial\Phi(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ is the projection matrix onto the tangent space $T_{\boldsymbol{\theta}}(\Gamma)$.*

Lemma 18 (Lemma C.4 of [37]) *For any $\boldsymbol{\theta} \in \Gamma$, $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v} \in T_{\boldsymbol{\theta}}(\Gamma)$, it holds that*

$$\partial^2\Phi(\boldsymbol{\theta})[\boldsymbol{v},\boldsymbol{u}] = -\partial\Phi(\boldsymbol{\theta})\nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{v}, \nabla^2\mathcal{L}(\boldsymbol{\theta})^+\boldsymbol{u}] - \nabla^2\mathcal{L}(\boldsymbol{\theta})^+\nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{v}, \partial\Phi(\boldsymbol{\theta})\boldsymbol{u}].$$

Lemma 19 (Lemma C.6 of [37]) *For any $\boldsymbol{\theta} \in \Gamma$ and $\boldsymbol{\Sigma} \in \operatorname{span}\{\boldsymbol{u}\boldsymbol{u}^\top \mid \boldsymbol{u} \in T_{\boldsymbol{\theta}}^{\perp}(\Gamma)\}$,*

$$\left\langle \partial^2\Phi(\boldsymbol{\theta}), \boldsymbol{\Sigma} \right\rangle = -\partial\Phi(\boldsymbol{\theta})\nabla^3\mathcal{L}(\boldsymbol{\theta})[\mathcal{V}_{\nabla^2\mathcal{L}(\boldsymbol{\theta})}(\boldsymbol{\Sigma})].$$

Lemma 20 *For all $\boldsymbol{\theta} \in \Gamma$, $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{\theta}}(\Gamma)$, it holds that*

$$\partial\Phi(\boldsymbol{\theta})\nabla^3\mathcal{L}[\boldsymbol{v}\boldsymbol{u}^\top] = \mathbf{0}. \tag{25}$$

Proof This proof is inspired by Lemma C.4 of [37]. For any $\boldsymbol{\theta} \in \Gamma$, consider a parameterized smooth curve $\boldsymbol{v}(t), t \geq 0$ on Γ such that $\boldsymbol{v}(0) = \boldsymbol{\theta}$ and $\boldsymbol{v}'(0) = \boldsymbol{v}$. Let $\boldsymbol{P}_{\|}(t) = \partial\Phi(\boldsymbol{v}(t))$, $\boldsymbol{P}_{\perp}(t) = \boldsymbol{I} - \partial\Phi(\boldsymbol{v}(t))$ and $\boldsymbol{H}(t) = \nabla^2\mathcal{L}(\boldsymbol{v}(t))$. By Lemma C.1 and 4.3 in [37],

$$\boldsymbol{H}(t) = \boldsymbol{P}_{\perp}(t)\boldsymbol{H}(t).$$

Take the derivative with respect to t on both sides,

$$\begin{aligned}\boldsymbol{H}'(t) &= \boldsymbol{P}_{\perp}(t)\boldsymbol{H}'(t) + \boldsymbol{P}_{\perp}'(t)\boldsymbol{H}(t)\\ \Rightarrow \boldsymbol{P}_{\|}(t)\boldsymbol{H}'(t) &= \boldsymbol{P}_{\perp}'(t)\boldsymbol{H}(t) = -\boldsymbol{P}_{\|}'(t)\boldsymbol{H}(t).\end{aligned}$$

At $t = 0$, we have

$$\boldsymbol{P}_{\|}(0)\boldsymbol{H}'(0) = -\boldsymbol{P}_{\|}'(0)\boldsymbol{H}(0). \tag{26}$$

WLOG let $\boldsymbol{H}(0) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d), \in \mathbb{R}^{d\times d}$, where $\lambda_i = 0$ for all $m < i \leq d$. Therefore $\boldsymbol{P}_{\perp}(0) = \begin{bmatrix}\boldsymbol{I}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$, $\boldsymbol{P}_{\|}(0) = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}_{d-m}\end{bmatrix}$. Decompose $\boldsymbol{P}_{\|}'(0)$, $\boldsymbol{H}(0)$ and $\boldsymbol{H}'(0)$as follows.

$$\boldsymbol{P}_{\|}'(0) = \begin{bmatrix}\boldsymbol{P}_{\|,11}'(0) & \boldsymbol{P}_{\|,12}'(0)\\ \boldsymbol{P}_{\|,21}'(0) & \boldsymbol{P}_{\|,22}'(0)\end{bmatrix}, \boldsymbol{H}(0) = \begin{bmatrix}\boldsymbol{H}_{11}(0) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, \boldsymbol{H}'(0) = \begin{bmatrix}\boldsymbol{H}_{11}'(0) & \boldsymbol{H}_{12}'(0)\\ \boldsymbol{H}_{21}'(0) & \boldsymbol{H}_{22}'(0)\end{bmatrix}.$$

Substituting the decomposition into (26), we have

$$\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{H}_{21}'(0) & \boldsymbol{H}_{22}'(0)\end{bmatrix} = -\begin{bmatrix}\boldsymbol{P}_{\|,11}'(0)\boldsymbol{H}_{11}(0) & \mathbf{0}\\ \boldsymbol{P}_{\|,21}'(0)\boldsymbol{H}_{11}(0) & \mathbf{0}\end{bmatrix}.$$

Therefore, $\boldsymbol{H}_{22}'(0) = \mathbf{0}$ and

$$\boldsymbol{P}_{\|}(0)\boldsymbol{H}'(0) = -\boldsymbol{P}_{\|}'(0)\boldsymbol{H}(0) = -\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{H}_{21}'(0) & \mathbf{0}\end{bmatrix}.$$

Any $\boldsymbol{u} \in T_{\boldsymbol{\theta}}(\Gamma)$ can be decomposed as $\boldsymbol{u} = \begin{bmatrix}\mathbf{0}, \boldsymbol{u}_2\end{bmatrix}^\top$ where $\boldsymbol{u}_2 \in \mathbb{R}^{d-m}$. With this decomposition, we have $\boldsymbol{P}_{\|}(0)\boldsymbol{H}'(0)\boldsymbol{u} = \mathbf{0}$. Also, note that $\boldsymbol{H}'(0) = \nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{v}]$. Hence,

$$\partial\Phi(\boldsymbol{\theta})\nabla^3\mathcal{L}(\boldsymbol{\theta})[\boldsymbol{v}\boldsymbol{u}^T] = \mathbf{0}.$$

■

J.3. Preliminary Lemmas for GD and GF

In this subsection, we introduce a few useful preliminary lemmas about gradient descent and gradient flow. Before presenting the lemmas, we introduce some notations and assumptions that will be used in this subsection.

Assume that the loss function $\mathcal{L}(\boldsymbol{\theta})$ is ρ-smooth and μ-PL in an open, convex neighborhood U of a local minimizer $\boldsymbol{\theta}^*$. Denote by $\mathcal{L}^* := \mathcal{L}(\boldsymbol{\theta}^*)$ the minimum value for simplicity. Let ϵ' be the radius of the open ϵ'-ball centered at $\boldsymbol{\theta}^*$ such that $B^{\epsilon'}(\boldsymbol{\theta}^*) \subseteq U$. We also define a potential function $\tilde{\Psi}(\boldsymbol{\theta}) := \sqrt{\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^*}$.

Consider gradient descent iterates $\{\hat{\boldsymbol{u}}_t\}_{t\in\mathbb{N}}$ following the update rule $\hat{\boldsymbol{u}}_{t+1} = \hat{\boldsymbol{u}}_t - \eta\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)$. We first introduce the descent lemma for gradient descent.

Lemma 21 (Descent lemma for GD) *If $\hat{\boldsymbol{u}}_t \in U$ and $\eta \leq \frac{1}{\rho}$, then*

$$\frac{\eta}{2}\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2^2 \leq \mathcal{L}(\hat{\boldsymbol{u}}_t) - \mathcal{L}(\hat{\boldsymbol{u}}_{t+1}),$$

and

$$\mathcal{L}(\hat{\boldsymbol{u}}_{t+1}) - \mathcal{L}^* \leq (1-\mu\eta)(\mathcal{L}(\hat{\boldsymbol{u}}_t) - \mathcal{L}^*).$$

Proof By ρ-smoothness,

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{u}}_{t+1}) &\leq \mathcal{L}(\hat{\boldsymbol{u}}_t) + \langle\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t), \hat{\boldsymbol{u}}_{t+1} - \hat{\boldsymbol{u}}_t\rangle + \frac{\rho\eta^2}{2}\|\hat{\boldsymbol{u}}_{t+1} - \hat{\boldsymbol{u}}_t\|_2^2 \\ &= \mathcal{L}(\hat{\boldsymbol{u}}_t) - \eta(1 - \frac{\rho\eta}{2})\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2^2 \\ &\leq \mathcal{L}(\hat{\boldsymbol{u}}_t) - \frac{\eta}{2}\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2^2\end{aligned}$$

By the definition of μ-PL, we have

$$\mathcal{L}(\hat{\boldsymbol{u}}_{t+1}) - \mathcal{L}^* \leq (1-\mu\eta)(\mathcal{L}(\hat{\boldsymbol{u}}_t) - \mathcal{L}^*).$$

■

Then we prove the Lipschitzness of $\tilde{\Psi}(\boldsymbol{\theta})$.

Lemma 22 (Lipschitzness of $\tilde{\Psi}(\boldsymbol{\theta})$) *$\tilde{\Psi}(\boldsymbol{\theta})$ is $\sqrt{2\rho}$-Lipschitz for $\boldsymbol{\theta} \in U$. That is, for any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in U$,*

$$|\tilde{\Psi}(\boldsymbol{\theta}_1) - \tilde{\Psi}(\boldsymbol{\theta}_2)| \leq \sqrt{2\rho}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Proof Fix $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Denote by $\boldsymbol{\theta}(t) := (1-t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2$ the convex combination of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ where $t \in [0,1]$. Further define $f(t) := \tilde{\Psi}(\boldsymbol{\theta}(t))$. Below we consider two cases.

Case 1. If $\forall t \in (0,1)$, $f(t) > 0$, then $f(t)$ is differentiable on $(0,1)$.

$$\begin{aligned}
|\tilde{\Psi}(\boldsymbol{\theta}_2) - \tilde{\Psi}(\boldsymbol{\theta}_1)| &= |f(1) - f(0)| \\
&= \left| \int_0^1 f'(t) \mathrm{d}t \right| \\
&= \left| \int_0^1 \left\langle \nabla \tilde{\Psi}(\boldsymbol{\theta}(t)), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \right\rangle \mathrm{d}t \right| \\
&= \left| \int_0^1 \frac{\langle \nabla \mathcal{L}(\boldsymbol{\theta}(t)), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle}{\sqrt{\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*}} \mathrm{d}t \right| \\
&\le \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2 \int_0^1 \frac{\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|_2}{\sqrt{\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*}} \mathrm{d}t.
\end{aligned}$$

By ρ-smoothness of $\mathcal{L}$, for all $\boldsymbol{\theta} \in U$,

$$\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \le 2\rho \left(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}^* \right).$$

Since $\sqrt{\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*} > 0$ for all $t \in (0,1)$, $\frac{\|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|_2}{\sqrt{\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}^*}} \le \sqrt{2\rho}$. Therefore,

$$|\tilde{\Psi}(\boldsymbol{\theta}_2) - \tilde{\Psi}(\boldsymbol{\theta}_1)| \le \sqrt{2\rho_2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2.$$

Case 2. If $\exists t' \in (0,1)$ such that $f(t') = 0$, then

$$\begin{aligned}
|\tilde{\Psi}(\boldsymbol{\theta}_2) - \tilde{\Psi}(\boldsymbol{\theta}_1)| &= |f(1) - f(0)| \\
&= \left| (1-t') \frac{f(1) - f(t')}{1-t'} + t' \left(\frac{f(t') - f(0)}{t'} \right) \right| \\
&\le \max \left(\frac{f(1)}{1-t'}, \frac{f(0)}{t'} \right).
\end{aligned}$$

Since $\boldsymbol{\theta}(t')$ minimizes $\mathcal{L}$ in an open set, $\nabla \mathcal{L}(\boldsymbol{\theta}(t')) = \mathbf{0}$. By ρ-smoothness of $\mathcal{L}$, for all $\boldsymbol{\theta} \in U$,

$$\mathcal{L}(\boldsymbol{\theta}) \le \mathcal{L}^* + \frac{\rho}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}(t')\|_2^2 \quad \Rightarrow \quad \tilde{\Psi}(\boldsymbol{\theta}) \le \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta} - \boldsymbol{\theta}(t')\|_2.$$

Therefore,

$$\begin{aligned}
f(1) &\le \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}(t')\|_2 = (1-t') \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2 \\
f(0) &\le \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}(t')\|_2 = t' \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2.
\end{aligned}$$

Then we have

$$|\tilde{\Psi}(\boldsymbol{\theta}_2) - \tilde{\Psi}(\boldsymbol{\theta}_1)| \le \sqrt{\frac{\rho}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2.$$

Combining case 1 and case 2, we conclude the proof. ■

Below we introduce a lemma that relates the movement of one step gradient descent to the change of the potential function.

Lemma 23 (Lemma G.1 in [40]) *If $\hat{\boldsymbol{u}}_t \in U$ and $\eta \le 1/\rho_2$ then*

$$\tilde{\Psi}(\hat{\boldsymbol{u}}_t) - \tilde{\Psi}(\hat{\boldsymbol{u}}_{t+1}) \ge \frac{\sqrt{2\mu}}{4}\eta\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2.$$

Proof

$$\begin{aligned}\tilde{\Psi}(\hat{\boldsymbol{u}}_t) - \tilde{\Psi}(\hat{\boldsymbol{u}}_{t+1}) &= \frac{\mathcal{L}(\hat{\boldsymbol{u}}_t) - \mathcal{L}(\hat{\boldsymbol{u}}_{t+1})}{\tilde{\Psi}(\hat{\boldsymbol{u}}_t) + \tilde{\Psi}(\hat{\boldsymbol{u}}_{t+1})} \\ &\ge \frac{\mathcal{L}(\hat{\boldsymbol{u}}_{t+1}) - \mathcal{L}(\hat{\boldsymbol{u}}_t)}{2\tilde{\Psi}(\hat{\boldsymbol{u}}_t)} \\ &\ge \frac{\eta(1-\rho_2\eta/2)\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2^2}{2\tilde{\Psi}(\hat{\boldsymbol{u}}_t)},\end{aligned}$$

where the two inequalities uses Theorem 21. By μ-PL, $\tilde{\Psi}(\hat{\boldsymbol{u}}_t) \le \frac{1}{\sqrt{2\mu}}\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2$. Therefore, we have $\tilde{\Psi}(\hat{\boldsymbol{u}}_t) - \tilde{\Psi}(\hat{\boldsymbol{u}}_{t+1}) \ge \frac{\sqrt{2\mu}}{2}(1-\eta\rho/2)\eta\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2 \ge \frac{\sqrt{2\mu}}{4}\eta\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_t)\|_2$. ■

Based on Theorem 23, we have the following lemma that bounds the movement of GD over multiple steps.

Lemma 24 (Bounding the movement of GD) *If $\hat{\boldsymbol{u}}_0$ is initialized such that $\|\hat{\boldsymbol{u}}_0 - \boldsymbol{\theta}^*\|_2 \le \frac{1}{4}\sqrt{\frac{\mu}{\rho}}\epsilon'$, then for all $t \ge 0$, $\hat{\boldsymbol{u}}_t \in B^{\epsilon'}(\boldsymbol{\theta}^*)$ and*

$$\|\hat{\boldsymbol{u}}_t - \hat{\boldsymbol{u}}_0\|_2 \le \sqrt{\frac{8}{\mu}}\tilde{\Psi}(\hat{\boldsymbol{u}}_0).$$

Proof We prove the proposition by induction. When $t = 0$, it trivially holds. Assume that the proposition holds for $\hat{\boldsymbol{u}}_\tau$, $0 \le \tau < t$. For step t, since $\hat{\boldsymbol{u}}_\tau \in B^{\epsilon'}(\boldsymbol{\theta}^*)$, we apply Theorem 23 and obtain

$$\|\hat{\boldsymbol{u}}_t - \hat{\boldsymbol{u}}_0\|_2 \le \eta\sum_{\tau=0}^{t-1}\|\nabla\mathcal{L}(\hat{\boldsymbol{u}}_\tau)\|_2 \le \sqrt{\frac{8}{\mu}}\left(\tilde{\Psi}(\hat{\boldsymbol{u}}_0) - \tilde{\Psi}(\hat{\boldsymbol{u}}_t)\right) \le \sqrt{\frac{8}{\mu}}\tilde{\Psi}(\hat{\boldsymbol{u}}_0).$$

Further by ρ-smoothness of $\mathcal{L}(\cdot)$,

$$\|\hat{\boldsymbol{u}}_t - \hat{\boldsymbol{u}}_0\|_2 \le \sqrt{\frac{8}{\mu}}\tilde{\Psi}(\hat{\boldsymbol{u}}_0) \le 2\sqrt{\frac{\rho}{\mu}}\|\hat{\boldsymbol{u}}_0 - \boldsymbol{\theta}^*\|_2 \le \frac{1}{2}\epsilon'.$$

Therefore, $\|\hat{\boldsymbol{u}}_t - \boldsymbol{\theta}^*\|_2 \le \|\hat{\boldsymbol{u}}_t - \hat{\boldsymbol{u}}_0\|_2 + \|\hat{\boldsymbol{u}}_0 - \boldsymbol{\theta}^*\|_2 < \epsilon'$, which concludes the proof. ■

Finally, we introduce a lemma adapted from Thm. D.4 of which bounds the movement of GF. [40].

Lemma 25 *Assume that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2 < \sqrt{\frac{\mu}{\rho}}\epsilon'$. The gradient flow $\boldsymbol{\theta}(t) = -\frac{\mathrm{d}\mathcal{L}(\boldsymbol{\theta}(t))}{\mathrm{d}t}$ starting at $\boldsymbol{\theta}_0$ converges to a point in U and*

$$\left\|\boldsymbol{\theta}_0 - \lim_{t\to+\infty}\boldsymbol{\theta}(t)\right\|_2 \le \sqrt{\frac{2}{\mu}}\sqrt{\mathcal{L}(\boldsymbol{\theta}_0) - \mathcal{L}^*} \le \sqrt{\frac{\rho}{\mu}}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$$

Proof Let $T := \inf\{t : \boldsymbol{\theta} \notin U\}$. Then for all $t < T$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}^*\right)^{1/2} &= \frac{1}{2}\left(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}^*\right)^{-1/2}\cdot\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}),\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\right\rangle \\ &= -\frac{1}{2}(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}^*)^{-1/2}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2\|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\|_2.\end{aligned}$$

By μ-PL, $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \geq \sqrt{2\mu(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}^*)}$. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}^*\right)^{1/2} \leq -\frac{\sqrt{2\mu}}{2}\|\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\|_2.$$

Integrating both sides, we have

$$\int_0^T \|\frac{\mathrm{d}\boldsymbol{\theta}(\tau)}{\mathrm{d}\tau}\|\mathrm{d}\tau \leq \frac{2}{\sqrt{2\mu}}(\mathcal{L}(\boldsymbol{\theta}_0)-\mathcal{L}^*)^{1/2} \leq \sqrt{\frac{\rho}{\mu}}\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|_2 < \epsilon',$$

where the second inequality uses ρ-smoothness of $\mathcal{L}$. Therefore, $T = +\infty$ and $\boldsymbol{\theta}(t)$ converges to some point in U. ■

J.4. Construction of working zones

In this subsection, we provide the proof for Theorem 14. Note that the notions of $\mathcal{Z}^\epsilon$, $\mathcal{M}^{\epsilon_4}$, ρ_2, ρ_3, ν_1 and ν_2 defined in the proof will be useful in the remaining part of this section.

Proof [Proof of Theorem 14] Let $\bar{\boldsymbol{\theta}}^{(0)}$ be initialized such that $\Phi(\bar{\boldsymbol{\theta}}^{(0)}) \in \Gamma$. Let $\mathcal{Z}$ be the set of all points on the gradient flow trajectory starting from $\bar{\boldsymbol{\theta}}^{(0)}$ and $\mathcal{Z}^\epsilon$ be the ϵ-neighborhood of $\mathcal{Z}$, where ϵ is a positive constant. Since the gradient flow converges to $\boldsymbol{\phi}^{(0)}$, $\mathcal{Z}$ and $\mathcal{Z}^\epsilon$ are bounded.

We construct four nested working zones. By Lemma H.3 in [40], there exists an ϵ_3-neighborhood of Γ, Γ^{ϵ_3}, such that $\mathcal{L}$ satisfies μ-PL for some $\mu > 0$. Let $\mathcal{M}$ be the convex hull of $\Gamma^{\epsilon_3} \cup \mathcal{Z}^\epsilon$ and $\mathcal{M}^{\epsilon_4}$ be the ϵ_4-neighborhood of $\mathcal{M}$ where ϵ_4 is a positive constant. Then $\mathcal{M}^{\epsilon_4}$ is bounded.

Define $\rho_2 = \sup_{\boldsymbol{\theta}\in\mathcal{M}^{\epsilon_4}} \|\nabla^2\mathcal{L}(\boldsymbol{\theta})\|_2$ and $\rho_3 = \sup_{\mathcal{M}^{\epsilon_4}} \|\nabla^3\mathcal{L}(\boldsymbol{\theta})\|_2$. By Theorem 25, we can construct an ϵ_2-neighborhood of Γ where $\epsilon_2 < \sqrt{\frac{\mu}{\rho_2}}\epsilon_3$ such that all GF starting in Γ^{ϵ_2} converges to Γ. By Falconer [10], $\Phi(\cdot)$ is $\mathcal{C}^2$ in Γ^{ϵ_3}. Define $\nu_1 = \sup_{\boldsymbol{\theta}\in\Gamma^{\epsilon_3}} \|\partial\Phi(\boldsymbol{\theta})\|_2$ and $\nu_2 = \sup_{\boldsymbol{\theta}\in\Gamma^{\epsilon_3}} \|\partial^2\Phi(\boldsymbol{\theta})\|_2$. We also construct an ϵ_1 neighborhood of Γ, Γ^{ϵ_1}, where $\epsilon_1 \leq \frac{1}{2}\epsilon_2 < \frac{1}{2}\sqrt{\frac{\mu}{\rho_2}}\epsilon_3$ such that all $\boldsymbol{\theta} \in \Gamma^{\epsilon_1}$ has an ϵ_1 neighborhood where Φ is well defined. Finally, by Theorem 24, there exists an ϵ_0-neighborhood of Γ where $\epsilon_0 \leq \frac{1}{4}\sqrt{\frac{\mu}{\rho_2}}\epsilon_1$ such that all gradient descent iterates starting in Γ^{ϵ_0} with $\eta \leq \frac{1}{\rho_2}$ will stay in Γ^{ϵ_1}. ■

When analyzing the limiting dynamics of Local SGD, we will show that all $\boldsymbol{\theta}_{k,t}^{(s)}$ stays in Γ^{ϵ_2}, $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_1}$, $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$ with high probability after $\mathcal{O}(\log\frac{1}{\eta})$ rounds.

J.5. High Probability Bounds for Phase 1: Iterate Approaching the Manifold

The approaching phase can be further divided two subphases. In the first subphase, $\bar{\boldsymbol{\theta}}^{(0)}$ is initialized such that $\boldsymbol{\phi}^{(0)} \in \Gamma$. We will show that after a constant number of rounds s_0, $\bar{\boldsymbol{\theta}}^{(s_0)}$ goes to the inner part of Γ^{ϵ_0} such that $\|\bar{\boldsymbol{\theta}}^{(s_0)} - \boldsymbol{\phi}^{(0)}\|_2 \le c\epsilon_0$ with high probability, where $0 < c < 1$ and the constants will be specified later (see Appendix J.5.2). In the second subphase, we show that the iterate can reach within $\tilde{\mathcal{O}}(\sqrt{\eta})$ distance from Γ after $\mathcal{O}(\log \frac{1}{\eta})$ rounds with high probability (see Appendix J.5.3).

J.5.1. ADDITIONAL NOTATIONS

Consider an auxiliary sequence $\{\tilde{\boldsymbol{u}}_t^{(s)}\}$ where $\tilde{\boldsymbol{u}}_0^{(s)} = \bar{\boldsymbol{\theta}}^{(s)}$ and $\tilde{\boldsymbol{u}}_{t+1}^{(s)} = \tilde{\boldsymbol{u}}_t^{(s)} - \eta\nabla\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}), 0 \le t \le H-1$. Define $\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)} := \boldsymbol{\theta}_{k,t}^{(s)} - \tilde{\boldsymbol{u}}_t^{(s)}$ to be the difference between the local iterate and the gradient descent iterate. Notice that $\tilde{\boldsymbol{\Delta}}_{k,0}^{(s)} = 0$, for all k and s.

Consider a gradient flow $\{\boldsymbol{u}(t)\}_{t\ge 0}$ with the initial condition $\boldsymbol{u}(0) = \bar{\boldsymbol{\theta}}^{(0)}$ and converges to $\boldsymbol{\phi}^{(0)} \in \Gamma$. For simplicity, let $\boldsymbol{u}_t^{(s)} := \boldsymbol{u}(s\alpha + t\eta)$ be the gradient flow after s rounds plus t steps. Let s_0 be the smallest number such that $\|\boldsymbol{u}_0^{(s_0)} - \boldsymbol{\phi}^{(0)}\|_2 \le \frac{1}{4}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0$. Note that s_0 is a constant independent of η.

In this subsection, the minimum value of the loss in Appendix J.3 corresponds to the loss value on Γ, i.e., $\mathcal{L}^* = \mathcal{L}(\boldsymbol{\phi}), \forall \boldsymbol{\phi} \in \Gamma$.

We also define the following sequence $\{\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\}_{t=0}^{H}$ that will be used in the proof. Define

$$\tilde{\boldsymbol{Z}}_{k,t}^{(s)} := \sum_{\tau=0}^{t-1}\left(\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\right)\boldsymbol{z}_{k,\tau}^{(s)}, \qquad \tilde{\boldsymbol{Z}}_{k,0}^{(s)} = \boldsymbol{0}.$$

J.5.2. PROOF FOR SUBPHASE 1

First, we have the following lemma about the concentration of $\tilde{\boldsymbol{Z}}_{k,t}^{(s)}$.

Lemma 26 (Concentration property of $\{\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\}_{t=0}^{H}$) *Given $\bar{\boldsymbol{\theta}}^{(s)}$ such that $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_3} \cup \mathcal{Z}^{\epsilon}$ for all $0 \le t \le H$, then with probability at least $1-\delta$,*

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \le \tilde{C}_1\sigma_{\max}\sqrt{2H\log\frac{2HK}{\delta}}, \qquad \forall 0 \le t \le H, k \in [K],$$

where $\tilde{C}_1 := \exp(\alpha\rho_2)$.

Proof For each $\tilde{\boldsymbol{Z}}_{k,t}^{(s)}$, construct a sequence $\{\tilde{\boldsymbol{Z}}_{k,t,t'}^{(s)}\}_{t'=0}^{t}$:

$$\tilde{\boldsymbol{Z}}_{k,t,t'}^{(s)} := \sum_{\tau=0}^{t'-1}\left(\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\right)\boldsymbol{z}_{k,\tau}^{(s)}, \qquad \tilde{\boldsymbol{Z}}_{k,t,0}^{(s)} = \boldsymbol{0}.$$

Since $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_3} \cup \mathcal{Z}^{\epsilon}$, we have $\|\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})\|_2 \le \rho_2$ for all $0 \le t \le H$. Then, for all τ and t,

$$\|\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\|_2 \le (1+\rho_2\eta)^H \le \exp(\alpha\rho_2) = \tilde{C}_1.$$

Notice that for all $0 \le t \le H$, $\{\tilde{\boldsymbol{Z}}^{(s)}_{k,t,t'}\}_{t'=0}^{t}$ is a martingale with $\|\tilde{\boldsymbol{Z}}^{(s)}_{k,t,t'} - \tilde{\boldsymbol{Z}}^{(s)}_{k,t,t'-1}\|_2 \le \tilde{C}_1\sigma_{\max}$. By Azuma-Hoeffding's inequality,

$$\mathbb{P}(\|\tilde{\boldsymbol{Z}}^{(s)}_{k,t}\|_2 \ge \epsilon') \le 2\exp\left(\frac{-\epsilon'^2}{2t\left(\tilde{C}_1\sigma_{\max}\right)^2}\right) \le 2\exp\left(\frac{-\epsilon'^2}{2H\left(\tilde{C}_1\sigma_{\max}\right)^2}\right).$$

Taking a union bound on all $k \in [K]$ and $0 \le t \le H$, we can conclude that with probability at least $1-\delta$,

$$\|\tilde{\boldsymbol{Z}}^{(s)}_{k,t}\|_2 \le \tilde{C}_1\sigma_{\max}\sqrt{2H\log\frac{2HK}{\delta}}, \qquad \forall 0 \le t \le H, k \in [K].$$

■

The following lemma states that the gradient descent iterates will closely track the gradient flow with the same initial point.

Lemma 27 *Denote $G := \sup_{t\ge 0}\|\nabla\mathcal{L}(\boldsymbol{u}(t))\|_2$ as the upper bound of the gradient on the gradient flow trajectory. If $\|\tilde{\boldsymbol{u}}^{(s)}_t - \boldsymbol{u}^{(s)}_t\|_2 = \mathcal{O}(\sqrt{\eta})$, then for all $0 \le t \le H$, the closeness of $\tilde{\boldsymbol{u}}^{(s)}_t$ and $\boldsymbol{u}^{(s)}_t$ is bounded by*

$$\|\tilde{\boldsymbol{u}}^{(s)}_t - \boldsymbol{u}^{(s)}_t\|_2 \le \tilde{C}_1\|\tilde{\boldsymbol{u}}^{(s)}_0 - \boldsymbol{u}^{(s)}_0\|_2 + \tilde{C}_1\eta G,$$

where $\tilde{C}_1 = \exp(\alpha\rho_2)$.

Proof We prove by induction that

$$\|\tilde{\boldsymbol{u}}^{(s)}_t - \boldsymbol{u}^{(s)}_t\|_2 \le (1+\rho_2\eta)^t\|\tilde{\boldsymbol{u}}^{(s)}_0 - \boldsymbol{u}^{(s)}_0\|_2 + \rho_2\eta^2 G\sum_{\tau=0}^{t-1}(1+\rho_2\eta)^\tau. \tag{27}$$

When $t=0$, (27) holds trivially. Assume that (27) holds for $0 \le \tau \le t$, then

$$\begin{aligned}\tilde{\boldsymbol{u}}^{(s)}_{t+1} - \boldsymbol{u}^{(s)}_{t+1} &= \tilde{\boldsymbol{u}}^{(s)}_t - \eta\nabla\mathcal{L}(\tilde{\boldsymbol{u}}^{(s)}_t) - \left(\boldsymbol{u}_t - \int_{s\alpha+t\eta}^{s\alpha+(t+1)\eta}\nabla\mathcal{L}(\boldsymbol{u}(v))dv\right)\\ &= \tilde{\boldsymbol{u}}^{(s)}_t - \boldsymbol{u}_t - \eta\left(\nabla\mathcal{L}(\tilde{\boldsymbol{u}}^{(s)}_t) - \nabla\mathcal{L}(\boldsymbol{u}^{(s)}_t)\right)\\ &\quad - \int_{s\alpha+t\eta}^{s\alpha+(t+1)\eta}\left(\nabla\mathcal{L}(\boldsymbol{u}^{(s)}_t) - \nabla\mathcal{L}(\boldsymbol{u}(v))\right)dv.\end{aligned}$$

By smoothness of $\mathcal{L}$,

$$\begin{aligned}\|\nabla\mathcal{L}(\boldsymbol{u}^{(s)}_t) - \nabla\mathcal{L}(\boldsymbol{u}(v))\|_2 &\le \rho_2\|\boldsymbol{u}^{(s)}_t - \boldsymbol{u}(v)\|_2\\ &\le \rho_2\int_{s\alpha+t\eta}^{v}\|\nabla\mathcal{L}(\boldsymbol{u}(w))\|_2 dw\\ &\le \rho_2\eta G.\end{aligned}$$

Since $\rho_2^2\eta^2 G\sum_{\tau=0}^{t-1}(1+\rho_2\eta)^\tau \le \eta G(1+\rho_2\eta)^t \le \exp(\alpha\rho_2)\eta G$, then $\|\tilde{\boldsymbol{u}}_t^{(s)} - \boldsymbol{u}_t^{(s)}\|_2 = \mathcal{O}(\sqrt{\eta})$, which implies that $\tilde{\boldsymbol{u}}_t^{(s)} \in \mathcal{M}^{\epsilon_4}$. Hence, $\|\nabla\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}) - \mathcal{L}(\boldsymbol{u}_t^{(s)})\|_2 \le \rho_2\|\tilde{\boldsymbol{u}}_t^{(s)} - \boldsymbol{u}_t^{(s)}\|_2$.

By triangle inequality,

$$\begin{aligned}\|\tilde{\boldsymbol{u}}_{t+1}^{(s)} - \boldsymbol{u}_{t+1}^{(s)}\|_2 &\le (1+\rho_2\eta)\|\tilde{\boldsymbol{u}}_t^{(s)} - \boldsymbol{u}_t^{(s)}\|_2 + \rho_2\eta^2 G \\ &\le (1+\rho_2\eta)^{t+1}\|\tilde{\boldsymbol{u}}_t^{(s)} - \boldsymbol{u}_t^{(s)}\|_2 + \rho_2\eta^2 G\sum_{\tau=0}^{t}(1+\rho_2\eta)^\tau,\end{aligned}$$

which concludes the induction step. Appling $1+\rho_2\eta \le \exp(\rho_2\eta)$, we have the lemma. ■

Utilizing the concentration probability of $\{\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\}$, we can obtain the following lemma which implies that the Local SGD iterates will closely track the gradient descent iterates with high probability.

Lemma 28 *Given $\bar{\boldsymbol{\theta}}^{(s)}$ such that $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_3} \cup \mathcal{Z}^{\epsilon}$ for all $0 \le t \le H$, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, there exists a constant $\tilde{C}_3$ such that*

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \tilde{\boldsymbol{u}}_t^{(s)}\|_2 \le \tilde{C}_3\sqrt{\eta\log\frac{1}{\eta\delta}}, \quad \forall 0 \le t \le H, k \in [K],$$

and

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \le \tilde{C}_3\sqrt{\eta\log\frac{1}{\eta\delta}}.$$

Proof Since $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_3} \cup \mathcal{Z}^{\epsilon}$ for all $0 \le t \le H$, we have $\|\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})\|_2 \le \rho_2$. According to the update rule for $\boldsymbol{\theta}_{k,t}^{(s)}$ and $\tilde{\boldsymbol{u}}_t^{(s)}$,

$$\boldsymbol{\theta}_{k,t+1}^{(s)} = \boldsymbol{\theta}_{k,t}^{(s)} - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \eta\boldsymbol{z}_{k,t}^{(s)}, \tag{28}$$

$$\tilde{\boldsymbol{u}}_{t+1}^{(s)} = \tilde{\boldsymbol{u}}_t^{(s)} - \eta\nabla\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}). \tag{29}$$

Subtracting (29) from (28) gives

$$\begin{aligned}\tilde{\boldsymbol{\Delta}}_{k,t+1}^{(s)} &= \tilde{\boldsymbol{\Delta}}_{k,t}^{(s)} - \eta(\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \nabla\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})) - \eta\boldsymbol{z}_{k,t}^{(s)} \\ &= (\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}))\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)} - \eta\boldsymbol{z}_{k,t}^{(s)} + \eta\tilde{\boldsymbol{v}}_{k,t}^{(s)}. \end{aligned}\tag{30}$$

Here, $\tilde{\boldsymbol{v}}_{k,t}^{(s)} = (1-\beta_{k,t}^{(s)})\boldsymbol{\theta}_{k,t}^{(s)} + \beta_{k,t}^{(s)}\tilde{\boldsymbol{u}}_{k,t}^{(s)}$, where $\beta_{k,t}^{(s)} \in (0,1)$ depends on $\boldsymbol{\theta}_{k,t}^{(s)}$ and $\tilde{\boldsymbol{u}}_t^{(s)}$. Therefore, $\|\tilde{\boldsymbol{v}}_{k,t}^{(s)}\|_2 \le \frac{\rho_3}{2}\|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2^2$ if $\boldsymbol{\theta}_{k,t}^{(s)} \in \mathcal{M}^{\epsilon_4}$. Applying (30) t times, we have

$$\begin{aligned}\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)} = \left[\prod_{\tau=0}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_\tau^{(s)}))\right]\tilde{\boldsymbol{\Delta}}_{k,0}^{(s)} - \eta\sum_{\tau=0}^{t-1}\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\boldsymbol{z}_{k,\tau}^{(s)} \\ + \eta\sum_{\tau=0}^{t-1}\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\tilde{\boldsymbol{v}}_{k,\tau}^{(s)}.\end{aligned}$$

By Cauchy-Schwartz inequality, triangle inequality and the definition of $\tilde{\boldsymbol{Z}}_{k,t}^{(s)}$, if for all $0 \leq \tau \leq t-1$ and $k \in [K]$, $\boldsymbol{\theta}_{k,\tau}^{(s)} \in \mathcal{M}^{\epsilon_4}$, then we have

$$\|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2 \leq \eta \|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 + \frac{1}{2}\eta\rho_3 \sum_{\tau=0}^{t-1} \tilde{C}_1 \|\tilde{\boldsymbol{\Delta}}_{k,\tau}^{(s)}\|_2^2. \tag{31}$$

Applying Theorem 26 and substituting in the value of H, we have that with probability at least $1-\delta$,

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_1 \sigma_{\max} \sqrt{\frac{2\alpha}{\eta} \log \frac{2\alpha K}{\eta\delta}}, \qquad \forall k \in K, 0 \leq t \leq H. \tag{32}$$

Now we show by induction that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when (32) holds, there exists a constant $\tilde{C}_2 > 2\sigma_{\max}\sqrt{2\alpha}\tilde{C}_1$ such that $\|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_2 \sqrt{\eta \log \frac{2\alpha K}{\eta\delta}}$.

When $t = 0$, $\tilde{\boldsymbol{\Delta}}_{k,0}^{(s)} = 0$. Assume that $\|\tilde{\boldsymbol{\Delta}}_{k,\tau}^{(s)}\|_2 \leq \tilde{C}_2 \sqrt{\eta \log \frac{2\alpha K}{\eta\delta}}$, for all $k \in [K], 0 \leq \tau \leq t-1$. Then for all $0 \leq \tau \leq t-1$, $\boldsymbol{\theta}_{k,\tau}^{(s)} \in \mathcal{M}^{\epsilon_4}$. Therefore, we can apply (31) and obtain

$$\begin{aligned} \|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2 &\leq \eta \|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 + \frac{1}{2}\eta\rho_3 \sum_{\tau=0}^{t-1} \tilde{C}_1 \|\tilde{\boldsymbol{\Delta}}_{k,\tau}^{(s)}\|_2^2 \\ &\leq \tilde{C}_1 \sigma_{\max} \sqrt{2\alpha\eta \log \frac{2\alpha K}{\eta\delta}} + \frac{1}{2}\tilde{C}_1 \tilde{C}_2^2 \sigma_{\max}^2 \alpha\rho_3\eta \log \frac{2\alpha K}{\eta\delta}. \end{aligned}$$

Given that $\tilde{C}_2 \geq 2\sigma_{\max}\sqrt{2\alpha}\tilde{C}_1$ and $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when η is sufficiently small, $\|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_2 \sqrt{\eta \log \frac{2\alpha K}{\eta\delta}}$.

To sum up, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, $\|\tilde{\boldsymbol{\Delta}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_2 \sqrt{\eta \log \frac{2\alpha K}{\eta\delta}}$ for all $k \in [K]$, $0 \leq t \leq H$. By triangle inequality,

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \leq \frac{1}{K} \sum_{k \in [K]} \|\tilde{\boldsymbol{\Delta}}_{k,H}^{(s)}\|_2 \leq \tilde{C}_2 \sqrt{\eta \log \frac{2\alpha K}{\eta\delta}}.$$

■

The combination of Theorem 27 and Theorem 28 leads to the following lemma, which states that the Local SGD iterate will enter Γ^{ϵ_1} after s_0 rounds with high probability.

Lemma 29 *Given $\bar{\boldsymbol{\theta}}^{(0)}$ such that $\Phi(\bar{\boldsymbol{\theta}}^{(0)}) \in \Gamma$, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, there exists a positive constant $\tilde{C}_4$ such that with probability at least $1-\delta$,*

$$\|\bar{\boldsymbol{\theta}}^{(s_0)} - \boldsymbol{\phi}^{(0)}\|_2 \leq \frac{1}{4}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0 + \tilde{C}_4 \sqrt{\eta \log \frac{1}{\eta\delta}}.$$

Proof First, we prove by induction that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_1 \sigma_{\max} \sqrt{2H \log \frac{2HKs_0}{\delta}}, \qquad \forall 0 \leq t \leq H, k \in [K], 0 \leq s < s_0, \tag{33}$$

the closeness of $\bar{\boldsymbol{\theta}}^{(s)}$ and $\boldsymbol{u}_0^{(s)}$ is bounded by

$$\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{u}_0^{(s)}\|_2 \leq \sum_{l=1}^{s} \tilde{C}_1^l \left(\eta G + \tilde{C}_3\sqrt{\eta \log \frac{s_0}{\eta\delta}}\right), \qquad \forall 0 \leq s \leq s_0. \tag{34}$$

When $s = 0$, $\bar{\boldsymbol{\theta}}^{(0)} = \boldsymbol{u}_0^{(0)}$. Assume that (34) holds for round s. Then by Theorem 27, for all $0 \leq t \leq H$,

$$\begin{aligned}\|\tilde{\boldsymbol{u}}_t^{(s)} - \boldsymbol{u}_t^{(s)}\|_2 &\leq \tilde{C}_1\|\tilde{\boldsymbol{u}}_0^{(s)} - \boldsymbol{u}_0^{(s)}\|_2 + \tilde{C}_1\eta G \\ &= \tilde{C}_1\|\bar{\boldsymbol{\theta}}_0^{(s)} - \boldsymbol{u}_0^{(s)}\|_2 + \tilde{C}_1\eta G \\ &\leq \sum_{l=1}^{s} \tilde{C}_1^{l+1} \left(\eta G + \tilde{C}_3\sqrt{\eta \log \frac{s_0}{\eta\delta}}\right) + \tilde{C}_1\eta G.\end{aligned}$$

Therefore, for sufficiently small η, $\tilde{\boldsymbol{u}}_t^{(s)} \in \mathscr{Z}^\epsilon$, $\forall 0 \leq t \leq H$. Combing the above inequality with Theorem 28, we have

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}^{(s+1)} - \boldsymbol{u}_0^{(s+1)}\|_2 &= \|\bar{\boldsymbol{\theta}}^{(s+1)} - \boldsymbol{u}_H^{(s)}\|_2 \\ &\leq \|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 + \|\tilde{\boldsymbol{u}}_H^{(s)} - \boldsymbol{u}_H^{(s)}\|_2 \\ &\leq \sum_{l=1}^{s+1} \tilde{C}_1^{l+1} \left(\eta G + \tilde{C}_3\sqrt{\eta \log \frac{s_0}{\eta\delta}}\right),\end{aligned}$$

which concludes the induction.

Therefore, when (33) holds, there exists a positive constant $\tilde{C}_4$ such that

$$\|\bar{\boldsymbol{\theta}}^{(s_0)} - \boldsymbol{u}_0^{(s_0)}\|_2 \leq \tilde{C}_4\sqrt{\eta \log \frac{1}{\eta\delta}}.$$

By definition of $\boldsymbol{u}_0^{(s_0)}$,

$$\|\bar{\boldsymbol{\theta}}^{(s_0)} - \boldsymbol{\phi}^{(0)}\|_2 \leq \frac{1}{4}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0 + \tilde{C}_4\sqrt{\eta \log \frac{1}{\eta\delta}}.$$

Finally, according to Theorem 26, (33) holds with probability at least $1 - \delta$. ■

J.5.3. PROOF FOR SUBPHASE 2

In subphase 2, we show that the iterate can reach within $\tilde{\mathcal{O}}(\sqrt{\eta})$ distance from Γ after $\mathcal{O}(\log \frac{1}{\eta})$ rounds with high probability. The following lemma manifests how the potential function $\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)})$ evolves after one round.

Lemma 30 *Given $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1 - \delta$,*

$$\boldsymbol{\theta}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}, \quad \tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) \leq \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta \log \frac{1}{\eta\delta}}, \quad \forall k \in [K], 0 \leq t \leq H$$

and

$$\bar{\boldsymbol{\theta}}^{(s+1)} \in \Gamma^{\epsilon_2}, \quad \Psi(\bar{\boldsymbol{\theta}}^{(s+1)}) \leq \exp(-\alpha\mu/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta \log \frac{1}{\eta\delta}},$$

where $\tilde{C}_5$ is a positive constant.

Proof *Since $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, then for all $0 \leq t \leq H$, $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_1}$ by the definition of the working zone. By Theorem 21, for $\eta \leq \frac{1}{\rho_2}$,*

$$\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}) - \mathcal{L}^* \leq (1-\mu\eta)^t \left(\mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*\right) \leq \mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*, \quad \forall 0 \leq t \leq H.$$

Specially, for $t = H$,

$$\mathcal{L}(\tilde{\boldsymbol{u}}_H^{(s)}) - \mathcal{L}^* \leq (1-\mu\eta)^{\frac{\alpha}{\eta}} \left(\mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*\right) \leq \exp(-\alpha\mu)(\mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*).$$

Therefore,

$$\tilde{\Psi}(\tilde{\boldsymbol{u}}_H^{(s)}) \leq \exp(-\alpha\mu/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}).$$

According to the proof of Theorem 28, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_1\sigma_{\max}\sqrt{\frac{2\alpha}{\eta}\log\frac{2\alpha K}{\eta\delta}}, \qquad \forall k \in [K], 0 \leq t \leq H, \tag{35}$$

there exists a constant $\tilde{C}_3$ such that

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \tilde{\boldsymbol{u}}_t^{(s)}\|_2 \leq \tilde{C}_3\sqrt{\eta\log\frac{1}{\eta\delta}}, \quad \forall 0 \leq t \leq H, k \in [K],$$

and

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \leq \tilde{C}_3\sqrt{\eta\log\frac{1}{\eta\delta}}.$$

Since $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_1}$, $\forall 0 \leq t \leq H$, $\bar{\boldsymbol{\theta}}^{(s+1)} \in \Gamma^{\epsilon_2}$ and $\bar{\boldsymbol{\theta}}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}$, $\forall 0 \leq t \leq H$, $k \in [K]$.

By Theorem 22, $\tilde{\Psi}(\cdot)$ is $\sqrt{2\rho_2}$-Lipschitz in $\mathcal{M}^{\epsilon_4}$. Therefore, when (35) holds, there exists a constant $\tilde{C}_5 := \sqrt{2\rho_2}\tilde{C}_3$ such that

$$\begin{aligned}\tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) &\leq \tilde{\Psi}(\tilde{\boldsymbol{u}}_t^{(s)}) + \sqrt{2\rho_2}\|\boldsymbol{\theta}_{k,t}^{(s)} - \tilde{\boldsymbol{u}}_t^{(s)}\|_2 \\ &\leq \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta\log\frac{1}{\eta\delta}},\end{aligned}$$

and

$$\begin{aligned}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s+1)}) &\leq \tilde{\Psi}(\tilde{\boldsymbol{u}}_H^{(s)}) + \sqrt{2\rho_2}\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \\ &\leq \exp(-\alpha\mu/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta\log\frac{1}{\eta\delta}}.\end{aligned}$$

Finally, by Theorem 26, (35) holds with probability at least $1-\delta$. ■

We are thus led to the following lemma which characterizes the evolution of the potential $\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)})$ and $\tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)})$ over multiple rounds.

Lemma 31 *Given* $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \leq \frac{1}{2}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0$, *for* $\delta = \mathcal{O}(\mathrm{poly}(\eta))$ *and any integer* $1 \leq R \leq R_{\mathrm{tot}}$, *with probability at least* $1-\delta$,

$$\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}, \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \leq \exp(-\alpha\mu s/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) + \frac{1}{1-\exp(-\alpha\mu/2)}\tilde{C}_5\sqrt{\eta \log\frac{R}{\eta\delta}}, \forall 0 \leq s \leq R. \tag{36}$$

Furthermore,

$$\bar{\boldsymbol{\theta}}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}, \quad \tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) \leq \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}}, \quad \forall 0 \leq t \leq H, 0 \leq s < R, k \in [K]. \tag{37}$$

Proof We prove induction that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \leq \tilde{C}_1\sigma_{\max}\sqrt{\frac{2\alpha}{\eta}\log\frac{2R\alpha K}{\eta\delta}}, \qquad \forall k \in [K], 0 \leq t \leq H, 0 \leq s < R, \tag{38}$$

then for all $0 \leq s \leq R$, (36) and (37) hold.

When $s = 0$, $\bar{\boldsymbol{\theta}}^{(0)} \in \Gamma^{\epsilon_0}$ and (36) trivially holds. By Theorem 30, (37) holds. Assume that (36) and (37) hold for round $s-1$. Then for round s, by Theorem 30, $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_2}$ and

$$\begin{aligned}\Psi(\bar{\boldsymbol{\theta}}^{(s)}) &\leq \exp(-\alpha\mu/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s-1)}) + \tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}} \\ &\leq \exp(-\alpha\mu s/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) + \frac{1}{1-\exp(-\alpha\mu/2)}\tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}},\end{aligned}$$

where the second inequality comes from the induction hypothesis. By Theorem 25,

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2 &\leq \frac{2}{\sqrt{2\mu}}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \\ &\leq \frac{2}{\sqrt{2\mu}}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) + \frac{2}{\sqrt{2\mu}(1-\exp(-\alpha\mu/2))}\tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}} \\ &\leq \frac{1}{2}\epsilon_0 + \frac{2}{\sqrt{2\mu}(1-\exp(-\alpha\mu/2))}\tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}}.\end{aligned}$$

Here, the last inequality uses $\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) \leq \sqrt{\frac{\rho_2}{2}}\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(0)}\|_2 \leq \frac{1}{2}\sqrt{\frac{\mu}{2}}\epsilon_0$. Hence, when η is sufficiently small, $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$. Still by Theorem 30, $\bar{\boldsymbol{\theta}}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}$ and

$$\tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) \leq \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) + \tilde{C}_5\sqrt{\eta\log\frac{R}{\eta\delta}}.$$

Finally, according to Theorem 26, (38) holds with probability at least $1-\delta$.

■

The following corollary is a direct consequence of Theorem 31 and Theorem 25.

Corollary 32 *Let $s_1 := \lceil \frac{20}{\alpha\mu} \log \frac{1}{\eta} \rceil$. Given $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \le \frac{1}{2}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0$, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$,*

$$\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s_1)}) \le \tilde{C}_6 \sqrt{\eta \log \frac{1}{\eta\delta}}, \quad \|\bar{\boldsymbol{\theta}}^{(s_1)} - \boldsymbol{\phi}^{(s_1)}\|_2 \le \tilde{C}_6 \sqrt{\eta \log \frac{1}{\eta\delta}}, \tag{39}$$

where $\tilde{C}_6$ is a constant.

Proof Substituting in $R = s_1$ to Theorem 31 and applying $\|\bar{\boldsymbol{\theta}}^{(s_1)} - \boldsymbol{\phi}^{(s)}\|_2 \le \sqrt{\frac{2}{\mu}}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s_1)})$ for $\bar{\boldsymbol{\theta}}^{(s_1)} \in \Gamma^{\epsilon_0}$, we have the lemma. ■

Finally, we provide a high probability bound for the movement of the projection on the manifold after s_1 rounds $\|\boldsymbol{\phi}^{(s_1)} - \boldsymbol{\phi}^{(0)}\|_2$.

Lemma 33 *Let $s_1 := \lceil \frac{20}{\alpha\mu} \log \frac{1}{\eta} \rceil$. Given $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \le \frac{1}{2}\sqrt{\frac{\mu}{\rho_2}}\epsilon_0$. For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$,*

$$\|\boldsymbol{\phi}^{(s_1)} - \boldsymbol{\phi}^{(0)}\|_2 \le \tilde{C}_8 \log\frac{1}{\eta}\sqrt{\eta \log \frac{1}{\eta\delta}}.$$

Proof From Theorem 31, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \le \tilde{C}_1 \sigma_{\max}\sqrt{\frac{2\alpha}{\eta}\log\frac{2 s_1 \alpha K}{\eta\delta}}, \qquad \forall k \in [K], 0 \le t \le H, 0 \le s < s_1, \tag{40}$$

then $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, for all $0 \le s \le s_1$. By the definition of Γ^{ϵ_0}, $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_1}$, for all $0 \le t \le H, 0 \le s \le s_1$. By triangle inequality, $\|\boldsymbol{\phi}^{(s_1)} - \boldsymbol{\phi}^{(0)}\|_2$ can be decomposed as follows.

$$\begin{aligned}\|\boldsymbol{\phi}^{(s_1)} - \boldsymbol{\phi}^{(0)}\|_2 &\le \sum_{s=0}^{s_1-1} \|\boldsymbol{\phi}^{(s+1)} - \boldsymbol{\phi}^{(s)}\|_2 \\ &\le \sum_{s=0}^{s_1-1} \|\Phi(\tilde{\boldsymbol{u}}_H^{(s)}) - \Phi(\tilde{\boldsymbol{u}}_0^{(s)})\|_2 + \sum_{s=0}^{s_1-1} \|\Phi(\bar{\boldsymbol{\theta}}^{(s+1)}) - \Phi(\tilde{\boldsymbol{u}}_H^{(s)})\|_2. \end{aligned} \tag{41}$$

By Theorem 28, when (40) hold , then for all $0 \le s < s_1 - 1$,

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \le \tilde{C}_3 \sqrt{\eta \log\frac{s_1}{\eta\delta}}.$$

This implies that $\bar{\boldsymbol{\theta}}^{(s+1)} \in B^{\epsilon_1}(\tilde{\boldsymbol{u}}_H^{(s)})$. Since for all $\boldsymbol{\theta} \in \Gamma^{\epsilon_2}$, $\|\partial\Phi(\boldsymbol{\theta})\|_2 \le \nu_1$, then $\Phi(\cdot)$ is ν_1-Lipschitz in $B^{\epsilon_1}(\tilde{\boldsymbol{u}}_H^{(s)})$. This gives

$$\begin{aligned}\|\Phi(\bar{\boldsymbol{\theta}}^{(s+1)}) - \Phi(\tilde{\boldsymbol{u}}_H^{(s)})\|_2 &\le \nu_1 \|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \\ &\le \nu_1 \tilde{C}_3 \sqrt{\eta \log \frac{s_1}{\eta\delta}}. \end{aligned} \tag{42}$$

Then we analyze $\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2$. By Theorem 24 and the definition of Γ^{ϵ_0} and Γ^{ϵ_1}, there exists $\boldsymbol{\phi} \in \Gamma$ such that $\tilde{\boldsymbol{u}}_t^{(s)} \in B^{\epsilon_1}(\boldsymbol{\phi})$, $\forall 0 \leq t \leq H$. Therefore, we can expand $\Phi(\tilde{\boldsymbol{u}}_{t+1}^{(s)})$ as follows:

$$\begin{aligned}
\Phi(\tilde{\boldsymbol{u}}_{t+1}^{(s)}) &= \Phi(\tilde{\boldsymbol{u}}_t^{(s)} - \eta \nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})) \\
&= \Phi(\tilde{\boldsymbol{u}}_t^{(s)}) - \eta \partial \Phi(\tilde{\boldsymbol{u}}^{(s)}) \nabla \mathcal{L}(\boldsymbol{u}_t^{(s)}) + \frac{\eta^2}{2} \partial^2 \Phi(\hat{\boldsymbol{u}}_t^{(s)})[\nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}), \nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})] \\
&= \Phi(\tilde{\boldsymbol{u}}_t^{(s)}) + \frac{\eta^2}{2} \partial^2 \Phi \left(c_t^{(s)} \tilde{\boldsymbol{u}}_t^{(s)} + (1 - c_t^{(s)}) \tilde{\boldsymbol{u}}_{t+1}^{(s)} \right) [\nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}), \nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})],
\end{aligned}$$

where $c_t^{(s)} \in (0, 1)$. Then we have

$$\begin{aligned}
\|\Phi(\tilde{\boldsymbol{u}}_H^{(s)}) - \Phi(\tilde{\boldsymbol{u}}_0^{(s)})\|_2 &\leq \frac{\eta^2}{2} \sum_{t=0}^{H-1} \|\partial^2 \Phi(\Big(c_t^{(s)} \tilde{\boldsymbol{u}}_t^{(s)} + (1 - c_t^{(s)}) \tilde{\boldsymbol{u}}_{t+1}^{(s)}\Big))[\nabla \mathcal{L}(\tilde{\boldsymbol{u}}^{(s)}), \nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})]\|_2 \\
&\leq \frac{\eta^2}{2} \nu_2 \sum_{t=0}^{H-1} \|\nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})\|_2^2.
\end{aligned}$$

By Theorem 21, $\frac{\eta}{2} \|\nabla \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})\|_2^2 \leq \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}) - \mathcal{L}(\tilde{\boldsymbol{u}}_{t+1}^{(s)})$. Therefore,

$$\begin{aligned}
\|\Phi(\tilde{\boldsymbol{u}}_H^{(s)}) - \Phi(\tilde{\boldsymbol{u}}_0^{(s)})\|_2 &\leq \eta \nu_2 (\mathcal{L}(\tilde{\boldsymbol{u}}_0^{(s)}) - \mathcal{L}(\tilde{\boldsymbol{u}}_H^{(s)})) \\
&\leq \eta \nu_2 [\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)})]^2 \\
&\leq \nu_2 \eta \left[2 \exp(-\alpha s \mu) \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) + \frac{\tilde{C}_5^2 \eta}{(1 - \exp(-\alpha \mu / 2))^2} \log \frac{s_1}{\eta \delta} \right], \qquad (43)
\end{aligned}$$

where the last inequality uses Cauchy-Schwartz inequality and Theorem 31. Summing up (43) , we obtain

$$\begin{aligned}
\sum_{s=0}^{s_1 - 1} \|\Phi(\tilde{\boldsymbol{u}}_H^{(s)}) - \Phi(\tilde{\boldsymbol{u}}_0^{(s)})\|_2 &\leq \nu_2 \eta \left[2 \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) \sum_{s=0}^{s_1 - 1} \exp(-\alpha \mu s) + \frac{s_1 \tilde{C}_5^2 \eta}{(1 - \exp(-\alpha \mu / 2))^2} \log \frac{s_1}{\eta \delta} \right] \\
&\leq \tilde{C}_7 \eta \log \frac{1}{\eta} \log \frac{1}{\eta \delta}, \qquad (44)
\end{aligned}$$

where $\tilde{C}_7$ is a constant. Substituting (42) and (44) into (41), for sufficiently small η, we have

$$\begin{aligned}
\|\boldsymbol{\phi}^{(s_1)} - \boldsymbol{\phi}^{(0)}\|_2 &\leq \nu_1 \tilde{C}_3 s_1 \sqrt{\eta \log \frac{s_1}{\eta \delta}} + \tilde{C}_7 \eta \log \frac{1}{\eta} \log \frac{1}{\eta \delta} \\
&\leq \tilde{C}_8 \log \frac{1}{\eta} \sqrt{\eta \log \frac{1}{\eta \delta}},
\end{aligned}$$

where $\tilde{C}_8$ is a constant. Finally, according to Theorem 26, (40) holds with probability at least $1 - \delta$. ■

J.6. High Probability Bounds for Phase 2: Iterates Staying Close to Manifold

In this subsection, we show that $\|\boldsymbol{x}_{k,t}^{(s)}\|_2 = \tilde{\mathcal{O}}(\sqrt{\eta})$ and $\|\bar{\boldsymbol{\theta}}^{(s+r)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 = \tilde{\mathcal{O}}(\eta^{0.5 - 0.5\beta})$, $\forall 0 \leq r \leq R_{\mathrm{grp}}$ with high probability.

J.6.1. ADDITIONAL NOTATIONS

Before presenting the lemmas, we define the following martingale $\{\boldsymbol{m}_{k,t}^{(s)}\}_{t=0}^{H}$ that will be useful in the proof:

$$\boldsymbol{m}_{k,t}^{(s)} := \sum_{\tau=0}^{t-1} \boldsymbol{z}_{k,\tau}^{(s)}, \quad \boldsymbol{m}_{k,0} = \boldsymbol{0}.$$

We also define $\tilde{\boldsymbol{P}} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as an extension of $\partial\Phi$:

$$\tilde{\boldsymbol{P}}(\boldsymbol{\theta}) := \begin{cases} \partial\Phi(\boldsymbol{\theta}), & \text{if } \boldsymbol{\theta} \in \Gamma^{\epsilon_2}, \\ \boldsymbol{0}, & \text{otherwise.} \end{cases}$$

Finally, we define a martingale $\{\boldsymbol{Z}_t^{(s)} : s \geq 0, 0 \leq t \leq H\}$:

$$\boldsymbol{Z}_t^{(s)} := \frac{1}{K} \sum_{k\in[K]} \sum_{r=0}^{s-1} \sum_{\tau=0}^{H-1} \tilde{\boldsymbol{P}}(\bar{\boldsymbol{\theta}}^{(r)}) \boldsymbol{z}_{k,t}^{(r)} + \frac{1}{K} \sum_{k\in[K]} \sum_{\tau=0}^{t-1} \tilde{\boldsymbol{P}}(\bar{\boldsymbol{\theta}}^{(s)}) \boldsymbol{z}_{k,t}^{(s)}, \quad \boldsymbol{Z}_0^{(0)} = \boldsymbol{0}.$$

J.6.2. PROOF FOR THE HIGH PROBABILITY BOUNDS

A direct application of Azuma-Hoeffding's inequality yields the following lemma.

Lemma 34 (Concentration property of $\boldsymbol{m}_{k,t}^{(s)}$) *With probability at least $1-\delta$, the following holds:*

$$\|\boldsymbol{m}_{k,t}^{(s)}\|_2 \leq \tilde{C}_9 \sqrt{\frac{1}{\eta} \log \frac{1}{\eta\delta}}, \quad \forall 0 \leq t \leq H, k \in [K], 0 \leq s < R_{\mathrm{grp}},$$

where $\tilde{C}_9$ is a constant.

Proof Notice that $\|\boldsymbol{m}_{k,t+1}^{(s)} - \boldsymbol{m}_{k,t}^{(s)}\|_2 \leq \sigma_{\max}$. Then by Azuma-Hoeffdings inequality,

$$\mathbb{P}(\|\boldsymbol{m}_{k,t}^{(s)}\|_2 \geq \epsilon') \leq 2 \exp\left(-\frac{\epsilon'^2}{2t\sigma_{\max}^2}\right).$$

Taking union bound on K clients, H local steps and R_{grp} rounds, we obtain that the following inequality holds with probability at least $1-\delta$:

$$\|\boldsymbol{m}_{k,t}^{(s)}\|_2 \leq \sigma_{\max} \sqrt{2H \log \frac{2KHR_{\mathrm{grp}}}{\delta}}, \quad \forall 0 \leq t \leq H, k \in [K], 0 \leq s < R_{\mathrm{grp}}.$$

Substituting in $H = \frac{\alpha}{\eta}$ and $R_{\mathrm{grp}} = \lfloor \frac{1}{\alpha\eta^{\beta}} \rfloor$ yields the lemma. ■

Again applying Azuma-Hoeffding's inequality, we have the following lemma about the concentration property of $\boldsymbol{Z}_t^{(s)}$.

Lemma 35 (Concentration property of $\boldsymbol{Z}_t^{(s)}$) *With probability at least $1-\delta$, the following inequality holds:*

$$\|\boldsymbol{Z}_H^{(s)}\|_2 \leq \tilde{C}_{12} \eta^{-0.5-0.5\beta} \sqrt{\log \frac{1}{\eta\delta}}, \quad \forall 0 \leq s < R_{\mathrm{grp}}.$$

Proof Notice that $\|\boldsymbol{Z}_{t+1}^{(s)} - \boldsymbol{Z}_{t}^{(s)}\|_2 \leq \nu_2\sigma_{\max}, \forall 0 \leq t \leq H-1$ and $\|\boldsymbol{Z}_{0}^{(s+1)} - \boldsymbol{Z}_{H}^{(s)}\|_2 \leq \nu_2\sigma_{\max}$. By Azuma-Hoeffding's inequality,

$$\mathbb{P}(\|\boldsymbol{Z}_t^{(s)}\|_2 \geq \epsilon') \leq 2\exp\left(-\frac{\epsilon'^2}{2(sH+t)\nu_2^2\sigma_{\max}^2}\right).$$

Taking union bound on R_{grp} rounds, we obtain that the following inequality holds with probability at least $1-\delta$:

$$\|\boldsymbol{Z}_H^{(s)}\|_2 \leq \sigma_{\max}\nu_2\sqrt{2HR_{\mathrm{grp}}\log\frac{2R_{\mathrm{grp}}}{\delta}}, \quad \forall 0 \leq s < R_{\mathrm{grp}}.$$

Substituting in $H = \frac{\alpha}{\eta}$ and $R_{\mathrm{grp}} = \lfloor \frac{1}{\alpha\eta^{\beta}} \rfloor$ yields the lemma. ■

We proceed to present a direct corollary of Theorem 31 which provides a bound for the potential function over R_{grp} rounds.

Lemma 36 *Given $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \leq C_0\sqrt{\eta\log\frac{1}{\eta}}$ where C_0 is a constant, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$,*

$$\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}, \quad \tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \leq C_1\sqrt{\eta\log\frac{1}{\eta\delta}}, \quad \forall 0 \leq s < R_{\mathrm{grp}}, \tag{45}$$

and

$$\bar{\boldsymbol{\theta}}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}, \quad \tilde{\Psi}(\bar{\boldsymbol{\theta}}_{k,t}^{(s)}) \leq C_1\sqrt{\eta\log\frac{1}{\eta\delta}}, \quad \forall 0 \leq s < R_{\mathrm{grp}}, 0 \leq t \leq H, k \in [K], \tag{46}$$

where C_1 is a constant that can depend on C_0.

Furthermore,

$$\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})}) \leq \tilde{C}_{10}\sqrt{\eta\log\frac{1}{\eta\delta}},$$

where $\tilde{C}_9$ is a constant independent of C_0.

Proof By ρ_2-smoothness of $\mathcal{L}$, $\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) \leq C_0\sqrt{\frac{\eta\rho_2}{2}\log\frac{1}{\eta}}$. Substituting $R_{\mathrm{grp}} = \lfloor\frac{1}{\alpha\eta^{\beta}}\rfloor$ and $\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(0)}) \leq C_0\sqrt{\frac{\eta\rho_2}{2}\log\frac{1}{\eta}}$ into Theorem 31, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, (45) and (46) where C_1 is a constant that can depend on C_0.

Furthermore, for round $\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})}$,

$$\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})}) \leq \exp(-\mathcal{O}(\eta^{-\beta})) + \frac{1}{1-\exp(-\alpha\mu/2)}\tilde{C}_5\sqrt{\eta\log\frac{R_{\mathrm{grp}}}{\eta\delta}} \leq \tilde{C}_{10}\sqrt{\eta\log\frac{1}{\eta\delta}},$$

where $\tilde{C}_9$ is a constant independent of C_0. ■

Lemma 37 *Given $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \le C_0\sqrt{\eta \log \frac{1}{\eta}}$ where C_0 is a constant, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, for all $0 \le s_0 < R_{\mathrm{grp}}, 0 \le t \le H$, $k \in [K]$,*

$$\|\boldsymbol{x}_{k,t}^{(s)}\|_2 \le C_2\sqrt{\eta \log \frac{1}{\eta\delta}}, \quad \|\bar{\boldsymbol{x}}_{H}^{(s)}\|_2 \le C_2\sqrt{\eta \log \frac{1}{\eta\delta}},$$
$$\|\bar{\boldsymbol{\theta}}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \le C_2\sqrt{\eta \log \frac{1}{\eta\delta}}, \quad \|\bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \le C_2\sqrt{\eta \log \frac{1}{\eta\delta}}.$$

where C_2 is a constant that can depend C_0. Furthermore,

$$\|\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})} - \boldsymbol{\phi}^{(R_{\mathrm{grp}})}\|_2 \le \tilde{C}_{11}\sqrt{\eta \log \frac{1}{\eta\delta}},$$

where $\tilde{C}_{11}$ is a constant independent of C_0.

Proof Decomposing $\boldsymbol{x}_{k,t}^{(s)}$ by triangle inequality, we have

$$\|\boldsymbol{x}_{k,t}^{(s)}\|_2 \le \|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 + \|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2.$$

We first bound $\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2$. By Theorem 36, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1 - \frac{\delta}{2}$,

$$\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \le C_1\sqrt{\eta \log \frac{2}{\eta\delta}}, \forall 0 \le s < R_{\mathrm{grp}}, \tag{47}$$
$$\tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) \le C_1\sqrt{\eta \log \frac{2}{\eta\delta}}, \quad \forall 0 \le s < R_{\mathrm{grp}}, 0 \le t \le H, \tag{48}$$

and

$$\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})}) \le \tilde{C}_{10}\sqrt{\eta \log \frac{2}{\eta\delta}}, \tag{49}$$

where C_2 is a constant that may depend on C_0 and $\tilde{C}_{10}$ is a constant independent of C_0. When (47) and (49) hold, by Theorem 25,

$$\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2 \le \sqrt{\frac{2}{\mu}}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \le C_1\sqrt{\frac{2\eta}{\mu} \log \frac{2}{\eta\delta}}, \tag{50}$$
$$\|\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})} - \boldsymbol{\phi}^{(R_{\mathrm{grp}})}\|_2 \le \sqrt{\frac{2}{\mu}}\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(R_{\mathrm{grp}})}) \le \tilde{C}_{10}\sqrt{\frac{2\eta}{\mu} \log \frac{2}{\eta\delta}}. \tag{51}$$

Then we bound $\|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2$. By the update rule, we have

$$\boldsymbol{\theta}_{k,t}^{(s)} = \bar{\boldsymbol{\theta}}^{(s)} - \eta\sum_{\tau=0}^{t-1}\nabla\mathcal{L}(\boldsymbol{\theta}_{k,\tau}^{(s)}) - \eta\sum_{\tau=0}^{t-1}\boldsymbol{z}_{k,\tau}^{(s)} = \bar{\boldsymbol{\theta}}^{(s)} - \eta\sum_{\tau=0}^{t-1}\nabla\mathcal{L}(\boldsymbol{\theta}_{k,\tau}^{(s)}) - \eta\boldsymbol{m}_{k,t}^{(s)}.$$

Still by triangle inequality, we have

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \le \eta\sum_{\tau=0}^{t-1}\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k,\tau}^{(s)})\|_2 + \eta\|\boldsymbol{m}_{k,t}^{(s)}\|_2.$$

Due to ρ_2-smoothness of $\mathcal{L}$, when (48) holds,

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}_{k,\tau}^{(s)})\|_2 \leq \sqrt{2\rho_2}\tilde{\Psi}(\boldsymbol{\theta}_{k,\tau}^{(s)}) \leq C_1\sqrt{2\rho_2 \eta \log \frac{2}{\eta\delta}}. \tag{52}$$

By Theorem 34, with probability at least $1-\frac{\delta}{2}$,

$$\|\boldsymbol{m}_{k,t}^{(s)}\|_2 \leq \tilde{C}_9\sqrt{\frac{1}{\eta}\log\frac{2}{\eta\delta}}, \quad \forall 0 \leq t \leq H, k\in[K], 0 \leq s < R_{\mathrm{grp}}. \tag{53}$$

Combining (52) and (53), when (48) and (49) hold simultaneously, there exists a constant C_3 which can depend on C_0 such that

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \leq C_3\sqrt{\eta\log\frac{1}{\eta\delta}}, \quad \forall k\in[K], 0\leq t\leq H. \tag{54}$$

By triangle inequality,

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \leq C_3\sqrt{\eta\log\frac{1}{\eta\delta}}.$$

Combining (50), (51) and (54), we complete the proof. ■

Then we provide high probability bounds for the movement of $\boldsymbol{\phi}^{(s)}$ within R_{grp} rounds.

Lemma 38 *Given $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 \leq C_0\sqrt{\eta\log\frac{1}{\eta}}$ where C_0 is a constant, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$,*

$$\|\boldsymbol{\phi}^{(s)} - \boldsymbol{\phi}^{(0)}\|_2 \leq C_4\eta^{0.5-0.5\beta}\sqrt{\log\frac{1}{\eta\delta}}, \quad \forall 1\leq s\leq R_{\mathrm{grp}}.$$

where C_4 is a constant that can depend on C_0.

Proof By the update rule of Local SGD,

$$\boldsymbol{\theta}_{k,H}^{(s)} = \bar{\boldsymbol{\theta}}^{(s)} - \eta\sum_{t=0}^{H-1}\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \eta\sum_{t=0}^{H-1}\boldsymbol{z}_{k,t}^{(s)}$$

Averaging among K clients gives

$$\bar{\boldsymbol{\theta}}^{(s+1)} = \bar{\boldsymbol{\theta}}^{(s)} - \frac{\eta}{K}\sum_{t=0}^{H-1}\sum_{k\in[K]}\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) - \frac{\eta}{K}\sum_{t=0}^{H-1}\sum_{k\in[K]}\boldsymbol{z}_{k,t}^{(s)}.$$

By Theorem 37, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, the following holds with probability at least $1-\delta/3$,

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \leq C_2\sqrt{\eta\log\frac{3}{\eta\delta}}, \ \boldsymbol{\theta}_{k,t}^{(s)} \in B^{\epsilon_0}(\boldsymbol{\phi}^{(s)}), \ \forall 0\leq s< R_{\mathrm{grp}}, 0\leq t\leq H, k\in[K], \tag{55}$$

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2 \leq C_2\sqrt{\eta\log\frac{3}{\eta\delta}}, \quad \bar{\boldsymbol{\theta}}^{(s)}, \bar{\boldsymbol{\theta}}^{(s+1)} \in B^{\epsilon_0}(\boldsymbol{\phi}^{(s)}), \quad \forall 0\leq s< R_{\mathrm{grp}}. \tag{56}$$

When (55) and (56) hold, we can expand $\Phi(\bar{\boldsymbol{\theta}}^{(s+1)})$ as follows:

$$\begin{aligned}\boldsymbol{\phi}^{(s+1)} &= \boldsymbol{\phi}^{(s)} + \partial\Phi(\bar{\boldsymbol{\theta}}^{(s)})(\bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}) + \frac{1}{2}\partial^2\Phi(\tilde{\boldsymbol{\theta}}^{(s)})[\bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}, \bar{\boldsymbol{\theta}}^{(s+1)} - \bar{\boldsymbol{\theta}}^{(s)}] \\ &= \boldsymbol{\phi}^{(s)} \underbrace{-\frac{\eta}{K}\sum_{t=0}^{H-1}\sum_{k\in[K]}\partial\Phi(\bar{\boldsymbol{\theta}}^{(s)})\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)})}_{\mathcal{T}_1^{(s)}} \underbrace{-\frac{\eta}{K}\partial\Phi(\bar{\boldsymbol{\theta}}^{(s)})\sum_{t=0}^{H-1}\sum_{k\in[K]}\boldsymbol{z}_{k,t}^{(s)}}_{\mathcal{T}_2^{(s)}} \\ &\quad + \underbrace{\frac{1}{2}\partial^2\Phi(a^{(s)}\bar{\boldsymbol{\theta}}^{(s)} + (1-a^{(s)})\bar{\boldsymbol{\theta}}^{(s+1)})[\boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)}]}_{\mathcal{T}_3^{(s)}},\end{aligned}$$

where $a^{(s)} \in (0,1)$. Telescoping from round 0 to $s-1$, we have

$$\|\boldsymbol{\phi}^{(s)} - \boldsymbol{\phi}^{(0)}\|_2 = \sum_{r=0}^{s-1}\mathcal{T}_1^{(r)} + \sum_{r=0}^{s-1}\mathcal{T}_2^{(r)} + \sum_{r=0}^{s-1}\mathcal{T}_3^{(r)}.$$

From (56), we can bound $\|\mathcal{T}_3^{(s)}\|_2$ by $\|\mathcal{T}_3^{(s)}\|_2 \le \frac{1}{2}\nu_2 C_2^2\eta\log\frac{3}{\eta\delta}$. We proceed to bound $\|\mathcal{T}_1^{(s)}\|_2$. When (55) and (56) hold, we have

$$\begin{aligned}\partial\Phi(\bar{\boldsymbol{\theta}}^{(s)})\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) &= \partial\Phi(\boldsymbol{\theta}_{k,t}^{(s)})\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)}) + \partial^2\Phi(\hat{\boldsymbol{\theta}}_{k,t}^{(s)})[\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}, \nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)})] \\ &= \partial^2\Phi(b_{k,t}^{(s)}\bar{\boldsymbol{\theta}}^{(s)} + (1-b_{k,t}^{(s)})\hat{\boldsymbol{\theta}}_{k,t}^{(s)})[\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}, \nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)})],\end{aligned}$$

where $b_{k,t}^{(s)} \in (0,1)$. By Theorem 31, with probability at least $1-\delta/3$, the following holds:

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)})\|_2 \le \sqrt{2\rho_2}\tilde{\Psi}(\boldsymbol{\theta}_{k,t}^{(s)}) \le C_1\sqrt{2\rho_2\eta\log\frac{3}{\eta\delta}}, \forall k\in[K], 0\le t\le H, 0\le s < R_{\mathrm{grp}}. \quad (57)$$

When (55), (56) and (57) hold simultaneously, we have for all $0 \le s < R_{\mathrm{grp}}$,

$$\begin{aligned}\|\mathcal{T}_1^{(s)}\|_2 &\le \frac{\eta\nu_2}{K}\sum_{t=0}^{H-1}\|\boldsymbol{\theta}_{k,t}^{(s)} - \bar{\boldsymbol{\theta}}^{(s)}\|_2\|\nabla\mathcal{L}(\boldsymbol{\theta}_{k,t}^{(s)})\|_2 \\ &\le \frac{\alpha\nu_2\sqrt{2\rho_2}C_1C_2}{K}\eta\log\frac{3}{\eta\delta}.\end{aligned}$$

Finally, we bound $\|\sum_{r=0}^{s-1}\mathcal{T}_2^{(r)}\|_2$. By Theorem 35, the following inequality holds with probability at least $1-\delta/3$:

$$\|\boldsymbol{Z}_H^{(s)}\|_2 \le \tilde{C}_{12}\eta^{-0.5-0.5\beta}\sqrt{\log\frac{3}{\eta\delta}}, \quad \forall 0\le s < R_{\mathrm{grp}}. \quad (58)$$

When (55), (56) and (58) hold simultaneously, we have

$$\|\sum_{r=0}^{s}\mathcal{T}_2^{(r)}\|_2 = \eta\|\boldsymbol{Z}_H^{(s)}\|_2 \le \tilde{C}_{12}\eta^{0.5-0.5\beta}\sqrt{\log\frac{3}{\eta\delta}}, \quad \forall 0\le s < R_{\mathrm{grp}}$$

Combining the bounds for $\|\mathcal{T}_1^{(s)}\|_2$, $\|\sum_{r=0}^{s}\mathcal{T}_2^{(r)}\|_2$ and $\|\mathcal{T}_3^{(s)}\|_2$ and taking union bound, we obtain that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, the following inequality holds with probability at least $1-\delta$:

$$\|\boldsymbol{\phi}^{(s)} - \boldsymbol{\phi}^{(0)}\|_2 \le C_4 \eta^{0.5-0.5\beta}\sqrt{\log\frac{1}{\eta\delta}}, \quad \forall 1 \le s \le R_{\mathrm{grp}}.$$

where C_4 is a constant that can depend on C_0. ■

J.7. Summary of High Probability Bounds

Based on the results in Appendix J.5 and Appendix J.6, we summarize the dynamics of Local SGD iterates in this subsection. For convenience, we first introduce the definition of **global step** and **δ-good step**.

Definition 39 (Global step) *Define an index set $\mathcal{I} := \{(s,t) \mid s \ge 0, 0 \le t \le H\}$ with lexicographical order, which means $(s_1,t_1) \preceq (s_2,t_2)$ if and only if $s_1 < s_2$ or ($s_1 = s_2$ and $t_1 \le t_2$). A global step is indexed by (s,t) which corresponds to the t-th local step at round s.*

Definition 40 (δ-good step) *In the training process of Local SGD, we say the global step $(s,t) \preceq (R_{\mathrm{tot}}, 0)$ is δ-good if the following inequalities hold:*

$$\|\tilde{\boldsymbol{Z}}_{k,\tau}^{(r)}\|_2 \le \exp(\alpha\rho_2)\sigma_{\max}\sqrt{2H\log\frac{6HR_{\mathrm{tot}}K}{\delta}}, \qquad \forall k \in [K], (r,\tau) \preceq (s,t),$$
$$\|\boldsymbol{m}_{k,\tau}^{(r)}\|_2 \le \sigma_{\max}\sqrt{2H\log\frac{6KHR_{\mathrm{tot}}}{\delta}}, \qquad \forall k \in [K], (r,\tau) \preceq (s,t),$$
$$\|\boldsymbol{Z}_H^{(r)}\|_2 \le \sigma_{\max}\nu_2\sqrt{2HR_{\mathrm{grp}}\log\frac{2R_{\mathrm{tot}}}{\delta}}, \qquad \forall 0 \le r < s.$$

Applying the concentration properties of $\tilde{\boldsymbol{Z}}_{k,\tau}^{(r)}$, $\boldsymbol{m}_{k,\tau}^{(r)}$ and $\boldsymbol{Z}_H^{(r)}$ (Lemmas 35, 34 and 26) yields the following theorem.

Theorem 41 *For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, all global steps $(s,t) \preceq (R_{\mathrm{tot}}, 0)$ are δ-good.*

In the remainder of this subsection, we use $\mathcal{O}(\cdot)$ notation to hide constants independent of δ and η.

Now we are ready to present a summary of the dynamics of Local SGD when $\bar{\boldsymbol{\theta}}^{(0)}$ is initialized such that $\Phi(\bar{\boldsymbol{\theta}}^{(0)}) \in \Gamma$ and all global steps are δ-good. Phase 1 lasts for $s_0 + s_1 = \mathcal{O}(\log\frac{1}{\eta})$ rounds. At the end of phase 1, the iterate reaches within $\mathcal{O}(\sqrt{\eta\log\frac{1}{\eta\delta}})$ from Γ, i.e., $\|\bar{\boldsymbol{\theta}}^{(s_0+s_1)} - \boldsymbol{\phi}^{(s_0+s_1)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta\delta}})$. The movement of the projection on manifold over s_0+s_1 rounds, $\|\boldsymbol{\phi}^{(s_1+s_0)} - \boldsymbol{\phi}^{(0)}\|_2$, is bounded by $\mathcal{O}(\log\frac{1}{\eta}\sqrt{\eta\log\frac{1}{\eta\delta}})$.

After s_0+s_1 rounds, the dynamic enters phase 2 when the iterates stay close to Γ with $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_2}, \forall s_0+s_1 \le s \le R_{\mathrm{tot}}$ and $\boldsymbol{\theta}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}, \forall k \in [K], (s_0+s_1, 0) \preceq (s,t) \preceq (R_{\mathrm{tot}}, 0)$. Furthermore, $\|\boldsymbol{x}_{k,t}^{(s)}\|_2$ and $\|\bar{\boldsymbol{x}}_H^{(s)}\|_2$ satisfy the following equations:

$$\|\boldsymbol{x}_{k,t}^{(s)}\|_2 = \mathcal{O}(\sqrt{\eta\log\tfrac{1}{\eta\delta}}), \qquad \forall k \in [K], 0 \le t \le H, s_0+s_1 \le s < R_{\mathrm{tot}},$$
$$\|\bar{\boldsymbol{x}}_H^{(s)}\|_2 = \mathcal{O}(\sqrt{\eta\log\tfrac{1}{\eta\delta}}), \qquad \forall s_0+s_1 \le s < R_{\mathrm{tot}}.$$

Moreover, the movement of the manifold projection within R_{grp} rounds can be bounded as follows:

$$\|\boldsymbol{\phi}^{(s+r)} - \boldsymbol{\phi}^{(s)}\|_2 = \mathcal{O}(\eta^{0.5-0.5\beta}\sqrt{\log\frac{1}{\eta\delta}}), \quad \forall 1 \le r \le R_{\mathrm{grp}}.$$

Finally, we provide a theorem which states that $\bar{\boldsymbol{\theta}}^{(s)}$ stays within $\tilde{\mathcal{O}}(\sqrt{\eta})$ from the manifold after $\mathcal{O}(\log\frac{1}{\eta})$ rounds with high probability. This theorem is a direct consequence of the lemmas in Appendix J.5 and J.6.

Theorem 42 *For $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, for all $\mathcal{O}(\log\frac{1}{\eta}) \le s \le \lfloor T/(H\eta^2)\rfloor$,*

$$\Phi(\bar{\boldsymbol{\theta}}^{(s)}) \in \Gamma, \qquad \|\bar{\boldsymbol{\theta}}^{(s)} - \Phi(\bar{\boldsymbol{\theta}}^{(s)})\|_2 = \mathcal{O}\left(\sqrt{\eta\log\frac{1}{\eta\delta}}\right),$$

where $\mathcal{O}(\cdot)$ hides constants independent of η and δ.

Proof [Proof for Theorem 42] By Lemmas 29, 37 and Theorem 32, for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, when all global steps are δ-good, $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_2}, \forall s_0 + s_1 \le s \le R_{\mathrm{tot}}$ and $\boldsymbol{\theta}_{k,t}^{(s)} \in \Gamma^{\epsilon_2}, \forall k \in [K], (s_0+s_1, 0) \preceq (s,t) \preceq (R_{\mathrm{tot}}, 0)$ and $\|\boldsymbol{x}_{k,t}^{(s)}\|_2$, $\|\bar{\boldsymbol{x}}_H^{(s)}\|_2$ satisfy the following equations:

$$\|\boldsymbol{x}_{k,t}^{(s)}\|_2 = \mathcal{O}(\sqrt{\eta\log\tfrac{1}{\eta\delta}}), \qquad \forall k \in [K], 0 \le t \le H, s_0+s_1 \le s < R_{\mathrm{tot}},$$
$$\|\bar{\boldsymbol{x}}_H^{(s)}\|_2 = \mathcal{O}(\sqrt{\eta\log\tfrac{1}{\eta\delta}}), \qquad \forall s_0+s_1 \le s < R_{\mathrm{tot}}.$$

Hence $\|\bar{\boldsymbol{x}}_0^{(R_{\mathrm{tot}})}\|_2 = \mathcal{O}(\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(R_{\mathrm{tot}})})) = \mathcal{O}(\|\bar{\boldsymbol{x}}_H^{(R_{\mathrm{tot}}-1)}\|_2) = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta\delta}})$ by smoothness of $\mathcal{L}$ and Theorem 25. According to Theorem 41, with probability at least $1-\delta$, all global steps are δ-good, thus completing the proof. ■

J.8. Proof of Theorem 5

In this subsection, we explicitly derive the dependency of the approximation error on α. The proofs are quite similar to those in Appendix J.5 and hence we only state the key proof idea for brevity. With the same method as the proofs in Appendix J.5.2, we can show that with high probability, $\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2 \le \frac{1}{2}\sqrt{\frac{\mu}{\rho_2}}$ after $s_0' = \mathcal{O}(1)$ rounds. Below we focus on the dynamics of Local SGD thereafter. We first remind the readers of the definition of $\{\tilde{\boldsymbol{Z}}_{k,t}^{s}\}$:

$$\tilde{\boldsymbol{Z}}_{k,t}^{(s)} := \sum_{\tau=0}^{t-1}\left(\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\right)\boldsymbol{z}_{k,\tau}^{(s)}, \qquad \tilde{\boldsymbol{Z}}_{k,0}^{(s)} = \boldsymbol{0}.$$

We have the following lemma that controls the norm of the matrix product $\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))$.

Lemma 43 *Given $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, then there exists a positive constant C_3' independent of α such that for all $0 \le \tau < t \le H$,*

$$\|\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\|_2 \le C_3'.$$

Proof Since $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, then $\tilde{\boldsymbol{u}}_t^{(s)} \in \Gamma^{\epsilon_1}$ for all $0 \le t \le H$. We first bound the minimum eigenvalue of $\nabla^2 \mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})$. Due to the PL condition, by Theorem 21, for $\eta \le \frac{1}{\rho_2}$,

$$\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}) - \mathcal{L}^* \le (1-\mu\eta)^t \left(\mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*\right) \le \exp(-\mu t\eta)(\mathcal{L}(\bar{\boldsymbol{\theta}}^{(s)}) - \mathcal{L}^*), \quad \forall 0 \le t \le H.$$

Therefore,

$$\tilde{\Psi}(\tilde{\boldsymbol{u}}_t^{(s)}) \le \exp(-\mu t\eta/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}).$$

Let $C_1' = \rho_3\sqrt{\frac{\rho_2}{\mu}}$. By Weyl's inequality,

$$\begin{aligned}
|\lambda_{\min}(\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}))| &= |\lambda_{\min}(\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)})) - \lambda_{\min}(\nabla^2\mathcal{L}(\Phi(\tilde{\boldsymbol{u}}_t^{(s)}))| \\
&\le \rho_3\|\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_t^{(s)}) - \nabla^2\mathcal{L}(\Phi(\tilde{\boldsymbol{u}}_t^{(s)}))\|_2 \\
&\le \rho_3\|\tilde{\boldsymbol{u}}_t^{(s)} - \Phi(\tilde{\boldsymbol{u}}_t^{(s)})\|_2 \\
&\le \rho_3\sqrt{\frac{2}{\mu}}\exp(-\mu t\eta/2)\tilde{\Psi}(\bar{\boldsymbol{\theta}}^{(s)}) \\
&\le C_1'\exp(-\mu t\eta/2)\epsilon_0,
\end{aligned}$$

where the last two inequalities use Lemmas 25 and 22 respectively. Therefore, for all $0 \le t \le H$ and $0 \le \tau \le t-1$,

$$\begin{aligned}
\|\prod_{l=\tau+1}^{t-1}(\boldsymbol{I} - \eta\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)}))\|_2 &\le \prod_{l=\tau+1}^{t-1}(1+\eta|\lambda_{\min}\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)})|) \\
&\le \prod_{l=0}^{\infty}(1+\eta|\lambda_{\min}\nabla^2\mathcal{L}(\tilde{\boldsymbol{u}}_l^{(s)})|) \\
&\le \exp(\eta\epsilon_0 C_1'\sum_{l=0}^{\infty}\exp(-\mu l\eta/2)). \qquad (59)
\end{aligned}$$

For sufficiently small η, there exists a constant C_2' such that

$$\sum_{l=0}^{\infty}\exp(-\mu l\eta/2)) = \frac{1}{1-\exp(-\mu\eta/2)} \le \frac{C_2'}{\eta}. \qquad (60)$$

Substituting (60) into (59), we obtain the lemma. ■

Based on Theorem 43, we obtain the following lemma about the concentration property of $\tilde{\boldsymbol{Z}}_{k,t}^{(s)}$, which can be derived in the same way as Theorem 26.

Lemma 44 *Given $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_0}$, then with probability at least $1-\delta$,*

$$\|\tilde{\boldsymbol{Z}}_{k,t}^{(s)}\|_2 \le C_3'\sigma_{\max}\sqrt{\frac{2\alpha}{\eta}\log\frac{2\alpha K}{\eta\delta}}, \qquad \forall 0 \le t \le H, k \in [K],$$

where C_3' is defined in Theorem 43.

The following lemma can be derived analogously to Theorem 28 but the error bound is tighter in terms of its dependency on α.

Lemma 45 *Given $\bar{\boldsymbol{\theta}}^{(s)} \in \Gamma^{\epsilon_1}$, then for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$, with probability at least $1-\delta$, there exists a constant C_4' independent of α such that*

$$\|\boldsymbol{\theta}_{k,t}^{(s)} - \tilde{\boldsymbol{u}}_t^{(s)}\|_2 \le C_4' \sqrt{\alpha\eta \log \frac{\alpha}{\eta\delta}}, \quad \forall 0 \le t \le H, k \in [K],$$

and

$$\|\bar{\boldsymbol{\theta}}^{(s+1)} - \tilde{\boldsymbol{u}}_H^{(s)}\|_2 \le C_4' \sqrt{\alpha\eta \log \frac{\alpha}{\eta\delta}}.$$

Then, similar to Theorem 31, we can show that for $\delta = \mathcal{O}(\mathrm{poly}(\eta))$ and simultaneously all $s \ge s_0' + s_1'$ where $s_1' = \mathcal{O}(\frac{1}{\alpha}\log\frac{1}{\eta})$, it holds with probability at least $1-\delta$ that $\|\bar{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\phi}^{(s)}\|_2 = \mathcal{O}(\sqrt{\alpha\eta\log\frac{\alpha}{\eta\delta}})$. Note that to eliminate the dependency of the second term's denominator on α in (37), we can discuss the cases of $\alpha > c_0$ and $\alpha < c_0$ respectively where c_0 can be an arbitrary positive constant independent of α. For the case of $\alpha < c_0$ group $\lceil \frac{c_0}{\alpha} \rceil$ rounds together and repeat the arguments in this subsection to analyze the closeness between Local SGD and GD iterates as well as the evolution of loss.

J.9. Computing the Moments for Phase 2

In this subsection, we compute the first and second moments for the movement of manifold projection after R_{grp} rounds of Local SGD. Since the randomness in training might drive the iterate out of the working zone, making the dynamic intractable, we analyze a more well-behaved sequence $\{\hat{\boldsymbol{\theta}}_{k,t}^{(s)} : (s,t) \preceq (R_{\mathrm{tot}}, 0), k \in [K]\}$ which is equal to $\{\hat{\boldsymbol{\theta}}_{k,t}^{(s)}\}$ with high probability. Specifically, $\hat{\boldsymbol{\theta}}_{k,t}^{(s)}$ equal to $\boldsymbol{\theta}_{k,t}^{(s)}$ if the global step (s,t) is η^{100}-good and is set as a point $\boldsymbol{\phi}_{\mathrm{null}} \in \Gamma$ otherwise. Denote by $\mathcal{E}_t^{(s)}$ the event $\{$global step (s,t) is η^{100}-good$\}$. Then $\hat{\boldsymbol{\theta}}_{k,t}^{(s)} = \boldsymbol{\theta}_{k,t}^{(s)} \mathbb{1}_{\mathcal{E}_t^{(s)}} + \boldsymbol{\phi}_{\mathrm{null}} \mathbb{1}_{\bar{\mathcal{E}}_t^{(s)}}$ and $\{\hat{\boldsymbol{\theta}}_{k,t}^{(s)}\}$ satisfy the following update rule:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{k,t+1}^{(s)} &= \boldsymbol{\theta}_{k,t+1}^{(s)} \mathbb{1}_{\mathcal{E}_{t+1}^{(s)}} + \boldsymbol{\phi}_{\mathrm{null}} \mathbb{1}_{\bar{\mathcal{E}}_{t+1}^{(s)}} &(61)\\ &= \hat{\boldsymbol{\theta}}_{k,t}^{(s)} - \eta\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{k,t}^{(s)}) - \eta \boldsymbol{z}_{k,t}^{(s)} \underbrace{- \mathbb{1}_{\bar{\mathcal{E}}_{t+1}^{(s)}} (\hat{\boldsymbol{\theta}}_{k,t}^{(s)} - \eta\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}_{k,t}^{(s)}) - \eta \boldsymbol{z}_{k,t}^{(s)}) + \mathbb{1}_{\bar{\mathcal{E}}_{t+1}^{(s)}} \boldsymbol{\phi}_{\mathrm{null}}}_{:\hat{\boldsymbol{e}}_{k,t}^{(s)}}. &(62) \end{aligned}$$

By Theorem 41, with probability at least $1-\eta^{100}$, $\hat{\boldsymbol{\theta}}_{k,t}^{(s)} = \boldsymbol{\theta}_{k,t}^{(s)}, \forall k \in [K], (s,t) \preceq (R_{\mathrm{tot}}, 0)$. Similar to $\{\boldsymbol{\theta}_{k,t}^{(s)}\}$, we define the following variables with respect to $\{\hat{\boldsymbol{\theta}}_{k,t}^{(s)}\}$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{avg}}^{(s+1)} := \frac{1}{K}\sum_{k\in[K]} \hat{\boldsymbol{\theta}}_{k,H}^{(s)}, \quad \hat{\boldsymbol{\phi}}^{(s)} := \Phi(\hat{\boldsymbol{\theta}}_{\mathrm{avg}}^{(s)}),$$

$$\hat{\boldsymbol{x}}_{k,t}^{(s)} := \hat{\boldsymbol{\theta}}_{k,t}^{(s)} - \hat{\boldsymbol{\phi}}^{(s)}, \quad \hat{\boldsymbol{x}}_{\mathrm{avg},0}^{(s)} := \hat{\boldsymbol{\theta}}_{\mathrm{avg}}^{(s)} - \hat{\boldsymbol{\phi}}^{(s)}, \quad \hat{\boldsymbol{x}}_{\mathrm{avg},H}^{(s)} := \frac{1}{K}\sum_{k\in[K]} \hat{\boldsymbol{x}}_{k,H}^{(s)}.$$

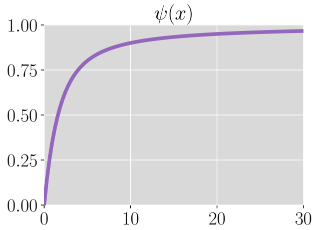


Figure 7: A plot of $\psi(x)$

Notice that $\hat{\boldsymbol{x}}_{k,0}^{(s)} = \hat{\boldsymbol{x}}_{\mathrm{avg},0}^{(s)}$ for all $k \in [K]$. Finally, we introduce the following mapping $\boldsymbol{\Psi}(\boldsymbol{\theta}) : \Gamma \to \mathbb{R}^{d\times d}$, which is closely related to $\widehat{\boldsymbol{\Psi}}$ defined in Theorem 4.

Definition 46 *For $\boldsymbol{\theta} \in \Gamma$, we define the mapping $\boldsymbol{\Psi}(\boldsymbol{\theta}) : \Gamma \to \mathbb{R}^{d\times d}$:*

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \sum_{i,j\in[d]} \psi(\eta H(\lambda_i + \lambda_j)) \left\langle \boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{v}_i \boldsymbol{v}_j^\top \right\rangle \boldsymbol{v}_i \boldsymbol{v}_j^\top,$$

where $\lambda_i, \boldsymbol{v}_i$ are the i-th eigenvalue and eigenvector of $\nabla^2\mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{v}_i$'s form an orthonormal basis of $\mathbb{R}^d$. Additionally, $\psi(x) := \frac{e^{-x}-1+x}{x}$ and $\psi(0) = 0$; see Figure 7 for a plot.

Remark 47 *Intuitively, $\boldsymbol{\Psi}(\boldsymbol{\theta})$ rescales the entries of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ in the eigenbasis of $\nabla^2\mathcal{L}(\boldsymbol{\theta})$. When $\nabla^2\mathcal{L}(\boldsymbol{\theta}) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d\times d}$, where $\lambda_i = 0$ for all $m < i \leq d$, $\Psi(\boldsymbol{\Sigma}_0)_{i,j} = \psi(\eta H(\lambda_i + \lambda_j))\Sigma_{0,i,j}$. Note that $\boldsymbol{\Psi}(\boldsymbol{\theta})$ can also be written as*

$$\mathrm{vec}(\boldsymbol{\Psi}(\boldsymbol{\theta})) = \psi(\eta H(\nabla^2\mathcal{L}(\boldsymbol{\theta}) \oplus \nabla^2\mathcal{L}(\boldsymbol{\theta})))\mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\theta})),$$

where $\oplus$ denotes the Kronecker sum $\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I}_d + \boldsymbol{I}_d \otimes \boldsymbol{B}$, $\mathrm{vec}(\cdot)$ is the vectorization operator of a matrix and $\psi(\cdot)$ is interpreted as a matrix function.

Now we are ready to present the result about the moments of $\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(s)}$.

Theorem 48 *For $s_0 + s_1 \leq s \leq R_{\mathrm{tot}} - R_{\mathrm{grp}}$ and $0 < \beta < 0.5$, the first and second moments of $\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(s)}$ are as follows:*

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(s)} \mid \hat{\boldsymbol{\phi}}^{(s)}, \mathcal{E}_{s,0}^{(s)}] = \frac{\eta^{1-\beta}}{2B}&\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(s)})[\boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}^{(s)}) + (K-1)\Psi(\hat{\boldsymbol{\phi}}^{(s)})] \\ &+ \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta),\end{aligned} \tag{63}$$

$$\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(s)})^\top \mid \hat{\boldsymbol{\phi}}^{(s)}, \mathcal{E}_0^{(s)}] = \frac{\eta^{1-\beta}}{B}\boldsymbol{\Sigma}_{\parallel}(\hat{\boldsymbol{\phi}}^{(s)}) + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta), \tag{64}$$

where $\tilde{\mathcal{O}}(\cdot)$ hides log terms and constants independent of η.

Remark 49 *By Theorem 41 and the definition of $\hat{\boldsymbol{\theta}}_{k,t}^{(s)}$, (63) and (64) still hold when we replace $\hat{\boldsymbol{\phi}}^{(s)}$ with $\boldsymbol{\phi}^{(s)}$ and replace $\hat{\boldsymbol{\phi}}^{(s+R_{\mathrm{grp}})}$ with $\boldsymbol{\phi}^{(s+R_{\mathrm{grp}})}$.*

We shall have Theorem 48 if we prove the following theorem, which directly gives Theorem 48 with a simple shift of index. For brevity, denote by $\Delta\hat{\boldsymbol{\phi}}^{(s)} := \hat{\boldsymbol{\phi}}^{(s)} - \hat{\boldsymbol{\phi}}^{(0)}$, $\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}^{(0)})$, $\boldsymbol{\Sigma}_{0,\parallel} := \boldsymbol{\Sigma}_{\parallel}(\hat{\boldsymbol{\phi}}^{(0)})$.

Theorem 50 *Given $\|\hat{\boldsymbol{\theta}}^{(0)}_{\mathrm{avg}} - \hat{\boldsymbol{\phi}}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$, for $0 < \beta < 0.5$, the first and second moments of $\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}$ are as follows:*

$$\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}] = \frac{\eta^{1-\beta}}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_0 + (K-1)\boldsymbol{\Psi}(\hat{\boldsymbol{\phi}}^{(0)})] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta),$$
$$\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})\top}] = \frac{\eta^{1-\beta}}{B}\boldsymbol{\Sigma}_{0,\parallel} + \tilde{\mathcal{O}}(\eta^{1.5-1.5\beta}) + \tilde{\mathcal{O}}(\eta).$$

We will prove Theorem 50 in the remainder of this subsection. For convenience, we introduce more notations that will be used throughout the proof. Let $\boldsymbol{H}_0 := \nabla^2\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})$. By Assumption 3.2, $\mathrm{rank}(\boldsymbol{H}_0) = m$. WLOG, assume $\boldsymbol{H}_0 = \mathrm{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d\times d}$, where $\lambda_i = 0$ for all $m < i \le d$ and $\lambda_1 \ge \lambda_2 \cdots \ge \lambda_m$. By Theorem 17, $\partial\Phi(\hat{\boldsymbol{\phi}}^{(0)})$ is the projection matrix onto the tangent space $T_{\hat{\boldsymbol{\phi}}^{(0)}}(\Gamma)$ (i.e. the null space of $\nabla^2\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})$) and therefore, $\partial\Phi(\hat{\boldsymbol{\phi}}^{(0)}) = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}_{d-m}\end{bmatrix}$. Let $\boldsymbol{P}_{\parallel} := \partial\Phi(\hat{\boldsymbol{\phi}}^{(0)})$ and $\boldsymbol{P}_{\perp} := \boldsymbol{I}_d - \boldsymbol{P}_{\parallel}$.

Let $\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}} := \mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}]$, $\hat{\boldsymbol{q}}^{(s)}_t := \mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{k,t}]$ and $\hat{\boldsymbol{B}}^{(s)}_t := \mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{k,t}\Delta\hat{\boldsymbol{\phi}}^{(s)}]$. The latter two notations are independent of k since for all $k \in [K]$, $\hat{\boldsymbol{\theta}}^{(s)}_{k,t}$'s are identically distributed. The following lemma computes the first and second moments of the one round movement.

Lemma 51 *Given $\|\hat{\boldsymbol{\theta}}^{(0)}_{\mathrm{avg}} - \hat{\boldsymbol{\phi}}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$, for $0 \le s < R_{\mathrm{grp}}$, the first and second moments of $\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}$ are as follows:*

$$\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] = \boldsymbol{P}_{\parallel}\hat{\boldsymbol{q}}^{(s)}_H + \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}^{(s)}_H] + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), \tag{65}$$
$$\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] = \boldsymbol{P}_{\parallel}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\parallel} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}). \tag{66}$$

Proof By Taylor expansion, we have

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{(s+1)} &= \Phi\left(\hat{\boldsymbol{\phi}}^{(s)} + \hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\right)\\ &= \hat{\boldsymbol{\phi}}^{(s)} + \partial\Phi(\hat{\boldsymbol{\phi}}^{(s)})\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H} + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(s)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}] + \mathcal{O}(\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3)\\ &= \hat{\boldsymbol{\phi}}^{(s)} + \partial\Phi(\hat{\boldsymbol{\phi}}^{(0)} + \Delta\hat{\boldsymbol{\phi}}^{(s)})\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H} + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)} + \Delta\hat{\boldsymbol{\phi}}^{(s)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}]\\ &\quad + \mathcal{O}(\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3)\\ &= \hat{\boldsymbol{\phi}}^{(s)} + \boldsymbol{P}_{\parallel}\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H} + \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}] + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}]\\ &\quad + \mathcal{O}(\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2^2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2 + \|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2 + \|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3).\end{aligned}$$

Rearrange the terms and we obtain:

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)} = \boldsymbol{P}_{\parallel}\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H} + \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}] + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}]\\ + \mathcal{O}(\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2^2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2 + \|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2 + \|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3).\end{aligned} \tag{67}$$

Moreover,

$$(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top = \boldsymbol{P}_\parallel \hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},H}\boldsymbol{P}_\parallel + \mathcal{O}(\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2). \tag{68}$$

Noticing that $\hat{\boldsymbol{x}}^{(s)}_{k,H}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}$ are identically distributed for all $k \in [K]$, we have $\mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}] = \frac{1}{K}\sum_{k\in[K]}\mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{k,H}\Delta\hat{\boldsymbol{\phi}}^{(s)\top}] = \hat{\boldsymbol{B}}^{(s)}_H$. Then taking expectation of both sides of (67) gives

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] &= \boldsymbol{P}_\parallel \hat{\boldsymbol{q}}^{(s)}_H + \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}^{(s)}_H] + \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}] \\ &\quad + \mathcal{O}(\mathbb{E}[\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2^2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2] + \mathbb{E}[\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2] + \mathbb{E}[\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3]).\end{aligned}$$

Again taking expectation of both sides of (68) yields

$$\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \Delta\hat{\boldsymbol{\phi}}^{(s)\top})] = \boldsymbol{P}_\parallel \hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_\parallel + \mathcal{O}(\mathbb{E}[\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2]).$$

By Lemmas 37 and 38, the following holds simultaneously with probability at least $1 - \eta^{100}$:

$$\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2 = \tilde{\mathcal{O}}(\eta^{0.5-0.5\beta}), \quad \|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2 = \tilde{\mathcal{O}}(\eta^{0.5}).$$

Furthermore, since for all $k \in [K]$ and $(s,t) \preceq (R_{\mathrm{tot}}, 0)$, $\hat{\boldsymbol{\theta}}^{(s)}_{k,t}$ stays in Γ^{ϵ_2} which is a bounded set, $\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2$ and $\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2$ are also bounded. Therefore, we have

$$\mathbb{E}[\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2^2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2] = \tilde{\mathcal{O}}(\eta^{1.5-\beta}), \tag{69}$$

$$\mathbb{E}[\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^2] = \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), \tag{70}$$

$$\mathbb{E}[\|\hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},H}\|_2^3] = \tilde{\mathcal{O}}(\eta^{1.5}), \tag{71}$$

which concludes the proof. ■

We compute $\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}$, $\hat{\boldsymbol{q}}^{(s)}_t$ and $\hat{\boldsymbol{B}}^{(s)}_t$ by solving a set of recursions, which is formulated in the following lemma. Additionally, define $\hat{\boldsymbol{A}}^{(s)}_t := \mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{k,t}\hat{\boldsymbol{x}}^{(s)\top}_{k,t}]$ and $\hat{\boldsymbol{M}}^{(s)}_t := \mathbb{E}[\hat{\boldsymbol{x}}^{(s)}_{k,t}\hat{\boldsymbol{x}}^{(s)}_{k,l}]$, $(k \neq l)$.

Lemma 52 *Given $\|\hat{\boldsymbol{\theta}}^{(0)}_{\mathrm{avg}} - \hat{\boldsymbol{\phi}}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$, for $0 \le s < R_{\mathrm{grp}}$ and $0 \le t < H$, we have the following recursions.*

$$\hat{\boldsymbol{q}}^{(s)}_{t+1} = \hat{\boldsymbol{q}}^{(s)}_t - \eta\boldsymbol{H}_0\hat{\boldsymbol{q}}^{(s)}_t - \eta\nabla^3\mathcal{L}(\boldsymbol{\phi}^{(0)})[\hat{\boldsymbol{B}}^{(s)}_t] - \frac{\eta}{2}\nabla^3\mathcal{L}(\boldsymbol{\phi}^{(0)})[\hat{\boldsymbol{A}}^{(s)}_t] + \tilde{\mathcal{O}}(\eta^{2.5-\beta}), \tag{72}$$

$$\hat{\boldsymbol{A}}^{(s)}_{t+1} = \hat{\boldsymbol{A}}^{(s)}_t - \eta\boldsymbol{H}_0\hat{\boldsymbol{A}}^{(s)}_t - \eta\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{H}_0 + \frac{\eta^2}{B_{\mathrm{loc}}}\boldsymbol{\Sigma}_0 + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}), \tag{73}$$

$$\hat{\boldsymbol{M}}^{(s)}_{t+1} = \hat{\boldsymbol{M}}^{(s)}_t - \eta\boldsymbol{H}_0\hat{\boldsymbol{M}}^{(s)}_t - \eta\hat{\boldsymbol{M}}^{(s)}_t\boldsymbol{H}_0 + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}), \tag{74}$$

$$\hat{\boldsymbol{B}}^{(s)}_{t+1} = (\boldsymbol{I} - \eta\boldsymbol{H}_0)\hat{\boldsymbol{B}}^{(s)}_t + \tilde{\mathcal{O}}(\eta^{2.5-\beta}). \tag{75}$$

Moreover,

$$\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}} = \frac{1}{K}\hat{\boldsymbol{A}}^{(s)}_H + (1 - \frac{1}{K})\hat{\boldsymbol{M}}^{(s)}_H, \tag{76}$$

$$\hat{\boldsymbol{M}}^{(s+1)}_0 = \hat{\boldsymbol{A}}^{(s+1)}_0 = \boldsymbol{P}_\perp\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_\perp + \mathcal{O}(\eta^{1.5-0.5\beta}), \tag{77}$$

$$\hat{\boldsymbol{q}}^{(s+1)}_0 = \boldsymbol{P}_\perp\hat{\boldsymbol{q}}^{(s)}_H - \partial^2\Phi(\boldsymbol{\phi}^{(0)})[\hat{\boldsymbol{B}}^{(s)}_H] - \frac{1}{2}\partial^2\Phi(\boldsymbol{\phi}^{(0)})[\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), \tag{78}$$

$$\hat{\boldsymbol{B}}^{(s+1)}_0 = \boldsymbol{P}_\perp\hat{\boldsymbol{B}}^{(s)}_H + \boldsymbol{P}_\perp\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_\parallel + \tilde{\mathcal{O}}(\eta^{1.5-\beta}). \tag{79}$$

Proof We first derive the recursion for $\hat{\boldsymbol{q}}_t^{(s)}$. Recall the update rule for $\hat{\boldsymbol{\theta}}_{k,t}^{(s)}$:

$$\hat{\boldsymbol{\theta}}_{k,t+1}^{(s)} = \hat{\boldsymbol{\theta}}_{k,t}^{(s)} - \eta \nabla \mathcal{L}(\hat{\boldsymbol{\theta}}_{k,t}^{(s)}) - \eta \boldsymbol{z}_{k,t}^{(s)} + \hat{\boldsymbol{e}}_{k,t}^{(s)}.$$

Subtracting $\hat{\boldsymbol{\phi}}^{(s)}$ from both sides gives

$$\begin{aligned}
\hat{\boldsymbol{x}}_{k,t+1}^{(s)} &= \hat{\boldsymbol{x}}_{k,t}^{(s)} - \eta \nabla \mathcal{L}(\hat{\boldsymbol{\theta}}_{k,t}^{(s)}) - \eta \boldsymbol{z}_{k,t}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2) \\
&= \hat{\boldsymbol{x}}_{k,t}^{(s)} - \eta \left(\nabla^2 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(s)}) \hat{\boldsymbol{x}}_{k,t}^{(s)} + \frac{1}{2} \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(s)})[\hat{\boldsymbol{x}}_{k,t}^{(s)} \hat{\boldsymbol{x}}_{k,t}^{(s)\top}] + \mathcal{O}(\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^3) \right) \\
&\quad - \eta \boldsymbol{z}_{k,t}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2) \\
&= \hat{\boldsymbol{x}}_{k,t}^{(s)} - \eta \left(\nabla^2 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)}) + \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)}) \Delta \hat{\boldsymbol{\phi}}^{(s)} + \mathcal{O}(\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|^2) \right) \hat{\boldsymbol{x}}_{k,t}^{(s)} \\
&\quad - \frac{\eta}{2} \left(\nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)}) + \mathcal{O}(\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2) \right) [\hat{\boldsymbol{x}}_{k,t}^{(s)} \hat{\boldsymbol{x}}_{kt}^{(s)\top}] - \eta \boldsymbol{z}_{k,t}^{(s)} + \mathcal{O}(\eta \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^3 + \|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2) \\
&= \hat{\boldsymbol{x}}_{k,t}^{(s)} - \eta \boldsymbol{H}_0 \hat{\boldsymbol{x}}_{k,t}^{(s)} - \eta \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}_{k,t}^{(s)} \Delta \hat{\boldsymbol{\phi}}^{(s)\top}] - \frac{\eta}{2} \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{x}}_{k,t}^{(s)} \hat{\boldsymbol{x}}_{k,t}^{(s)\top}] - \eta \boldsymbol{z}_{k,t}^{(s)} \\
&\quad + \mathcal{O}(\eta \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^3 + \eta \|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2 + \eta \|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2^2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2 + \|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2), \qquad (80)
\end{aligned}$$

where the second and third equality perform Taylor expansion. Taking expectation on both sides gives

$$\begin{aligned}
\hat{\boldsymbol{q}}_{t+1}^{(s)} &= (\boldsymbol{I} - \eta \boldsymbol{H}_0) \hat{\boldsymbol{q}}_t^{(s)} - \eta \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{q}}_t^{(s)}] - \frac{\eta}{2} \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_t^{(s)}] \\
&\quad + \mathcal{O}\left(\eta \mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^3] + \eta \mathbb{E}[\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2] + \eta \mathbb{E}[\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2^2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2] + \mathbb{E}[\|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2] \right).
\end{aligned}$$

By Theorem 41, with probability at least $1 - \eta^{100}$, $\hat{\boldsymbol{e}}_{k,t}^{(s)} = \boldsymbol{0}$, $\forall k \in [K], (s,t) \preceq (R_{\mathrm{grp}}, 0)$. Also notice that both $\hat{\boldsymbol{\theta}}_{k,t}^{(s)}$ and $\boldsymbol{\phi}_{\mathrm{null}}$ belong to the bounded set Γ^{ϵ_2}. Therefore, $\|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2$ is bounded and we have $\mathbb{E}[\|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2] = \mathcal{O}(\eta^{100})$. Combining this with (69) to (71) yields (72).

Secondly, we derive the recursion for $\hat{\boldsymbol{B}}_t^{(s)}$. Multiplying both sides of (80) by $\Delta \hat{\boldsymbol{\phi}}^{(s)\top}$ and taking expectation, we have

$$\hat{\boldsymbol{B}}_{t+1}^{(s)} = (\boldsymbol{I} - \eta \boldsymbol{H}_0) \hat{\boldsymbol{B}}_t^{(s)} + \mathcal{O}(\eta \mathbb{E}[\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2 + \|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2^2 \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2 + \|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2]).$$

Still by Theorem 41 and (69) to (71), we have (75).

Thirdly, we derive the recursion for $\hat{\boldsymbol{A}}_t^{(s)}$. By (80), we have

$$\begin{aligned}
\hat{\boldsymbol{A}}_{t+1}^{(s)} &= \hat{\boldsymbol{A}}_t^{(s)} - \eta \boldsymbol{H}_0 \hat{\boldsymbol{A}}_t^{(s)} - \eta \hat{\boldsymbol{A}}_t^{(s)} \boldsymbol{H}_0 + \frac{\eta^2}{B_{\mathrm{loc}}} \boldsymbol{\Sigma}_0 + \mathcal{O}(\eta^2 \mathbb{E}[\|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2 + \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2]) \\
&\quad + \mathcal{O}(\eta \mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^3 + \|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2 \|\Delta \hat{\boldsymbol{\phi}}^{(s)}\|_2 + \|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2]) \\
&= (\boldsymbol{I} - \eta \boldsymbol{H}_0) \hat{\boldsymbol{A}}_t^{(s)} + \frac{\eta^2}{B_{\mathrm{loc}}} \boldsymbol{\Sigma}_0 + \tilde{\mathcal{O}}(\eta^{2.5 - 0.5\beta}),
\end{aligned}$$

which establishes (73).

Fourthly, we derive the recursion for $\hat{\boldsymbol{M}}_t^{(s)}$. Multiplying both sides of (80) by $\hat{\boldsymbol{x}}_{l,t+1}^{(s)}$ and taking expectation, $l \neq k$, we obtain

$$\begin{aligned}\hat{\boldsymbol{M}}_{t+1}^{(s)} = \hat{\boldsymbol{M}}_t^{(s)} - \eta \boldsymbol{H}_0 \hat{\boldsymbol{M}}_t^{(s)} - \eta \hat{\boldsymbol{M}}_t^{(s)} \boldsymbol{H}_0 + \mathcal{O}(\eta \mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2 \|\hat{\boldsymbol{x}}_{l,t}^{(s)}\|_2 \|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2]) \\ + \mathcal{O}(\eta \mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2 \|\hat{\boldsymbol{x}}_{l,t}^{(s)}\|_2 + \|\hat{\boldsymbol{e}}_{k,t}^{(s)}\|_2]).\end{aligned}$$

By a similar argument to the proof of Theorem 51, we have

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2^2 \|\hat{\boldsymbol{x}}_{l,t}^{(s)}\|_2] &= \tilde{\mathcal{O}}(\eta^{1.5}), \\ \mathbb{E}[\|\hat{\boldsymbol{x}}_{k,t}^{(s)}\|_2 \|\hat{\boldsymbol{x}}_{l,t}^{(s)}\|_2 \|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2] &= \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),\end{aligned}$$

which yields (74).

Now we proceed to prove (76) to (79). By definition of $\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}$,

$$\begin{aligned}\hat{\boldsymbol{A}}_{\text{avg}}^{(s)} &= \frac{1}{K^2}\mathbb{E}[(\sum_{k\in[K]} \hat{\boldsymbol{x}}_{k,H}^{(s)})(\sum_{k\in[K]} \hat{\boldsymbol{x}}_{k,H}^{(s)})^\top] \\ &= \frac{1}{K^2}\sum_{k\in[K]} \mathbb{E}[\hat{\boldsymbol{x}}_{k,H}^{(s)} \hat{\boldsymbol{x}}_{k,H}^{(s)\top}] + \frac{1}{K^2}\sum_{k,l\in[K],k\neq l} \mathbb{E}[\hat{\boldsymbol{x}}_{k,H}^{(s)} \hat{\boldsymbol{x}}_{l,H}^{(s)\top}] \\ &= \frac{1}{K}\hat{\boldsymbol{A}}_H^{(s)} + (1-\frac{1}{K})\hat{\boldsymbol{M}}_H^{(s)},\end{aligned}$$

which demonstrates (76). Then we derive (77). By definition of $\hat{\boldsymbol{x}}_{\text{avg},0}^{(s+1)}$,

$$\begin{aligned}\hat{\boldsymbol{x}}_{\text{avg},0}^{(s+1)} &= \hat{\boldsymbol{\phi}}^{(s)} + \hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} - \Phi(\hat{\boldsymbol{\phi}}^{(s)} + \hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}) \\ &= \hat{\boldsymbol{\phi}}^{(s)} + \hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} - \left(\hat{\boldsymbol{\phi}}^{(s)} + \partial\Phi(\hat{\boldsymbol{\phi}}^{(s)})\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^2)\right) \\ &= \hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} - \left(\boldsymbol{P}_{\|} + \mathcal{O}(\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2)\right)\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^2) \\ &= \boldsymbol{P}_{\perp}\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^2 + \|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2). \qquad (81)\end{aligned}$$

Hence,

$$\begin{aligned}\hat{\boldsymbol{M}}_0^{(s+1)} = \hat{\boldsymbol{A}}_0^{(s+1)} &= \mathbb{E}[\hat{\boldsymbol{x}}_{\text{avg},0}^{(s)}\hat{\boldsymbol{x}}_{\text{avg},0}^{(s)\top}] \\ &= \boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}\boldsymbol{P}_{\perp} + \mathcal{O}(\mathbb{E}[\|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^3 + \|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^2\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2]).\end{aligned}$$

By (69) and (71), we obtain (77). By (67),

$$\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)} = \boldsymbol{P}_{\|}\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)} + \mathcal{O}(\|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2\|\Delta\hat{\boldsymbol{\phi}}^{(s)}\|_2 + \|\hat{\boldsymbol{x}}_{\text{avg},H}^{(s)}\|_2^2). \qquad (82)$$

Combining (81) and (82) gives

$$\mathbb{E}[\hat{\boldsymbol{x}}_{\text{avg},0}^{(s)}(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] = \boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}\boldsymbol{P}_{\|} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

Therefore,

$$\begin{aligned}\hat{\boldsymbol{B}}_0^{(s+1)} &= \mathbb{E}[\hat{\boldsymbol{x}}_{\text{avg},0}^{(s+1)}\Delta\hat{\boldsymbol{\phi}}^{(s+1)\top}] = \mathbb{E}[\hat{\boldsymbol{x}}_{\text{avg},0}^{(s+1)}(\Delta\hat{\boldsymbol{\phi}}^{(s)} + \hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] \\ &= \boldsymbol{P}_{\perp}\hat{\boldsymbol{B}}_H^{(s)} + \boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}\boldsymbol{P}_{\|} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

Finally, we apply Theorem 51 to derive (78).

$$\begin{aligned}
\hat{\boldsymbol{q}}_0^{(s+1)} &= \mathbb{E}[\hat{\boldsymbol{x}}_{\mathrm{avg},0}^{(s+1)}] = \mathbb{E}[\hat{\boldsymbol{x}}_{\mathrm{avg},H}^{(s)} - (\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})] \\
&= \hat{\boldsymbol{q}}_H^{(s)} - \boldsymbol{P}_{\parallel}\hat{\boldsymbol{q}}_H^{(s)} - \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_H^{(s)}] - \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_{\mathrm{avg}}^{(s)}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}) \\
&= \boldsymbol{P}_{\perp}\hat{\boldsymbol{q}}_H^{(s)} - \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_H^{(s)}] - \frac{1}{2}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_{\mathrm{avg}}^{(s)}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}),
\end{aligned}$$

which concludes the proof. ■

With the assumption that the hessian at $\hat{\boldsymbol{\phi}}^{(0)}$ is diagonal, we have the following corollary that formulates the recursions for each matrix element.

Corollary 53 *Given $\|\hat{\boldsymbol{\theta}}_{\mathrm{avg}}^{(0)} - \hat{\boldsymbol{\phi}}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, for $0 \le s < R_{\mathrm{grp}}$ and $0 \le t < H$, we have the following elementwise recursions.*

$$\hat{A}_{t+1,i,j}^{(s)} = (1 - (\lambda_i + \lambda_j)\,\eta)\hat{A}_{t,i,j}^{(s)} + \frac{\eta^2}{B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}), \tag{83}$$

$$\hat{M}_{t+1,i,j}^{(s)} = (1 - (\lambda_i + \lambda_j)\,\eta)\hat{M}_{t,i,j}^{(s)} + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}), \tag{84}$$

$$\hat{B}_{t+1,i,j}^{(s)} = (1 - \lambda_i\eta)\hat{B}_{t,i,j}^{(s)} + \tilde{\mathcal{O}}(\eta^{2.5-\beta}), \tag{85}$$

$$\hat{A}_{\mathrm{avg},i,j}^{(s)} = \frac{1}{K}(\hat{A}_{H,i,j}^{(s)} - \hat{M}_{H,i,j}^{(s)}) + \hat{M}_{H,i,j}^{(s)}, \tag{86}$$

$$\hat{M}_{0,i,j}^{(s+1)} = \hat{A}_{0,i,j}^{(s+1)} = \begin{cases} \hat{A}_{\mathrm{avg},i,j}^{(s)} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), & 1 \le i \le m, 1 \le j \le m, \\ \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), & \textit{otherwise}. \end{cases} \tag{87}$$

$$\hat{B}_{0,i,j}^{(s+1)} = \begin{cases} \hat{B}_{H,i,j}^{(s)} + \hat{A}_{\mathrm{avg},,i,j}^{(s)} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & 1 \le i \le m, m < j \le d, \\ \hat{B}_{H,i,j}^{(s)} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & 1 \le i \le m, 1 \le j \le m, \\ \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & m < i \le d. \end{cases} \tag{88}$$

Having formulated the recursions, we are ready to solve out the explicit expressions. We will split each matrix into four parts and them one by on. Specifically, a matrix $\boldsymbol{M}$ can be split into $\boldsymbol{P}_{\parallel}\boldsymbol{M}\boldsymbol{P}_{\parallel}$ in the tangent space of Γ at $\hat{\boldsymbol{\phi}}^{(0)}$, $\boldsymbol{P}_{\perp}\boldsymbol{M}\boldsymbol{P}_{\perp}$ in the normal space, along with $\boldsymbol{P}_{\parallel}\boldsymbol{M}\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\perp}\boldsymbol{M}\boldsymbol{P}_{\parallel}$ across both spaces.

We first compute the elements of $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_t^{(s)}\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_{\mathrm{avg}}^{(s)}\boldsymbol{P}_{\perp}$.

Lemma 54 (General formula for $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_t^{(s)}\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}_{\mathrm{avg}}^{(s)}\boldsymbol{P}_{\perp}$) *Let $R_0 := \lceil \frac{10}{\lambda_m \alpha} \log \frac{1}{\eta} \rceil$. Then for $1 \le i \le m, 1 \le j \le m$ and $R_0 \le s < R_{\mathrm{grp}}$,*

$$\begin{aligned}
\hat{A}_{\mathrm{avg},i,j}^{(s)} &= \frac{1}{(\lambda_i + \lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), \\
\hat{A}_{t,i,j}^{(s)} &= -\left(1 - \frac{1}{K}\right)\frac{(1 - (\lambda_i + \lambda_j)\eta)^t}{(\lambda_i + \lambda_j)B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \frac{\eta}{(\lambda_i + \lambda_j)B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).
\end{aligned}$$

For $s < R_0$, $\hat{A}_{t,i,j}^{(s)} = \tilde{\mathcal{O}}(\eta)$ and $\hat{A}_{\mathrm{avg},,i,j}^{(s)} = \tilde{\mathcal{O}}(\eta)$.

Proof For $1 \leq i \leq m, 1 \leq j \leq m$, $\lambda_i > 0, \lambda_j > 0$. By (83),

$$\begin{aligned}
\hat{A}^{(s)}_{t,i,j} &= (1-(\lambda_i+\lambda_j)\eta)^t \hat{A}^{(s)}_{0,i,j} + \sum_{\tau=0}^{t-1}(1-(\lambda_i+\lambda_j)\eta)^\tau \frac{\eta^2}{B_{\mathrm{loc}}}\Sigma_{0,i,j} \\
&\quad + \tilde{\mathcal{O}}(\sum_{\tau=0}^{t-1}(1-(\lambda_i+\lambda_j)\eta)^\tau \eta^{2.5-0.5\beta}) \\
&= (1-(\lambda_i+\lambda_j)\eta)^t \hat{A}^{(s)}_{0,i,j} + \frac{1-(1-(\lambda_i+\lambda_j)\eta)^t}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),
\end{aligned}$$

where the second inequality uses $\sum_{\tau=0}^{t-1}(1-(\lambda_i+\lambda_j)\eta)^\tau = \frac{1-(1-(\lambda_i+\lambda_j)\eta)^t}{(\lambda_i+\lambda_j)\eta} \leq \frac{1}{(\lambda_i+\lambda_j)\eta}$. By (84),

$$\begin{aligned}
\hat{M}^{(s)}_{t,i,j} &= (1-(\lambda_i+\lambda_j)\eta)^t \hat{M}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\sum_{\tau=0}^{t-1}(1-(\lambda_i+\lambda_j)\eta)^\tau \eta^{2.5-0.5\beta}) \\
&= (1-(\lambda_i+\lambda_j)\eta)^t \hat{A}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),
\end{aligned}$$

where the second equality uses $\boldsymbol{M}_0^{(s+1)} = \boldsymbol{A}_0^{(s+1)}$. By (86) and (87),

$$\begin{aligned}
\hat{A}^{(s)}_{\mathrm{avg},i,j} &= \frac{1-(1-(\lambda_i+\lambda_j)\eta)^H}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + (1-(\lambda_i+\lambda_j)\eta)^H \hat{A}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), \\
\hat{A}^{(s+1)}_{0,i,j} &= \hat{A}^{(s)}_{\mathrm{avg},i,j} + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}) \\
&= \frac{1-(1-(\lambda_i+\lambda_j)\eta)^H}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + (1-(\lambda_i+\lambda_j)\eta)^H \hat{A}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).
\end{aligned}$$

Then we obtain

$$\begin{aligned}
\hat{A}^{(s)}_{0,i,j} &= (1-(\lambda_i+\lambda_j)\eta)^{sH}\hat{A}^{(0)}_{0,i,j} + \frac{1-(1-(\lambda_i+\lambda_j)\eta)^H}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j}\sum_{r=0}^{s-1}(1-(\lambda_i+\lambda_j)\eta)^{rH} \\
&\quad + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}\sum_{r=R_0}^{s-1}(1-(\lambda_i+\lambda_j)\eta)^{rH}).
\end{aligned}$$

Notice that $|1-(\lambda_i+\lambda_j)\eta| < 1$ and

$$(1-(\lambda_i+\lambda_j)\eta)^H \leq \exp(-(\lambda_i+\lambda_j)\eta H) = \exp(-(\lambda_i+\lambda_j)\alpha). \tag{89}$$

Therefore,

$$\sum_{r=0}^{s-1}(1-(\lambda_i+\lambda_j)\eta)^{rH} = \frac{1-(1-(\lambda_i+\lambda_j)\eta)^{rH}}{1-(1-(\lambda_i+\lambda_j)\eta)^H} \leq \frac{1}{1-\exp(-(\lambda_i+\lambda_j)\alpha)}.$$

Then we have

$$\hat{A}^{(s)}_{0,i,j} = (1-(\lambda_i+\lambda_j)\eta)^{sH}\hat{A}^{(0)}_{0,i,j} + \frac{1-(1-(\lambda_i+\lambda_j)\eta)^{sH}}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

Finally, we demonstrate that for $s \geq R_0$, $\hat{A}^{(s)}_{0,i,j}$ and $\hat{A}^{(s)}_{\mathrm{avg},i,j}$ is approximately equal to $\frac{\eta}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\Sigma_{0,i,j}$. By (89), when $s \geq R_0$, $(1-(\lambda_i+\lambda_j)\eta)^{sH} = \mathcal{O}(\eta^{10})$, which gives

$$\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{1}{(\lambda_i+\lambda_j)KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),$$

$$A^{(s)}_{t,i,j} = -\left(1-\frac{1}{K}\right)\frac{(1-(\lambda_i+\lambda_j)\eta)^t}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \frac{\eta}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

For $s < R_0$, since $\hat{\boldsymbol{A}}^{(0)}_0 = \hat{\boldsymbol{x}}^{(s)}_{\mathrm{avg},0}\hat{\boldsymbol{x}}^{(s)\top}_{\mathrm{avg},0}$=$\tilde{\mathcal{O}}(\eta)$, we have $\hat{A}^{(s)}_{\mathrm{avg},,i,j} = \tilde{\mathcal{O}}(\eta)$ and $\hat{A}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta)$. ■

Secondly, we compute $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\perp}$.

Lemma 55 (General formula for $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\|}$ and $\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\|}$) *For $1 \leq i \leq m, m < j \leq d$,*

$$\hat{A}^{(s)}_{t,i,j} = \frac{1-(1-\lambda_i\eta)^t}{\lambda_i B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),$$

$$\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{1-(1-\lambda_i\eta)^H}{\lambda_i KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

Proof Note that for $1 \leq i \leq m, m < j \leq d$ and $\lambda_i > 0, \lambda_j = 0$. By (83) and (87),

$$\begin{aligned}\hat{A}^{(s)}_{t,i,j} &= (1-\lambda_i\eta)^t\hat{A}^{(s)}_{0,i,j} + \frac{1-(1-\lambda_i\eta)^t}{\lambda_i B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta})\\ &= \frac{1-(1-\lambda_i\eta)^t}{\lambda_i B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

By (84) and (87), $\hat{M}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta})$. Then,

$$\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{1-(1-\lambda_i\eta)^H}{\lambda_i KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

■

Similar to Theorem 55, we have the following lemma for the general formula of $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\perp}$.

Lemma 56 (General formula for $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\perp}$ and $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\perp}$) *For $m < i \leq d$ and $1 \leq j \leq m$,*

$$\hat{A}^{(s)}_{t,i,j} = \frac{1-(1-\lambda_j\eta)^t}{\lambda_j B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),$$

$$\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{1-(1-\lambda_j\eta)^H}{\lambda_j KB_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

Finally, we derive the general formula for $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\|}$ and $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\mathrm{avg}}\boldsymbol{P}_{\|}$.

Lemma 57 (General formula for $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}_t^{(s)}\boldsymbol{P}_{\|}$ and $\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}_{\mathrm{avg}}^{(s)}\boldsymbol{P}_{\|}$) *For $m < i \le d$ and $m < j \le d$,*

$$\begin{aligned}\hat{A}^{(s)}_{\mathrm{avg},i,j} &= \frac{H\eta^2}{KB_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),\\ \hat{A}^{(s)}_{t,i,j} &= \hat{A}^{(s)}_{0,i,j} + \frac{t\eta^2}{B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).\end{aligned}$$

Proof Note that for $m < i \le d$, $m < j \le d$ and $\lambda_i = \lambda_j = 0$. (83) is then simplified as

$$\hat{A}^{(s)}_{t+1,i,j} = \hat{A}^{(s)}_{t,i,j} + \frac{\eta^2}{B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-0.5\beta}).$$

Therefore,

$$\hat{A}^{(s)}_{t,i,j} = \hat{A}^{(s)}_{0,i,j} + \frac{t\eta^2}{B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}). \tag{90}$$

According to (84), $\hat{M}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta})$ for $m < i \le d$ and $m < j \le d$. Combining (84), (87) and (90) yields

$$\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{H\eta^2}{KB_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}).$$

■

Now, we move on to compute the general formula for $\hat{\boldsymbol{B}}_t^{(s)}$.

Lemma 58 (The general formula for $\boldsymbol{P}_{\perp}\hat{\boldsymbol{B}}_t^{(s)}\boldsymbol{P}_{\|}$) *Note that for $1 \le i \le m$ and $m < j \le d$, when $R_0 := \lceil \frac{10}{\lambda_m \alpha}\log\frac{1}{\eta}\rceil \le s < R_{\mathrm{grp}}$,*

$$\hat{B}^{(s)}_{t,i,j} = \frac{(1-\lambda_i\eta)^t}{\lambda_i K B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

For $s < R_0$, $\hat{B}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta)$.

Proof Note that for $1 \le i \le m$, $\lambda_i > 0$. By (85),

$$\hat{B}^{(s)}_{t+1,i,j} = (1-\lambda_i\eta)\hat{B}^{(s)}_{t,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-\beta}).$$

Hence,

$$\hat{B}^{(s)}_{t,i,j} = (1-\lambda_i\eta)^t\hat{B}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

According to (88),

$$\begin{aligned}\hat{B}^{(s+1)}_{0,i,j} &= \hat{B}^{(s)}_{H,i,j} + \hat{A}^{(s)}_{\mathrm{avg},,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-\beta})\\ &= (1-\lambda_i\eta)^H\hat{B}^{(s)}_{0,i,j} + \hat{A}^{(s)}_{\mathrm{avg},i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

Then we have

$$\begin{aligned}\hat{B}^{(s)}_{0,i,j} &= (1-\lambda_i\eta)^{sH}\hat{B}^{(0)}_{0,i,j} + \hat{A}^{(s)}_{\mathrm{avg},i,j}\sum_{r=0}^{s-1}(1-\lambda_i\eta)^{rH} + \tilde{\mathcal{O}}(\sum_{r=0}^{s-1}(1-\lambda_i\eta)^{rH}\eta^{1.5-\beta})\\ &= (1-\lambda_i\eta)^{sH}\hat{B}^{(0)}_{0,i,j} + \frac{1-(1-\lambda_i\eta)^{sH}}{1-(1-\lambda_i\eta)^{H}}\hat{A}^{(s)}_{\mathrm{avg},,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta})\\ &= \frac{1-(1-\lambda_i\eta)^{sH}}{1-(1-\lambda_i\eta)^{H}}\hat{A}^{(s)}_{\mathrm{avg},,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

where the second equality uses (89) and the last inequality uses $\hat{\boldsymbol{B}}^{(0)}_0 = \hat{\boldsymbol{x}}^{(0)}_{\mathrm{avg},0}\Delta\hat{\boldsymbol{\phi}}^{(0)} = \boldsymbol{0}$. For $s \geq R_0$, $\hat{A}^{(s)}_{\mathrm{avg},i,j} = \frac{1-(1-\lambda_i\eta)^H}{\lambda_i K B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta})$, which gives

$$\hat{B}^{(s)}_{0,i,j} = \frac{\eta}{\lambda_i K B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

Therefore,

$$\hat{B}^{(s)}_{t,i,j} = \frac{(1-\lambda_i\eta)^t}{\lambda_i K B_{\mathrm{loc}}}\eta\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

For $s < R_0$, $\hat{A}^{(s)}_{\mathrm{avg},,i,j} = \tilde{\mathcal{O}}(\eta)$ and therefore, $\hat{B}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta)$. ■

Lemma 59 (General formula for the elements of $\boldsymbol{P}_\perp\hat{\boldsymbol{B}}^{(s)}_t\boldsymbol{P}_\perp$) *For* $1 \leq i \leq m$ *and* $1 \leq j \leq m$, , $\hat{B}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta^{1.5-\beta})$.

Proof Note that for $1 \leq i \leq m$, $\lambda_i > 0$. By (85),

$$\hat{B}^{(s)}_{t+1,i,j} = (1-\lambda_i\eta)\hat{B}^{(s)}_{t,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-\beta}).$$

Hence,

$$\hat{B}^{(s)}_{t,i,j} = (1-\lambda_i\eta)^t\hat{B}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

By (88),

$$\begin{aligned}\hat{B}^{(s+1)}_{0,i,j} &= \hat{B}^{(s)}_{H,i,j} + \tilde{\mathcal{O}}(\eta^{2.5-\beta})\\ &= (1-\lambda_i\eta)^H\hat{B}^{(s)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta})\\ &= (1-\lambda_i\eta)^{sH}\hat{B}^{(0)}_{0,i,j} + \tilde{\mathcal{O}}(\sum_{r=0}^{s-1}(1-\lambda_i\eta)^{rH}\eta^{1.5-\beta})\\ &= (1-\lambda_i\eta)^{sH}\hat{B}^{(0)}_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta})\\ &= \tilde{\mathcal{O}}(\eta^{1.5-\beta}),\end{aligned}$$

where the last inequality uses $\hat{\boldsymbol{B}}^{(0)}_0 = \boldsymbol{0}$. ■

Lemma 60 (General formula for $\boldsymbol{P}_{\|}\hat{\boldsymbol{B}}_t^{(s)}$) *For $m < i \leq d$, $\hat{B}_{t,i,j}^{(s)} = \tilde{\mathcal{O}}(\eta^{1.5-\beta})$.*

Proof Note that $\lambda_i = 0$ for $m < i \leq d$. By (85) and (88),

$$\hat{B}_{t+1}^{(s)} = \hat{B}_t^{(s)} + \tilde{\mathcal{O}}(\eta^{2.5-\beta}), \quad \hat{B}_0^{(s)} = \tilde{\mathcal{O}}(\eta^{2.5-\beta}).$$

Therefore,

$$\hat{B}_t^{(s)} = t\tilde{\mathcal{O}}(\eta^{2.5-\beta}) + \hat{B}_0^{(s)} = \tilde{\mathcal{O}}(\eta^{1.5-\beta}).$$

■

Having obtained the expressions for $\hat{\boldsymbol{B}}_t^{(s)}$, $\hat{\boldsymbol{A}}_t^{(s)}$ and $\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}$, we now provide explicit expressions for the first and second moments of one round movements in the following two lemmas.

Lemma 61 *The expectation of one round movement is*

$$\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] = \begin{cases} \frac{H\eta^2}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_0 + \boldsymbol{\Psi}(\hat{\boldsymbol{\phi}}^{(0)})] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), & R_0 < s < R_{\text{grp}} \\ \tilde{\mathcal{O}}(\eta), & s \leq R_0 \end{cases}, \tag{91}$$

where $R_0 := \lceil \frac{10}{\lambda_m \alpha} \log \frac{1}{\eta} \rceil$.

Proof We first compute $\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}]$. By (65), we only need to compute $\boldsymbol{P}_{\|}\hat{\boldsymbol{q}}_H^{(s)}$ by relating it to these matrices. Multiplying both sides of (72) by $\boldsymbol{P}_{\|}$ gives

$$\boldsymbol{P}_{\|}\hat{\boldsymbol{q}}_{t+1}^{(s)} = \boldsymbol{P}_{\|}\hat{\boldsymbol{q}}_t^{(s)} - \eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_t^{(s)}] - \frac{\eta}{2}\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_t^{(s)}] + \tilde{\mathcal{O}}(\eta^{2.5-\beta}). \tag{92}$$

Similarly, according to (78), we have

$$\boldsymbol{P}_{\|}\hat{\boldsymbol{q}}_0^{(s+1)} = -\boldsymbol{P}_{\|}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_H^{(s)}] - \frac{1}{2}\boldsymbol{P}_{\|}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}). \tag{93}$$

Combining (92) and (93) yields

$$\begin{aligned} \boldsymbol{P}_{\|}\hat{\boldsymbol{q}}_H^{(s)} = &-\frac{1}{2}\boldsymbol{P}_{\|}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_{\text{avg}}^{(s-1)}] - \frac{\eta}{2}\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\sum_{t=0}^{H-1}\hat{\boldsymbol{A}}_t^{(s)}] \\ &- \eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\sum_{t=0}^{H-1}\hat{\boldsymbol{B}}_t^{(s)}] - \boldsymbol{P}_{\|}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_H^{(s-1)}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}). \end{aligned} \tag{94}$$

By Lemmas 54, 57 and 55, for $s \leq R_0 = \lfloor \frac{10}{\lambda\alpha} \log \frac{1}{\eta} \rfloor$, $\hat{\boldsymbol{A}}_t^{(s)} = \tilde{\mathcal{O}}(\eta)$, $\hat{\boldsymbol{A}}_{\text{avg}}^{(s)} = \tilde{\mathcal{O}}(\eta)$ and $\hat{\boldsymbol{B}}_t^{(s)} = \tilde{\mathcal{O}}(\eta)$. Therefore, $\mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] = \tilde{\mathcal{O}}(\eta)$. For $s > R_0$, $\hat{\boldsymbol{A}}_{\text{avg}}^{(s-1)} = \hat{\boldsymbol{A}}_{\text{avg}}^{(s)} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta})$. Substituting (94) into (65) gives

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] = &\underbrace{\frac{1}{2}\boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}_{\text{avg}}^{(s)}] + \boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}_H^{(s)}]}_{\mathcal{T}_1} \\ &\overbrace{-\eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\underbrace{\frac{1}{2}\sum_{t=0}^{H-1}\hat{\boldsymbol{A}}_t^{(s)} + \sum_{t=0}^{H-1}\hat{\boldsymbol{B}}_t^{(s)}}_{\mathcal{T}_3}]}^{\mathcal{T}_2} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}). \end{aligned}$$

Below we compute $\mathcal{T}_1$ and $\mathcal{T}_2$ for $s > R_0$ respectively. By Theorem 18,

$$\begin{aligned}\boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}\boldsymbol{P}_{\|}] = \boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}\boldsymbol{P}_{\perp}] = \boldsymbol{0},\\ \boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}\boldsymbol{P}_{\|}] = \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}\boldsymbol{P}_{\|}].\end{aligned}$$

By Theorem 19,

$$\boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}\boldsymbol{P}_{\perp}] = \boldsymbol{0}.$$

Therefore, for $s > R_0$,

$$\boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}] = \frac{H\eta^2}{2KB_{\text{loc}}}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})\Phi[\boldsymbol{\Sigma}_{0,\|}] + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),$$

where we apply Theorem 57. Similarly, for $s > R_0$,

$$\boldsymbol{P}_{\perp}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\hat{\boldsymbol{B}}^{(s)}_{H}] = \partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\|}\hat{\boldsymbol{B}}^{(s)}_{H}\boldsymbol{P}_{\|}] = \tilde{\mathcal{O}}(\eta^{1.5-\beta}),$$

where we apply Theorem 60. Hence,

$$\mathcal{T}_1 = \frac{H\eta^2}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_{0,\|}] + \tilde{\mathcal{O}}(\eta^{1.5-\beta}). \tag{95}$$

We move on to show that

$$\mathcal{T}_2 = \frac{H\eta^2}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{0,\|} + (K-1)\boldsymbol{\Psi}(\hat{\boldsymbol{\phi}}^{(0)})]. \tag{96}$$

Similar to the way we compute $\hat{\boldsymbol{A}}^{(s)}_t$, $\hat{\boldsymbol{A}}^{(s)}_{\text{avg}}$ and $\hat{\boldsymbol{B}}^{(s)}_t$, we compute $\mathcal{T}_2$ by splitting $\mathcal{T}_3$ into four matrices and then substituting them into the linear operator $-\eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\cdot]$ one by one. First, we show that

$$\begin{aligned}-\eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\perp}\mathcal{T}_3\boldsymbol{P}_{\perp}] = &\frac{H\eta^2}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_{0,\perp} + (K-1)\psi(\boldsymbol{\Sigma}_{0,\perp})]\\ &+ \tilde{\mathcal{O}}(\eta^{1.5-\beta}),\end{aligned} \tag{97}$$

where $\psi(\cdot)$ is interpreted as an *elementwise* matrix function here. By Lemmas 54 and 59, for $1 \le i \le m$, $1 \le j \le m$ and $s > R_0$,

$$\begin{aligned}\hat{A}^{(s)}_{t,i,j} &= -\left(1-\frac{1}{K}\right)\frac{(1-(\lambda_i+\lambda_j)\eta)^t}{(\lambda_i+\lambda_j)B_{\text{loc}}}\eta\Sigma_{0,i,j} + \frac{\eta}{(\lambda_i+\lambda_j)B_{\text{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}),\\ \hat{B}^{(s)}_{t,i,j} &= \tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{t=0}^{H-1} \hat{A}^{(s)}_{t,i,j} &= -\left(1-\frac{1}{K}\right)\frac{1-(1-(\lambda_i+\lambda_j)\eta)^H}{(\lambda_i+\lambda_j)^2 B_{\mathrm{loc}}}\Sigma_{0,i,j} + \frac{H\eta}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\Sigma_{0.,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{K(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\Sigma_{0.,i,j} \\ &\quad + \left(1-\frac{1}{K}\right)\frac{H\eta}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\underbrace{\left[1-\frac{1-(1-(\lambda_i+\lambda_j)\eta)^H}{H\eta(\lambda_i+\lambda_j)}\right]}_{\mathcal{T}_4}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}).\end{aligned}$$

$$\sum_{t=0}^{H-1} \hat{B}^{(s)}_{t,i,j} = \tilde{\mathcal{O}}(\eta^{0.5-\beta}),$$

Then we simplify $\mathcal{T}_4$. Notice that

$$\begin{aligned}(1-(\lambda_i+\lambda_i)\eta)^H &= \exp(-H(\lambda_i+\lambda_j)\eta)[1+\mathcal{O}(H\eta^2)] \\ &= \exp(-H(\lambda_i+\lambda_j)\eta)+\mathcal{O}(\eta).\end{aligned}$$

Therefore,

$$\mathcal{T}_4 = \psi((\lambda_i+\lambda_j)H\eta)+\mathcal{O}(\eta).$$

Substituting $\mathcal{T}_4$ back into the expression for $\sum_{t=0}^{H-1}\hat{A}^{(s)}_{t,i,j}$ gives

$$\sum_{t=0}^{H-1} \hat{A}^{(s)}_{t,i,j} = \frac{H\eta}{K(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\Sigma_{0.,i,j} + \left(1-\frac{1}{K}\right)\frac{H\eta\psi((\lambda_i+\lambda_j)H\eta)}{(\lambda_i+\lambda_j)B_{\mathrm{loc}}}\Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}).$$

Combining the elementwise results, we obtain the following matrix form expression:

$$\begin{aligned}-\eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\perp}\mathcal{T}_3\boldsymbol{P}_{\perp}] = &-\frac{H\eta^2}{2B}\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\mathcal{V}_{\boldsymbol{H}_0}(\boldsymbol{\Sigma}_{0,\perp}+(K-1)\psi(\boldsymbol{\Sigma}_{0,\perp}))] \\ &+\tilde{\mathcal{O}}(\eta^{1.5-\beta}).\end{aligned}$$

By Theorem 19, we have (97).

Secondly, we show that for $s > R_0$,

$$\begin{aligned}&-\eta\boldsymbol{P}_{\|}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\perp}\mathcal{T}_3\boldsymbol{P}_{\|}+\boldsymbol{P}_{\|}\mathcal{T}_3\boldsymbol{P}_{\perp}] \\ &= \frac{H\eta^2}{B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_{0,\perp,\|}+(K-1)\psi(\boldsymbol{\Sigma}_{0,\perp,\|})]+\tilde{\mathcal{O}}(\eta^{1.5-\beta}),\end{aligned} \tag{98}$$

where $\psi(\cdot)$ is interpreted as an *elementwise* matrix function here. By symmetry of $\hat{\boldsymbol{A}}^{(s)}_t$'s and $\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})$,

$$\frac{1}{2}\nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})\left[\sum_{t=0}^{H-1}\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\|}+\sum_{t=0}^{H-1}\boldsymbol{P}_{\|}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\perp}\right] = \nabla^3\mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})\left[\sum_{t=0}^{H-1}\boldsymbol{P}_{\perp}\hat{\boldsymbol{A}}^{(s)}_t\boldsymbol{P}_{\|}\right].$$

Therefore, we only have to evaluate

$$\nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)}) \left[\sum_{t=0}^{H-1} \boldsymbol{P}_{\perp} (\hat{\boldsymbol{A}}_t^{(s)} + \hat{\boldsymbol{B}}_t^{(s)}) \boldsymbol{P}_{\|} + \sum_{t=0}^{H-1} \boldsymbol{P}_{\|} \hat{\boldsymbol{B}}_t^{(s)} \boldsymbol{P}_{\perp} \right].$$

To compute the elements of $\sum_{t=0}^{H-1} \boldsymbol{P}_{\perp} (\hat{\boldsymbol{A}}_t^{(s)} + \hat{\boldsymbol{B}}_t^{(s)}) \boldsymbol{P}_{\|}$, we combine Lemmas 55 and 58 to obtain that for $1 \le i \le m$ and $m < j \le d$,

$$\begin{aligned} \sum_{t=0}^{H-1} \hat{A}_{t,i,j}^{(s)} &= \sum_{t=0}^{H-1} \frac{1-(1-\lambda_i \eta)^t}{\lambda_i B_{\mathrm{loc}}} \eta \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{\lambda_i B_{\mathrm{loc}}} \Sigma_{0,i,j} - \frac{1-(1-\lambda_i\eta)^H}{\lambda_i^2 B_{\mathrm{loc}}} \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{\lambda_i B_{\mathrm{loc}}} \left(1 - \frac{1-(1-\lambda_i\eta)^H}{\lambda_i H \eta} \right) \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{\lambda_i B_{\mathrm{loc}}} \psi(\lambda_i H \eta) \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}), \end{aligned}$$

and

$$\begin{aligned} \sum_{t=0}^{H-1} \hat{B}_{t,i,j}^{(s)} &= \sum_{t=0}^{H-1} \frac{(1-\lambda_i \eta)^t}{\lambda_i K B_{\mathrm{loc}}} \eta \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{1.5-\beta}), \\ &= \frac{1-(1-\lambda_i\eta)^H}{\lambda_i^2 K B_{\mathrm{loc}}} \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{\lambda_i K B_{\mathrm{loc}}} \Sigma_{0,i,j} - \frac{H\eta}{\lambda_i K B_{\mathrm{loc}}} \left(1 - \frac{1-(1-\lambda_i\eta)^H}{\lambda_i H \eta} \right) \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}) \\ &= \frac{H\eta}{\lambda_i K B_{\mathrm{loc}}} \Sigma_{0,i,j} - \frac{H\eta}{\lambda_i K B_{\mathrm{loc}}} \psi(\lambda_i H \eta) \Sigma_{0,i,j} + \tilde{\mathcal{O}}(\eta^{0.5-\beta}). \end{aligned}$$

Therefore, the matrix form of $\sum_{t=0}^{H-1} \boldsymbol{P}_{\perp} (\hat{\boldsymbol{A}}_t^{(s)} + \hat{\boldsymbol{B}}_t^{(s)}) \boldsymbol{P}_{\|}$ is

$$\sum_{t=0}^{H-1} \boldsymbol{P}_{\perp} (\hat{\boldsymbol{A}}_t^{(s)} + \hat{\boldsymbol{B}}_t^{(s)}) \boldsymbol{P}_{\|} = \frac{H\eta}{B} \mathcal{V}_{\boldsymbol{H}_0} \left(\boldsymbol{\Sigma}_{0,\perp,\|} + (K-1)\psi(\boldsymbol{\Sigma}_{0,\perp,\|}) \right) + \tilde{\mathcal{O}}(\eta^{0.5-\beta}),$$

where $\psi(\cdot)$ is interpreted as an *elementwise* matrix function here. Furthermore, by Theorem 60, $\sum_{t=0}^{H-1} \hat{\boldsymbol{B}}_t^{(s)} = \tilde{\mathcal{O}}(\eta^{0.5-\beta})$. Applying Theorem 18, we have (98). Finally, directly applying Theorem 20, we have

$$-\eta \boldsymbol{P}_{\|} \nabla^3 \mathcal{L}(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{P}_{\|} \mathcal{T}_3 \boldsymbol{P}_{\|}] = \mathbf{0}. \tag{99}$$

Notice that $\psi(\boldsymbol{\Sigma}_{0,\|}) = \mathbf{0}$ where $\psi(\cdot)$ operates on each element. Combining (97), (98) and (99), we obtain (96). By (95) and (96), we have (91). ■

Lemma 62 *The second moment of one round movement is*

$$\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] = \begin{cases} \frac{H\eta^2}{B}\boldsymbol{\Sigma}_{0,\|} + \tilde{\mathcal{O}}(\eta^{1.5-0.5\beta}), & R_0 \le s < R_{\mathrm{grp}} \\ \tilde{\mathcal{O}}(\eta), & s < R_0 \end{cases},$$

where $R_0 := \lceil \frac{10}{\lambda_m \alpha} \log \frac{1}{\eta} \rceil$.

Proof Directly apply Theorem 57 and Theorem 51 and we have the lemma. ■

With Lemmas 61 and 62, we are ready to prove Theorem 50.

Proof [Proof of Theorem 50.] We first derive $\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}]$. Recall that $R_{\mathrm{grp}} = \lfloor \frac{1}{\alpha\eta^\beta} \rfloor = \frac{1}{H\eta^{1+\beta}} + o(1)$ where $0 < \beta < 0.5$. By Theorem 61,

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(0)}] &= \sum_{s=0}^{R_0} \mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] + \sum_{s=R_0+1}^{R_{\mathrm{grp}}-1} \mathbb{E}[\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)}] \\ &= \frac{\eta^{1-\beta}}{2B}\partial^2\Phi(\hat{\boldsymbol{\phi}}^{(0)})[\boldsymbol{\Sigma}_0 + \boldsymbol{\Psi}(\hat{\boldsymbol{\phi}}^{(0)})] + \tilde{\mathcal{O}}(\eta^{1.5-2\beta}) + \tilde{\mathcal{O}}(\eta).\end{aligned}$$

Then we compute $\mathbb{E}[\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}\Delta\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})\top}]$.

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{s=0}^{R_{\mathrm{grp}}-1}(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})\right)\left(\sum_{s=0}^{R_{\mathrm{grp}}-1}(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})\right)^\top\right] \\ &= \sum_{s=0}^{R_{\mathrm{grp}}-1}\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})^\top] + \sum_{s\neq s'}\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})]\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s'+1)} - \hat{\boldsymbol{\phi}}^{(s')})^\top] \\ &= \frac{\eta^{1-\beta}}{B}\boldsymbol{\Sigma}_{0,\|} + \tilde{\mathcal{O}}(\eta) + \tilde{\mathcal{O}}(\eta^{1.5-1.5\beta}),\end{aligned}$$

where the last inequality uses $\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s+1)} - \hat{\boldsymbol{\phi}}^{(s)})]\mathbb{E}[(\hat{\boldsymbol{\phi}}^{(s'+1)} - \hat{\boldsymbol{\phi}}^{(s')})^\top] = \tilde{\mathcal{O}}(\eta^2)$. ■

J.10. Proof of Weak Approximation

We are now in a position to utilize the estimate of moments obtained in previous subsections to prove the closeness of the sequence $\{\boldsymbol{\phi}^{(s)}\}_{s=0}^{\lfloor T/(H\eta^2)\rfloor}$ and the SDE solution $\{\boldsymbol{\zeta} : t \in [0, T]\}$ in the sense of weak approximation. Recall the SDE that we expect the manifold projection $\{\Phi(\bar{\boldsymbol{\theta}}^{(s)})\}_{s=0}^{\lfloor T/(H\eta^2)\rfloor}$ to track:

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\|}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t}_{\text{(a) diffusion}} \underbrace{-\tfrac{1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{(b) drift-I}} \underbrace{-\tfrac{K-1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{(c) drift-II}}\Big), \tag{100}$$

According to Theorem 18 and Theorem 19, the drift term in total can be written as the following form:

$$\text{(b)} + \text{(c)} = \frac{1}{2B}\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{\Sigma}(\boldsymbol{\zeta}) + (K-1)\boldsymbol{\Psi}(\boldsymbol{\zeta})].$$

Then by definition of $P_{\boldsymbol{\zeta}}$, (100) is equivalent to the following SDE:

$$\mathrm{d}\boldsymbol{\zeta}(t) = \frac{1}{\sqrt{B}}\partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t + \frac{1}{2B}\partial^2\Phi(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}(\boldsymbol{\zeta}) + (K-1)\boldsymbol{\Psi}(\boldsymbol{\zeta})\right]\mathrm{d}t. \tag{101}$$

Therefore, we only have to show that $\boldsymbol{\phi}^{(s)}$ closely tracks $\{\boldsymbol{\zeta}(t)\}$ satisfying Equation (101). By Theorem 14, there exists an ϵ_3 neighborhood of Γ, Γ^{ϵ_3}, where $\Phi(\cdot)$ is $\mathcal{C}^\infty$-smooth. Due to compactness of Γ, Γ^{ϵ_3} is bounded and the mappings $\partial^2\Phi(\cdot)$, $\partial\Phi(\cdot)$, $\boldsymbol{\Sigma}^{1/2}(\cdot)$, $\boldsymbol{\Sigma}(\cdot)$ and $\boldsymbol{\Psi}(\cdot)$ are all Lipschitz in Γ^{ϵ_3}. By Kirszbraun theorem, both the drift and diffusion term of (101) can be extended as Lipschitz functions on $\mathbb{R}^d$. Therefore, the solution to the extended SDE exists and is unique. We further show that the solution, if initialized as a point on Γ, always stays on the manifold almost surely.

As a preparation, we first show that Γ has no boundary.

Lemma 63 *Under Assumptions 3.1 to 3.3, Γ has no boundary.*

Proof We prove by contradiction. If Γ has boundary $\partial\Gamma$, WLOG, for a point $\boldsymbol{p} \in \partial\Gamma$, let the Hessian at $\boldsymbol{p}$ be diagonal with the form $\nabla^2\mathcal{L}(\boldsymbol{p}) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ where $\lambda_i > 0$ for $1 \le i \le m$ and $\lambda_i = 0$ for $m < i \le d$.

Denote by $\boldsymbol{x}_{i:j} := (x_i, x_{i+1}, \cdots, x_j)$ $(i \le j)$ the $(j-i+1)$-dimensional vector formed by the i-th to j-th coordinates of $\boldsymbol{x}$. Since $\frac{\partial(\nabla\mathcal{L}(\boldsymbol{p}))}{\partial\boldsymbol{p}_{1:m}} = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$ is invertible, by the implicit function theorem, there exists an open neighborhood V of $\boldsymbol{p}_{m+1:d}$ such that $\nabla\mathcal{L}(\boldsymbol{v}) = \boldsymbol{0}$, $\forall \boldsymbol{v} \in V$. Then, $\mathcal{L}(\boldsymbol{v}) = \mathcal{L}(\boldsymbol{p}) = \min_{\boldsymbol{\theta}\in U}\mathcal{L}(\boldsymbol{\theta})$ and hence $V \subset \Gamma$, which contradicts with $\boldsymbol{p} \in \partial\Gamma$. ■

Therefore, Γ is a closed manifold (i.e., compact and without boundary). Then we have the following lemma stating that Γ is invariant for (101).

Lemma 64 *Let $\boldsymbol{\zeta}(t)$ be the solution to (101) with $\boldsymbol{\zeta}(0) \in \Gamma$, then $\boldsymbol{\zeta}(t) \in \Gamma$ for all $t \ge 0$. In other words, Γ is invariant for (101).*

Proof According to Filipović [12] and Du and Duan [9], for a closed manifold $\mathcal{M}$ to be viable for the SDE $\mathrm{d}\boldsymbol{X}(t) = F(\boldsymbol{X}(t))\mathrm{d}t + \boldsymbol{B}(\boldsymbol{X}(t))\mathrm{d}\boldsymbol{W}_t$ where $F : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{B} : \mathbb{R}^d \to \mathbb{R}^d$ are locally Lipschitz, we only have to verify the following Nagumo type consistency condition:

$$\mu(\boldsymbol{x}) := F(\boldsymbol{x}) - \frac{1}{2}\sum_j \mathrm{D}[B_j(\boldsymbol{x})]B_j(\boldsymbol{x}) \in T_{\boldsymbol{x}}(\mathcal{M}), \quad B_j(\boldsymbol{x}) \in T_{\boldsymbol{x}}(\mathcal{M}),$$

where $\mathrm{D}[\cdot]$ is the Jacobian operator and $B_j(\boldsymbol{x})$ denotes the j-th column of $\boldsymbol{B}(\boldsymbol{x})$.

In our context, since for $\boldsymbol{\phi} \in \Gamma$, $\partial\Phi(\boldsymbol{\phi})$ is a projection matrix onto $T_{\boldsymbol{\phi}}(\Gamma)$, each column of $\partial\Phi(\boldsymbol{\phi})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\phi})$ belongs to $T_{\boldsymbol{\phi}}(\Gamma)$, verifying the second condition. Denote by $\boldsymbol{P}_\perp(\boldsymbol{\phi}) := \boldsymbol{I}_d - \partial\Phi(\boldsymbol{\phi})$ the projection onto the normal space of Γ at $\boldsymbol{\phi}$. To verify the first condition, it suffices to show that $\boldsymbol{P}_\perp(\boldsymbol{\phi})\boldsymbol{\mu}(\boldsymbol{\phi}) = \boldsymbol{0}$. We evaluate $\sum_j \boldsymbol{P}_\perp(\boldsymbol{\phi})\mathrm{D}[B_j(\boldsymbol{\phi})]B_j(\boldsymbol{\phi})$ as follows.

$$\begin{aligned}\sum_j \boldsymbol{P}_\perp(\boldsymbol{\phi})\mathrm{D}[B_j(\boldsymbol{\phi})]B_j(\boldsymbol{\phi}) &= \frac{1}{B}\sum_j \mathrm{D}[\partial\Phi(\boldsymbol{\phi})\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\phi})]\partial\Phi(\boldsymbol{\phi})\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\phi}) \\ &= \frac{1}{B}\boldsymbol{P}_\perp(\boldsymbol{\phi})\sum_j \partial^2\Phi(\boldsymbol{\phi})[\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\phi}), \partial\Phi(\boldsymbol{\phi})\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\phi})] \\ &= -\frac{1}{B}\nabla^2\mathcal{L}(\boldsymbol{\phi})^+\nabla^3\mathcal{L}(\boldsymbol{\phi})[\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\phi})],\end{aligned} \tag{102}$$

where the last inequality uses Theorem 18. Again applying Theorem 18, we have

$$\boldsymbol{P}_{\perp}(\boldsymbol{\phi})F(\boldsymbol{\phi}) = -\frac{1}{2B}\nabla^2\mathcal{L}(\boldsymbol{\phi})^+\nabla^3\mathcal{L}(\boldsymbol{\phi})[\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\phi})]. \tag{103}$$

Combining (102) and (103), we can verify the first condition. ■

In order to establish Theorem 4, it suffices to prove the following theorem, which captures the closeness of $\boldsymbol{\phi}^{(s)}$ and $\boldsymbol{\zeta}(t)$ every R_{grp} rounds.

Theorem 65 *If $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$ and $\boldsymbol{\zeta}(0) = \boldsymbol{\phi}^{(0)} \in \Gamma$, then for $R_{\mathrm{grp}} = \lfloor \frac{1}{\alpha\eta^{0.75}} \rfloor$ every test function $g \in \mathcal{C}^3$,*

$$\max_{n=0,\cdots,\lfloor T/\eta^{0.75}\rfloor} \left| \mathbb{E}g(\boldsymbol{\phi}^{(nR_{\mathrm{grp}})}) - \mathbb{E}g(\boldsymbol{\zeta}(n\eta^{0.75})) \right| \leq C_g \eta^{0.25}(\log \tfrac{1}{\eta})^b,$$

where $C_g > 0$ is a constant independent of η but can depend on $g(\cdot)$ and $b > 0$ is a constant independent of η and $g(\cdot)$.

J.10.1. PRELIMINARIES AND ADDITIONAL NOTATIONS

We first introduce a general formulation for stochastic gradient algorithms (SGAs) and then specify the components of this formulation in our context. Consider the following SGA:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \eta_{\mathrm{e}}\boldsymbol{h}(\boldsymbol{x}_n, \boldsymbol{\xi}_n),$$

where $\boldsymbol{x}_n \in \mathbb{R}^d$ is the parameter, η_{e} is the learning rate, $\boldsymbol{h}(\cdot,\cdot)$ is the update which depends on $\boldsymbol{x}_n$ and a random vector $\boldsymbol{\xi}_n$ sampled from some distribution $\Xi(\boldsymbol{x}_n)$. Also consider the following Stochastic Differential Equation (SDE).

$$\mathrm{d}\boldsymbol{X}(t) = \boldsymbol{b}(\boldsymbol{X}(t))\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{X}(t))\mathrm{d}\boldsymbol{W}_t,$$

where $\boldsymbol{b}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is the drift function and $\boldsymbol{\sigma}(\cdot) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is the diffusion matrix.

Denote by $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}, s, t)$ the distribution of $\boldsymbol{X}(t)$ with the initial condition $\boldsymbol{X}(s) = \boldsymbol{x}$.Define

$$\tilde{\boldsymbol{\Delta}}(\boldsymbol{x}, n) := \boldsymbol{X}_{(n+1)\eta_{\mathrm{e}}} - \boldsymbol{x}, \qquad \text{where } \boldsymbol{X}_{(n+1)\eta_{\mathrm{e}}} \sim \mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}, n\eta_{\mathrm{e}}, (n+1)\eta_{\mathrm{e}}),$$

which characterizes the update in one step.

In our context, we view the movement of manifold projection over $R_{\mathrm{grp}} := \lfloor \frac{1}{\alpha\eta^{1-\beta}} \rfloor (\beta \in (0, 0.5))$ rounds as one update step. Hence the $\boldsymbol{\phi}^{(nR_{\mathrm{grp}})}$ corresponds to the discrete time random variable $\boldsymbol{x}_n$ corresponds to and $\boldsymbol{\zeta}(t)$ corresponds to the continuous time random variable $\boldsymbol{X}_t$. According to Theorem 48, we set

$$\eta_{\mathrm{e}} = \eta^{1-\beta}, \quad \boldsymbol{b}(\boldsymbol{\zeta}) = \frac{1}{2B}\partial^2\Phi(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}(\boldsymbol{\zeta}) + (K-1)\boldsymbol{\Psi}(\boldsymbol{\zeta})\right], \quad \boldsymbol{\sigma}(\boldsymbol{\zeta}) = \frac{1}{\sqrt{B}}\partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta}).$$

Due to compactness of Γ, $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ are Lipschitz on Γ.

As for the update in one step, $\tilde{\boldsymbol{\Delta}}(\cdot,\cdot)$ is defined in our context as:

$$\tilde{\boldsymbol{\Delta}}(\boldsymbol{\phi}, n) := \boldsymbol{\zeta}_{(n+1)\eta_{\mathrm{e}}} - \boldsymbol{\phi}, \qquad \text{where } \boldsymbol{\zeta}_{(n+1)\eta_{\mathrm{e}}} \sim \mathcal{P}_{\boldsymbol{\zeta}}(\boldsymbol{\phi}, n\eta_{\mathrm{e}}, (n+1)\eta_{\mathrm{e}}) \text{ and } \boldsymbol{\phi} \in \Gamma.$$

For convenience, we further define

$$\boldsymbol{\Delta}^{(n)} := \hat{\boldsymbol{\phi}}^{((n+1)R_{\mathrm{grp}})} - \hat{\boldsymbol{\phi}}^{(nR_{\mathrm{grp}})}, \qquad \tilde{\boldsymbol{\Delta}}^{(n)} := \tilde{\boldsymbol{\Delta}}(\hat{\boldsymbol{\phi}}^{(R_{\mathrm{grp}})}, n),$$
$$\boldsymbol{b}^{(n)} : = \boldsymbol{b}(\hat{\boldsymbol{\phi}}^{(nR_{\mathrm{grp}})}), \qquad \boldsymbol{\sigma}^{(n)} : = \boldsymbol{\sigma}(\hat{\boldsymbol{\phi}}^{(nR_{\mathrm{grp}})}).$$

We use $C_{g,i}$ to denote constants that can depend on the test function g and independent of η_{e}. The following lemma relates the moments of $\tilde{\boldsymbol{\Delta}}(\boldsymbol{\phi}, n)$ to $\boldsymbol{b}(\boldsymbol{\phi})$ and $\boldsymbol{\sigma}(\boldsymbol{\phi})$.

Lemma 66 *There exists a positive constant C_0 independent of η_{e} and g such that for all $\boldsymbol{\phi} \in \Gamma$,*

$$|\mathbb{E}[\tilde{\Delta}_i(\boldsymbol{\phi}, n)] - \eta_{\mathrm{e}} b_i(\boldsymbol{\phi})| \leq C_0 \eta_{\mathrm{e}}^2, \qquad \forall 1 \leq i \leq d,$$
$$|\mathbb{E}[\tilde{\Delta}_i(\boldsymbol{\phi}, n)\tilde{\Delta}_j(\boldsymbol{x}, n)] - \eta_{\mathrm{e}} \sum_{l=1}^{d} \sigma_{i,l}(\boldsymbol{\phi})\sigma_{l,j}(\boldsymbol{\phi})| \leq C_0 \eta_{\mathrm{e}}^2, \qquad \forall 1 \leq i, j \leq d,$$
$$\mathbb{E}\left[\left|\prod_{s=1}^{6} \tilde{\Delta}_{i_s}(\boldsymbol{\phi}, n)\right|\right] \leq C_0 \eta_{\mathrm{e}}^3, \qquad \forall 1 \leq i_1, \cdots, i_6 \leq d.$$

The lemma below states that the expectation of the test function is smooth with respect to the initial value.

Proof *Noticing that (i) the solution to (101) always stays on Γ almost surely if its initial value $\boldsymbol{\zeta}(0)$ belongs to Γ , (ii) $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ are $\mathcal{C}^\infty$ and (iii) Γ is compact, we can directly apply Lemma B.3 in [42] and Lemma 26 in [34] to obtain the above lemma.* ■

The following lemma states that the expectation of $g(\boldsymbol{\zeta}(t))$ for $g \in \mathcal{C}^3$ is smooth with respect to the initial value of the SDE solution.

Lemma 67 *Let $s \in [0, T]$, $\boldsymbol{\phi} \in \Gamma$ and $g \in \mathcal{C}^3$. For $t \in [s, T]$, define*

$$u(\boldsymbol{\phi}, s, t) := \mathbb{E}_{\boldsymbol{\zeta}_t \sim \mathcal{P}_{\boldsymbol{\zeta}}(\boldsymbol{\phi}, s, t)}[g(\boldsymbol{\zeta}_t)].$$

Then $u(\cdot, s, t) \in \mathcal{C}^3$ uniformly in s, t.

Proof A slight modification of Lemma B.4 in [42] will give the above lemma. ■

J.10.2. PROOF OF THE APPROXIMATION IN OUR CONTEXT

For $\beta \in (0, 0.5)$, define $\gamma_1 := \frac{1.5-2\beta}{1-\beta}, \gamma_2 := \frac{1}{1-\beta}$, and then $1 < \gamma_1 < 1.5$, $1 < \gamma_2 < 2$. We introduce the following lemma which serves as a key step to control the approximation error. Specifically, this lemma bounds the difference in one step change between the discrete process and the continuous one as well as the product of higher orders.

Lemma 68 *If $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta \log \frac{1}{\eta}})$, then there exist positive constants C_1 and b independent of η_{e} and g such that for all $0 \leq n < \lfloor T/\eta_{\mathrm{e}} \rfloor$,*

1.

$$|\mathbb{E}[\Delta_i^{(n)} - \tilde{\Delta}_i^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}| \leq C_1 \eta_{\mathrm{e}}^{\gamma_1} (\log \tfrac{1}{\eta_{\mathrm{e}}})^b + C_1 \eta_{\mathrm{e}}^{\gamma_2} (\log \tfrac{1}{\eta_{\mathrm{e}}})^b, \qquad \forall 1 \leq i \leq d,$$
$$|\mathbb{E}[\Delta_i^{(n)}\Delta_j^{(n)} - \tilde{\Delta}_i^{(n)}\tilde{\Delta}_j^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}| \leq C_1 \eta_{\mathrm{e}}^{\gamma_1} (\log \tfrac{1}{\eta_{\mathrm{e}}})^b + C_1 \eta_{\mathrm{e}}^{\gamma_2} (\log \tfrac{1}{\eta_{\mathrm{e}}})^b, \quad \forall 1 \leq i, j \leq d.$$

2.

$$\mathbb{E}\left[\left|\prod_{s=1}^{6}\Delta_{i_s}^{(n)}\right| \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}\right] \leq C_1^2\eta_{\mathrm{e}}^{2\gamma_1}(\log\tfrac{1}{\eta_{\mathrm{e}}})^{2b}, \qquad \forall 1\leq i_1,\cdots,i_6\leq d,$$

$$\mathbb{E}\left[\left|\prod_{s=1}^{6}\tilde{\Delta}_{i_s}^{(n)}\right| \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}\right] \leq C_1^2\eta_{\mathrm{e}}^{2\gamma_1}(\log\tfrac{1}{\eta_{\mathrm{e}}})^{2b}, \qquad \forall 1\leq i_1,\cdots,i_6\leq d.$$

Proof According to Appendix J.7, we have

$$\mathbb{E}\left[\left|\prod_{s=1}^{6}\Delta_{i_s}^{(n)}\right| \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}\right] = \tilde{\mathcal{O}}(\eta^{3-3\beta}).$$

Since $\gamma_1 < 1.5$ and $\gamma_2 < 2$, we can utilize Theorem 50 and conclude that there exist positive constants C_2 and b independent of η_{e} and g such that

$$\left|\mathbb{E}[\Delta_i^{(n)} - \eta_{\mathrm{e}} b_i^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right| \leq C_2\eta_{\mathrm{e}}^{\gamma_1}(\log\tfrac{1}{\eta_{\mathrm{e}}})^b + C_2\eta_{\mathrm{e}}^{\gamma_2}(\log\tfrac{1}{\eta_{\mathrm{e}}})^b, \forall 1\leq i\leq d, \tag{104}$$

$$\left|\mathbb{E}[\Delta_i^{(n)}\Delta_j^{(n)} - \eta_{\mathrm{e}}\sum_{l=1}^{d}\sigma_{i,l}^{(n)}\sigma_{l,j}^{(n)} \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right| \leq C_2\eta_{\mathrm{e}}^{\gamma_1}(\log\tfrac{1}{\eta_{\mathrm{e}}})^b + C_2\eta_{\mathrm{e}}^{\gamma_2}(\log\tfrac{1}{\eta_{\mathrm{e}}})^b, \forall 1\leq i,j\leq d, \tag{105}$$

$$\mathbb{E}\left[\left|\prod_{s=1}^{6}\Delta_{i_s}^{(n)}\right| \mid \mathcal{E}_0^{(nR_{\mathrm{grp}})}\right] \leq C_2^2\eta_{\mathrm{e}}^{2\gamma_1}(\log\tfrac{1}{\eta_{\mathrm{e}}})^{2b}, \quad \forall 1\leq i_1,\cdots,i_6\leq d. \tag{106}$$

Combining (104) - (106) with Theorem 66 gives the above lemma. ■

Lemma 69 *For a test function $g\in\mathcal{C}^3$, let $u_{l,n}(\boldsymbol{\phi}) := u(\boldsymbol{\phi}, l\eta_{\mathrm{e}}, n\eta_{\mathrm{e}}) = \mathbb{E}_{\boldsymbol{\zeta}_t\sim\mathcal{P}_{\boldsymbol{\zeta}}(\boldsymbol{\phi}, l\eta_{\mathrm{e}}, n\eta_{\mathrm{e}})}[g(\boldsymbol{\zeta}_t)]$. If $\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\phi}^{(0)}\|_2 = \mathcal{O}(\sqrt{\eta\log\frac{1}{\eta}})$, then for all $0\leq l\leq n-1$ and $1\leq n\leq \lfloor T/\eta_{\mathrm{e}}\rfloor$,*

$$\left|\mathbb{E}[u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)}) - u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\boldsymbol{\Delta}}^{(l+1)}) \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}]\right| \leq C_{g,1}(\eta_{\mathrm{e}}^{\gamma_1} + \eta_{\mathrm{e}}^{\gamma_2})\log(\tfrac{1}{\eta_{\mathrm{e}}})^b,$$

where $C_{g,1}$ is a positive constant independent of η and $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}$ but can depend on g.

Proof By Theorem 67, $u_{l,n}(\boldsymbol{\phi})\in\mathcal{C}^3$ for all l and n. That is, there exists $K(\cdot)\in G$ such that for all l, n, $u_{l,n}(\boldsymbol{\phi})$ and its partial derivatives up to the third order are bounded by $K(\boldsymbol{\phi})$.

By the law of total expectation and triangle inequality,

$$\begin{aligned}
&\left|\mathbb{E}[u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)}) - u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\boldsymbol{\Delta}}^{(l)})] \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}\right| \\
&\leq \underbrace{\left|\mathbb{E}[u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)}) - u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\boldsymbol{\Delta}}^{(l)}) \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}]\right|}_{\mathcal{A}_1} \\
&\quad + \underbrace{\eta^{100}\mathbb{E}[|u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)})| \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \bar{\mathcal{E}}_0^{(lR_{\mathrm{grp}})}]}_{\mathcal{A}_2} \\
&\quad + \underbrace{\eta^{100}\mathbb{E}[|u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\boldsymbol{\Delta}}^{(l)})| \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \bar{\mathcal{E}}_0^{(lR_{\mathrm{grp}})}]}_{\mathcal{A}_3}.
\end{aligned}$$

We first bound $\mathcal{A}_2$ and $\mathcal{A}_3$. Since $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} \in \Gamma$, both $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)}$ and $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\boldsymbol{\Delta}}^{(l)}$ belong to Γ. Due to compactness of Γ and smoothness of $u_{l+1,n}(\cdot)$ on Γ, there exist a positive constant $C_{g,2}$ such that $\mathcal{A}_2 + \mathcal{A}_3 \le C_{g,2}\eta^{100}$.

We proceed to bound $\mathcal{A}_1$. Expanding $u_{l+1,n}(\cdot)$ at $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}$ and by triangle inequality,

$$\begin{aligned}
\mathcal{A}_1^{(s)} \le & \underbrace{\sum_{i=1}^{d} \left| \mathbb{E}[\frac{\partial u_{l+1,n}}{\partial \phi_i}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}) \left(\Delta_i^{(l)} - \tilde{\Delta}_i^{(l)} \right) \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})} \right|}_{\mathcal{B}_1} \\
& + \underbrace{\frac{1}{2} \sum_{1 \le i,j \le d} \left| \mathbb{E}[\frac{\partial^2 u_{l+1,n}}{\partial \phi_i \partial \phi_j}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}) \left(\Delta_i^{(l)}\Delta_j^{(l)} - \tilde{\Delta}_i^{(l)}\tilde{\Delta}_j^{(l)} \right) \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}] \right|}_{\mathcal{B}_2} \\
& + |\mathcal{R}| + |\tilde{\mathcal{R}}|,
\end{aligned}$$

where the remainders $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are

$$\begin{aligned}
\mathcal{R} &= \frac{1}{6} \sum_{1 \le i,j,p \le d} \mathbb{E}[\frac{\partial^3 u_{l+1,n}}{\partial \phi_i \partial \phi_j \partial \phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \theta \boldsymbol{\Delta}^{(l)}) \Delta_i^{(l)} \Delta_j^{(l)} \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}], \\
\tilde{\mathcal{R}} &= \frac{1}{6} \sum_{1 \le i,j,p \le d} \mathbb{E}[\frac{\partial^3 u_{l+1,n}}{\partial \phi_i \partial \phi_j \partial \phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \tilde{\theta} \tilde{\boldsymbol{\Delta}}^{(l)}) \tilde{\Delta}_i^{(l)} \tilde{\Delta}_j^{(l)} \tilde{\Delta}_p^{(l)} \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}],
\end{aligned}$$

for some $\theta, \tilde{\theta} \in (0,1)$. Since $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}$ belongs to Γ which is compact, there exists a constant $C_{g,3}$ such that for all $1 \le i, j \le d, 0 \le l \le n-1, 1 \le n \le \lfloor T/\eta_{\mathrm{e}} \rfloor$,

$$|\frac{\partial u_{l+1,n}}{\partial \phi_i}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})})| \le C_{g,3}, \qquad |\frac{\partial^2 u_{l+1,n}}{\partial \phi_i \partial \phi_j}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})})| \le C_{g,3}.$$

By Theorem 68,

$$\mathcal{B}_1 \le d C_{g,3} C_1 (\eta_{\mathrm{e}}^{\gamma_1} + \eta_{\mathrm{e}}^{\gamma_2}) (\log \tfrac{1}{\eta_{\mathrm{e}}})^b, \qquad \mathcal{B}_2 \le \frac{d^2}{2} C_{g,3} C_1 (\eta_{\mathrm{e}}^{\gamma_1} + \eta_{\mathrm{e}}^{\gamma_2}) (\log \tfrac{1}{\eta_{\mathrm{e}}})^b.$$

Now we bound the remainders. By Cauchy-Schwartz inequality,

$$\begin{aligned}
& \left| \mathbb{E}[\frac{\partial^3 u_{l+1,n}}{\partial \phi_i \partial \phi_j \partial \phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \theta \boldsymbol{\Delta}^{(l)}) \Delta_i^{(l)} \Delta_j^{(l)} \Delta_p^{(l)} \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(lR_{\mathrm{grp}})}] \right| \\
& \le \left(\mathbb{E}\left[\left(\frac{\partial^3 u_{l+1,n}}{\partial \phi_i \partial \phi_j \partial \phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \theta \boldsymbol{\Delta}^{(l)}) \right)^2 \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(nR_{\mathrm{grp}})} \right] \right)^{1/2} \times \\
& \left(\mathbb{E}[(\Delta_i^{(l)} \Delta_j^{(l)} \Delta_p^{(l)})^2 \mid \hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}, \mathcal{E}_0^{(nR_{\mathrm{grp}})}] \right)^{1/2}.
\end{aligned}$$

Since $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}$ and $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \boldsymbol{\Delta}^{(l)}$ both belong to Γ which is compact, there exists a constant $C_{g,4}$ such that for all $1 \le i, j, p \le d$, $0 \le l \le n-1$ and $1 \le n \le \lfloor T/\eta_{\mathrm{e}} \rfloor$,

$$\left(\tfrac{\partial^3 u_{l+1,n}}{\partial \phi_i \partial \phi_j \partial \phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})} + \theta \boldsymbol{\Delta}^{(l)}) \right)^2 \le C_{g,4}^2.$$

Combining the above inequality with Theorem 68, we have

$$\left|\mathbb{E}[\frac{\partial^3 u_{l+1,n}}{\partial\phi_i\partial\phi_j\partial\phi_p}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}+\theta\boldsymbol{\Delta}^{(l)})\Delta_i^{(l)}\Delta_j^{(l)}\Delta_p^{(l)}\mid\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})},\mathcal{E}_0^{(lR_{\mathrm{grp}})}]\right|\leq C_{g,4}C_1\eta_{\mathrm{e}}^{\gamma_1}\log(\tfrac{1}{\eta_{\mathrm{e}}})^b.$$

Hence, for all $1\leq n\leq\lfloor T/\eta_{\mathrm{e}}\rfloor, 0\leq l\leq n-1$,

$$|\mathcal{R}|\leq\frac{d^3}{6}C_{g,4}C_1\eta_{\mathrm{e}}^{\gamma_1}\log(\tfrac{1}{\eta_{\mathrm{e}}})^b.$$

Similarly, we can show that there exists a constant $C_{g,5}$ such that for all $1\leq n\leq\lfloor T/\eta_{\mathrm{e}}\rfloor, 0\leq l\leq n-1$,

$$|\tilde{\mathcal{R}}|\leq\frac{d^3}{6}C_{g,5}C_1\eta_{\mathrm{e}}^{\gamma_1}\log(\tfrac{1}{\eta_{\mathrm{e}}})^b.$$

Combining the bounds on $\mathcal{A}_1$ to $\mathcal{A}_3$, we have the lemma. ■

Finally, we prove Theorem 65.

Proof For $0\leq l\leq n$, define the random variable $\hat{\boldsymbol{\zeta}}_{l,n}$ which follows the distribution $\mathcal{P}_{\boldsymbol{\zeta}}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})},l,n)$ conditioned on $\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}$. Therefore, $\mathbb{P}(\hat{\boldsymbol{\zeta}}_{n,n}=\hat{\boldsymbol{\phi}}^{(nR_{\mathrm{grp}})})=1$ and $\hat{\boldsymbol{\zeta}}_{0,n}\sim\boldsymbol{\zeta}_{n\eta_{\mathrm{e}}}$. Denote by $u(\boldsymbol{\phi},s,t):=\mathbb{E}_{\boldsymbol{\zeta}_t\sim\mathcal{P}_{\boldsymbol{\zeta}}(\boldsymbol{\phi},s,t)}[g(\boldsymbol{\zeta}_t)]$ and $\mathcal{T}_{l+1,n}:=u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}+\boldsymbol{\Delta}^{(l)},(l+1)\eta_{\mathrm{e}},n\eta_{\mathrm{e}})-u_{l+1,n}(\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})}+\tilde{\boldsymbol{\Delta}}^{(l)},(l+1)\eta_{\mathrm{e}},n\eta_{\mathrm{e}})$.

$$\begin{aligned}&\left|\mathbb{E}[g(\boldsymbol{\phi}^{(nR_{\mathrm{grp}})})]-\mathbb{E}[g(\boldsymbol{\zeta}(n\eta_{\mathrm{e}}))]\right|\\ \leq&\left|\mathbb{E}[g(\hat{\boldsymbol{\zeta}}_{n,n})-g(\hat{\boldsymbol{\zeta}}_{0,n})\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right|+\mathcal{O}(\eta^{100})\\ \leq&\sum_{l=0}^{n-1}\left|\mathbb{E}[g(\hat{\boldsymbol{\zeta}}_{l+1,n})-g(\hat{\boldsymbol{\zeta}}_{l,n})\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right|+\mathcal{O}(\eta^{100})\\ =&\sum_{l=0}^{n-1}\left|\mathbb{E}[u(\hat{\boldsymbol{\phi}}^{((l+1)R_{\mathrm{grp}})},(l+1)\eta_{\mathrm{e}},n\eta_{\mathrm{e}})-u(\hat{\boldsymbol{\zeta}}_{l,l+1},(l+1)\eta_{\mathrm{e}},n\eta_{\mathrm{e}})\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right|+\mathcal{O}(\eta^{100})\\ =&\sum_{l=0}^{n-1}\left|\mathbb{E}[\mathcal{T}_{l+1,n}\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]\right|+\mathcal{O}(\eta^{100}).\end{aligned}$$

Noticing that $\mathbb{E}[\mathcal{T}_{l+1,n}\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]=\mathbb{E}[\mathbb{E}[\mathcal{T}_{l+1,n}\mid\hat{\boldsymbol{\phi}}^{(lR_{\mathrm{grp}})},\mathcal{E}_0^{(lR_{\mathrm{grp}})}]\mid\mathcal{E}_0^{(nR_{\mathrm{grp}})}]$, we can apply Theorem 69 and obtain that for all $0\leq n\leq\lfloor T/\eta_{\mathrm{e}}\rfloor$,

$$\begin{aligned}\left|\mathbb{E}[g(\boldsymbol{\phi}^{(nR_{\mathrm{grp}})})]-\mathbb{E}[g(\boldsymbol{\zeta}(n\eta_{\mathrm{e}}))]\right|&\leq nC_{g,1}(\eta_{\mathrm{e}}^{\gamma_1}+\eta_{\mathrm{e}}^{\gamma_2})(\log\tfrac{1}{\eta_{\mathrm{e}}})^b\\ &\leq TC_{g,1}(\eta_{\mathrm{e}}^{\gamma_1-1}+\eta_{\mathrm{e}}^{\gamma_2-1})(\log\tfrac{1}{\eta_{\mathrm{e}}})^b.\end{aligned}$$

Notice that $\eta_{\mathrm{e}}^{\gamma_1}+\eta_{\mathrm{e}}^{\gamma_2}=\eta^{0.5-\beta}+\eta^{\beta}$ and T, $C_{g,1}$ are both constants that are independent of η_{e}. Let $\beta=0.25$ and we have Theorem 65. ■

Having established Theorem 65, we are thus led to prove Theorem 4.

Proof [Proof of Theorem 4] Denote by $s_{\text{cls}} = s_0 + s_1 = \mathcal{O}(\log \frac{1}{\eta})$, which is the time the global iterate $\bar{\boldsymbol{\theta}}^{(s)}$ will reach within $\tilde{\mathcal{O}}(\eta)$ from Γ with high probability. Define $\tilde{\boldsymbol{\zeta}}(t)$ to be the solution to the limiting SDE (101) conditioned on $\mathcal{E}_0^{(s_{\text{cls}})}$ and $\tilde{\boldsymbol{\zeta}}(0) = \boldsymbol{\phi}^{(s_{\text{cls}})}$. By Theorem 65, we have

$$\max_{n=0,\cdots,\lfloor T/\eta^{0.75}\rfloor} \left| \mathbb{E}[g(\boldsymbol{\phi}^{(nR_{\text{grp}}+s_{\text{cls}})}) - g(\tilde{\boldsymbol{\zeta}}(n\eta^{0.75})) \mid \boldsymbol{\phi}^{(s_{\text{cls}})}, \mathcal{E}_0^{(s_{\text{cls}})}] \right| \le C_g \eta^{0.25} (\log \tfrac{1}{\eta})^b,$$

where $R_{\text{grp}} = \lfloor \frac{1}{\alpha\eta^{0.75}} \rfloor$. Noticing that (i) $g \in \mathcal{C}^3$ (ii) $\boldsymbol{b}, \boldsymbol{\sigma} \in \mathcal{C}^\infty$ and (iii) $\boldsymbol{\zeta}(t), \tilde{\boldsymbol{\zeta}}(t) \in \Gamma, t \in [0, \infty)$ almost surely, we can conclude that given $\mathcal{E}_0^{(s_{\text{cls}})}$,

$$\|\boldsymbol{\zeta}(t) - \tilde{\boldsymbol{\zeta}}(t)\|_2 = \tilde{\mathcal{O}}(\sqrt{\eta}), \quad \forall t \in [0, T].$$

Then there exists positive constant b' independent of η and g, and C'_g which is independent of η but can depend on g such that

$$\max_{n=0,\cdots,\lfloor T/\eta^{0.75}\rfloor} \left| \mathbb{E}[g(\boldsymbol{\phi}^{(nR_{\text{grp}}+s_{\text{cls}})}) - g(\boldsymbol{\zeta}(n\eta^{0.75} + s_{\text{cls}} H\eta^2))] \right| \le C'_g \eta^{0.25} (\log \tfrac{1}{\eta})^{b'}.$$

We can view the random variable pairs $\{(\boldsymbol{\phi}^{(nR_{\text{grp}}+s_{\text{cls}})}, \boldsymbol{\zeta}_{n\eta^{0.75}+s_{\text{cls}}\alpha\eta}) : n = 0, \cdots, \lfloor T/\eta^{0.75} \rfloor\}$ as reference points and then approximate the value of $g(\boldsymbol{\phi}^{(s)})$ and $g(\boldsymbol{\zeta}(sH\eta^2))$ with the value at the nearest reference points. By Lemmas 33 and 38, for $0 \le r \le R_{\text{grp}}$ and $0 \le s \le R_{\text{tot}} - r$,

$$\mathbb{E}[\|\boldsymbol{\phi}^{(s+r)} - \boldsymbol{\phi}^{(s)}\|_2] = \tilde{\mathcal{O}}(\eta^{0.375}).$$

Since the values of $\boldsymbol{\phi}^{(s)}$ and $\boldsymbol{\zeta}$ are restricted to a bounded set, $g(\cdot)$ is Lipschitz on that set. Therefore, we have the theorem. ■

Appendix K. Deriving the Slow SDE for Label Noise Regularization

In this this section, we formulate how label noise regularization works and derive the theoretical results in Appendix E.

Consider training a model for C-class classification on dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i$ denotes the input and $y_i \in [C]$ denotes the label. Denote by Δ_+^{C-1} the $(C-1)$-open simplex. Let $f(\boldsymbol{\theta}; \boldsymbol{x}) \in \Delta_+^{C-1}$ be the model output on input $\boldsymbol{x}$ with parameter $\boldsymbol{\theta}$, whose j-th coordinate $f_j(\boldsymbol{\theta}; \boldsymbol{x})$ stands for the probability of $\boldsymbol{x}$ belonging to class j. Let $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$ be the cross entropy loss given input $\boldsymbol{x}$ and label y, i.e, $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) = -\log f_y(\boldsymbol{\theta}; \boldsymbol{x})$.

Adding label noise means replacing the true label y with a fresh noisy label $\hat{y}$ every time we access the sample. Specifically, $\hat{y}$ is set as the true label y with probability $1 - p$ and as any other label with probability $\frac{p}{C-1}$, where p is the fixed corruption probability. The training loss is defined as $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[\ell(\boldsymbol{\theta}; \boldsymbol{x}_i, \hat{y}_i)]$, where the expectation is taken over the stochasticity of $\hat{y}_i$. Notice that given a sample $(\boldsymbol{x}, y)$,

$$\mathbb{E}[\ell(\boldsymbol{\theta}; \boldsymbol{x}, \hat{y})] = -(1-p) \log f_y(\boldsymbol{\theta}; \boldsymbol{x}) - \frac{p}{C-1} \sum_{j \ne y} \log f_j(\boldsymbol{\theta}; \boldsymbol{x}). \tag{107}$$

By the property of cross-entropy loss, (107) attains its global minimum if and only if $f_j = \frac{p}{C-1}$, for all $j \in [C], j \ne y$ and $f_y = 1 - p$. Due to the large expressiveness of modern deep learning models, there typically exists a set $\mathcal{S}^* := \{\boldsymbol{\theta} \mid f_i(\boldsymbol{\theta}) = \mathbb{E}[\hat{y}_i], \forall i \in [N]\}$ such that all elements of $\mathcal{S}^*$ minimizes $\mathcal{L}(\boldsymbol{\theta})$. Then, the manifold Γ is a subset of $\mathcal{S}^*$. The following theorem relates the noise covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i\in[N]} \mathbb{E}[(\nabla\ell(\boldsymbol{\theta};\boldsymbol{x}_i,\hat{y}_i) - \nabla\mathcal{L}(\boldsymbol{\theta}))\,(\nabla\ell(\boldsymbol{\theta};\boldsymbol{x}_i,\hat{y}_i) - \nabla\mathcal{L}(\boldsymbol{\theta}))^\top]$ to the hessian $\nabla^2\mathcal{L}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{S}^*$.

Theorem 70 *If $f(\boldsymbol{\theta};\boldsymbol{x}_i,\hat{y}_i)$ is $\mathcal{C}^2$-smooth on $\mathbb{R}^d$ given any $i \in [N]$, $\hat{y}_i \in [C]$ and $\mathcal{S}^* \neq \varnothing$, then for all $\boldsymbol{\theta} \in \mathcal{S}^*$, $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \nabla^2\mathcal{L}(\boldsymbol{\theta})$.*

Proof Since $\mathcal{L}(\cdot)$ is $\mathcal{C}_2$-smooth, $\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$ for all $\boldsymbol{\theta} \in \mathcal{S}^*$. To prove the above theorem, it suffices to show that $\forall i \in [N]$, $\mathbb{E}[\nabla\ell(\boldsymbol{\theta};\boldsymbol{x}_i,\hat{y}_i)\nabla\ell(\boldsymbol{\theta};\boldsymbol{x}_i,\hat{y}_i)^\top] = \nabla^2\mathcal{L}(\boldsymbol{\theta})$. W.L.O.G, let $y = 1$ and therefore for all $\boldsymbol{\theta} \in S^*$

$$\begin{aligned} f_1(\boldsymbol{\theta};\boldsymbol{x}) &= 1 - p =: a_1, \\ f_j(\boldsymbol{\theta};\boldsymbol{x}) &= \frac{p}{C-1} =: a_2, \forall j > 1, j \in [C]. \end{aligned}$$

Additionally, let $h(x) := -\log(x), x \in \mathbb{R}^+$. The stochastic gradient $\nabla\ell(\boldsymbol{\theta};\boldsymbol{x},\hat{y})$ follows the distribution

$$\nabla\ell(\boldsymbol{\theta};\boldsymbol{x},\hat{y}) = \begin{cases} h'(a_1)\frac{\partial f_1}{\partial\boldsymbol{\theta}} & \text{w.p. } 1-p, \\ h'(a_2)\frac{\partial f_j}{\partial\boldsymbol{\theta}}, & \text{w.p. } \frac{p}{C-1}, \forall j \in [C], j > 1. \end{cases}$$

Then the covariance of the gradient noise is

$$\begin{aligned} \mathbb{E}[\nabla\ell(\boldsymbol{\theta};\boldsymbol{x},\hat{y})\nabla\ell(\boldsymbol{\theta};\boldsymbol{x},\hat{y})^\top] &= (1-p)(h'(a_1))^2 \frac{\partial f_1(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^*}\left(\frac{\partial f_1(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^*}\right)^\top \\ &\quad + \frac{p(h'(a_2))^2}{C-1}\sum_{j>1}\frac{\partial f_j(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^*}\left(\frac{\partial f_j(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^*}\right)^\top. \end{aligned}$$

Now we compute the hessian.

$$\begin{aligned} \nabla^2\mathcal{L}(\boldsymbol{\theta}) &= \underbrace{(1-p)h'(a_1)\frac{\partial^2 f_1}{\partial\boldsymbol{\theta}^2} + \frac{ph'(a_2)}{C-1}\sum_{j>1}\frac{\partial^2 f_j}{\partial\boldsymbol{\theta}^2}}_{\mathcal{T}} \\ &\quad + (1-p)h''(a_1)\frac{\partial f_1}{\partial\boldsymbol{\theta}}\left(\frac{\partial f_1}{\partial\boldsymbol{\theta}}\right)^\top + \frac{ph''(a_2)}{C-1}\sum_{j>1}\frac{\partial f_j}{\partial\boldsymbol{\theta}}\left(\frac{\partial f_j(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^\top. \end{aligned}$$

Since $\sum_{j\in[C]} f_i = 1$,

$$\frac{\partial^2 f_1}{\partial\boldsymbol{\theta}^2} = -\sum_{j>1}\frac{\partial^2 f_j}{\partial\boldsymbol{\theta}^2}. \tag{108}$$

Also, notice that $h'(x) = -\frac{1}{x}$. Therefore,

$$(1-p)h'(a_1) = \frac{ph'(a_2)}{C-1}. \tag{109}$$

Substituting (108) and (109) into the expression of $\mathcal{T}$ gives $\mathcal{T} = \mathbf{0}$, which simplifies $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ as the following form:

$$\nabla^2 \mathcal{L}(\boldsymbol{\theta}) = (1-p)h''(a_1)\frac{\partial f_1}{\partial \boldsymbol{\theta}}\left(\frac{\partial f_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top + \frac{ph''(a_2)}{C-1}\sum_{j>1}\frac{\partial f_j}{\partial \boldsymbol{\theta}}\left(\frac{\partial f_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top.$$

Again notice that $h''(x) = h'(x)$ for all $x \in \mathbb{R}^+$. Therefore, $\nabla^2 \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. ■

With the property $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \nabla^2 \mathcal{L}(\boldsymbol{\theta})$, the limiting SDE (4) can be greatly simplified.

Corollary 71 (Slow SDE for Local SGD with label noise) *For C-class classification task with cross-entropy loss, the slow SDE of Local SGD with label noise has the following form:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{1}{4B}\nabla_\Gamma\left(\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})) + (K-1)\cdot\frac{\mathrm{tr}(F(2H\eta\nabla^2\mathcal{L}(\boldsymbol{\zeta})))}{2H\eta}\right)\mathrm{d}t, \tag{110}$$

where $F(x) := \int_0^x \psi(y)\mathrm{d}y$ and is interpreted as a matrix function in (110). Additionally, $\nabla_\Gamma f$ stands for the gradient of a function f projected to the tangent space of Γ.

Proof Recall the general form of the slow SDE for Local SGD:

$$\mathrm{d}\boldsymbol{\zeta}(t) = \frac{1}{\sqrt{B}}\partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}(t) + \frac{1}{2B}\partial^2\Phi(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}(\boldsymbol{\zeta}) + (K-1)\boldsymbol{\Psi}(\boldsymbol{\zeta})\right]\mathrm{d}t, \tag{111}$$

where $\boldsymbol{\Psi}(\boldsymbol{\zeta})$ is defined in Definition 46. Since for $\boldsymbol{\zeta} \in \Gamma$, $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) = \nabla^2\mathcal{L}(\boldsymbol{\zeta})$, then

$$\partial\Phi(\boldsymbol{\zeta})\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\zeta}) = \mathbf{0}. \tag{112}$$

Now we show that

$$\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{\Sigma}(\boldsymbol{\zeta})] = -\nabla_\Gamma\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})). \tag{113}$$

Since $\nabla^2\mathcal{L}(\boldsymbol{\zeta}) = \boldsymbol{\Sigma}(\boldsymbol{\zeta})$, $\mathcal{V}_{\nabla^2\mathcal{L}(\boldsymbol{\zeta})}[\boldsymbol{\Sigma}] = \frac{1}{2}\boldsymbol{I}$. By Theorem 19,

$$\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{\Sigma}(\boldsymbol{\zeta})] = -\frac{1}{2}\partial\Phi(\boldsymbol{\zeta})\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\boldsymbol{I}] = -\frac{1}{2}\nabla_\Gamma\mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})).$$

Finally, we show that

$$\partial^2\Phi(\boldsymbol{\zeta})[\boldsymbol{\Psi}(\boldsymbol{\zeta})] = -\nabla_\Gamma\frac{1}{2H\eta}\mathrm{tr}(F(2H\eta\nabla^2\mathcal{L}(\boldsymbol{\zeta}))). \tag{114}$$

Define $\hat{\psi}(x) := x\psi(x) = e^{-x} - 1 + x$. By definition of $\Psi(\boldsymbol{\zeta})$, when $\boldsymbol{\Sigma}(\boldsymbol{\zeta}) = \nabla^2\mathcal{L}(\boldsymbol{\zeta})$, $\Psi(\boldsymbol{\zeta}) = \hat{\psi}(2\eta H\nabla^2\mathcal{L}(\boldsymbol{\zeta}))$, where $\hat{\psi}(\cdot)$ is interpreted as a matrix function. Since $\psi(2\eta H\nabla^2\mathcal{L}(\boldsymbol{\zeta})) \in \mathrm{span}\{\boldsymbol{u}\boldsymbol{u}^\top \mid \boldsymbol{u} \in T_{\boldsymbol{\zeta}}^{\perp}(\Gamma)\}$, by Theorem 19,

$$\partial^2\Phi(\boldsymbol{\zeta})[\Psi(\boldsymbol{\zeta})] = -\frac{1}{2}\partial\Phi(\boldsymbol{\zeta})\mathrm{tr}\psi(2\eta H\nabla^2\mathcal{L}(\boldsymbol{\zeta})).$$

By the chain rule, we have (114). Combining (112),(113) and (114) gives the corollary. ■

We further have the following corollary as H goes to infinity.

Lemma 72 *As the number of local steps H goes to infinity, the slow SDE of Local SGD with label noise* (110)*can be simplified as:*

$$\mathrm{d}\boldsymbol{\zeta}(t) = -\frac{K}{4B}\nabla_\Gamma \mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta}))\mathrm{d}t. \tag{115}$$

Proof We obtain the corollary by simply taking the limit. By L'Hospital's rule,

$$\lim_{x\to+\infty}\frac{F(ax)}{x} = \lim_{x\to+\infty}\frac{\mathrm{d}F(ax)}{\mathrm{d}x} = \lim_{x\to+\infty} a\psi(ax) = a.$$

Therefore,

$$\lim_{x\to+\infty}\frac{\mathrm{tr}(F(2H\eta\nabla^2\mathcal{L}(\boldsymbol{\zeta})))}{2H\eta} = \mathrm{tr}(\nabla^2\mathcal{L}(\boldsymbol{\zeta})). \tag{116}$$

Substituting (116) into (110)yields (115). ■

Appendix L. Experimental Details

In this section, we specify the experimental details that are omitted in the main text. Our experiments are conducted on CIFAR-10 [31] and ImageNet [51]. Our implementation of ResNet-56 [18] and VGG-16 [54] is based on the high-starred repository by Wei Yang [2] and we use the implementation of ResNet-50 from torchvision 0.3.1. We run all CIFAR-10 experiments with $B_{\mathrm{loc}} = 128$ on 8 NVIDIA Tesla P100 GPUs while ImageNet experiments are run on 8 NVIDIA A100 GPUS with $B_{\mathrm{loc}} = 32$. All ImageNet experiments are trained with ResNet-50.

We generally adopt the following training strategies. We do not add any momentum unless otherwise stated. We follow the suggestions by Jia et al. [24] and do not add weight decay to the bias and learnable parameters in the normalization layers. For all models with BatchNorm layers, we go through 100 batches of data with batch size B_{loc} to estimate the running mean and variance before evaluation. Experiments on both datasets follow the standard data augmentation pipeline in He et al. [18] except the label noise experiments. Additionally, we use FFCV [32] to accelerate data loading for ImageNet training.

Slightly different from the update rule of Local SGD in Section 1, we use the following sampling scheme unless otherwise stated. At the beginning of every epoch, the whole training dataset is shuffled and evenly partitioned into K shards. Each worker takes one shard and samples batches without replacement. When all workers pass their own shard, the next epoch begins and the whole dataset is reshuffled. An alternative view is that the workers always share the same dataset. For each epoch, they perform local steps by sampling batches of data without replacement until the dataset contains too few data to form a batch. Then another epoch starts with the dataset reloaded to the initial state. This sampling scheme is standard in practice and is also adopted by Lin et al. [39] and Goyal et al. [15].

[2] https://github.com/bearpaw/pytorch-classification

L.1. Post-local SGD Experiments in Section 1

CIFAR-10 experiments. We simulate 32 clients with $B = 4096$. We follow the linear scaling rule and linear learning rate warmup strategy suggested by Goyal et al. [15]. We first run 250 epochs of SGD with the learning rate gradually ramping up from 0.1 to 3.2 for the first 50 epochs. Resuming from the model obtained at epoch 250, we run Local SGD with $\eta = 0.32$. Note that we conduct grid search for the initial learning rate among $\{0.005, 0.01, 0.05, 0.1, 0.15, 0.2\}$ and choose the learning rate with which parallel SGD ($H = 1$) achieves the best test accuracy. We also make sure that the optimal learning rate resides in the middle of the set. The weight decay λ is set as 5×10^{-4}. As for the initialization scheme, we follow Lin et al. [39] and Goyal et al. [15]. Specifically, we use Kaiming Normal [17] for the weights of convolutional layers and initialize the weights of fully-connected layers by a Gaussian distribution with mean zero and standard deviation 0.01. The weights for normalization layers are initialized as one. All bias parameters are initialized as zero. We report the mean and standard deviation over 5 runs.

ImageNet experiments. We simulate 256 workers with $B = 8192$. We follow the linear scaling rule and linear learning rate warmup strategy suggested by Goyal et al. [15]. We first run 100 epochs of SGD where the learning rate linearly ramps up from 0.5 to 16 for the first 5 epochs and then decays by a factor of 0.1 at epoch 50. Resuming from epoch 100, we run Local SGD with $\eta = 0.16$. Note that we conduct grid search for the initial learning rate among $\{0.05, 0.1, 0.5, 1\}$ and choose the learning rate with which parallel SGD ($H = 1$) achieves the best test accuracy. We also make sure that the optimal learning rate resides in the middle of the set. The weight decay λ is set as 1×10^{-4} and we do not add any momentum. The initialization scheme follows the implementation of torchvision 0.3.1. We report the mean and standard deviation over 3 runs.

L.2. Experimental Details for Sections 2 and B.2

CIFAR-10 experiments. We use ResNet-56 for all CIFAR-10 experiments in the two sections. We simulate 32 workers with $B = 4096$ and set the weight decay as 5×10^{-4}. For Figure 2 (a) and (b), we set $\eta = 0.32$, which is the same as the learning rate after decay in Figure 1 (a). For Figure 2 (a), we adopt the same initialization scheme introduced in the corresponding paragraph in Appendix L.1. For Figure 2 (b), (e) and Figure 3 (c), we use the model at epoch 250 in Figure 1 (a) as the pre-trained model. Additionally, we use a training budget of 250 epochs for Figure 2 (e). In Figure 3 (e), we use Local SGD with momentum 0.9, where the momentum buffer is kept locally and never averaged. We run SGD with momentum 0.9 for 150 epochs to obtain the pre-trained model, where the learning rate ramps up from 0.05 to 1.6 linearly in the first 150 epochs. Note that we conduct grid search for the initial learning rate among $\{0.01, 0.05, 0.1, 0.15, 0.2\}$ and choose the learning rate with which parallel SGD ($H = 1$) achieves the highest test accuracy. We also make sure that the optimal learning rate resides in the middle of the set. Resuming from epoch 150, we run Local SGD $H = 1$ (i.e., SGD) and 24 with $\eta = 0.16$ and decay η by 0.1 at epoch 226. For Local SGD $H = 900$, we resume from the model at epoch 226 of $H = 24$ with $\eta = 0.016$. We report the mean and standard deviation over 3 runs for Figure 2 (a) (b) and Figure 3 (c), and over 5 runs for Figure 2 (e).

ImageNet experiments. We simulate 256 clients with $B = 8192$ and set the weight decay as 1×10^{-4}. In Figure 2 (d), both Local SGD and SGD start from the same random initialization. We warm up the learning rate from 0.1 to 3.2 in the first 5 epochs and decay the learning rate by a factor of 0.1 at epochs 50 and 100. For Figure 2 (c) (f) and Figure 3 (d), we use the model at epoch 100 in Figure 1 (b) as the pre-trained model. In Figure 2 (c), we set the learning rate as 0.16, which is the same as the learning rate after epoch 100 in Figure 1 (b). Finally, in Figure 2 (f) and Figure 3 (d), we report the mean and average over 3 runs.

L.3. Experiments on Reducing the Diffusion Term

CIFAR-10 experiments. The model we use is ResNet-56. We first run SGD with batch size 128 and learning rate $\eta = 0.5$ for 250 epochs to obtain the pre-trained model. The initialization scheme is the same as the corresponding paragraph in Appendix L.1. Resuming from epoch 250 with $\eta = 0.05$, we run Local SGD with $K = 16$ until epoch 6000 and run all other setups for the same number of iterations. We report the mean and standard deviation over 3 runs.

ImageNet experiments. We use the model at epoch 100 in Figure 1 (b) as the pre-trained model. Resuming from epoch 100 with $\eta = 0.032$, we run Local SGD with $K = 256$ until epoch 250 and run all other setups for the same number of iterations.

L.4. Local SGD with Label Noise Regularization

For the label noise experiments, we do not use data augmentation and use sampling with replacement. We simulate 32 clients with $B = 4096$ and set the corruption probability as 0.1. Below we list the training details for ResNet-56 and VGG-16 respectively.

ResNet-56. As for the model architecture, we replace the batch normalization layer in Yang's implementation with group normalization such that the training loss is independent of the sampling order. We also use Swish activation [49] in place of ReLU to ensure the smoothness of the loss function. We generate the pre-trained model by running label noise SGD with corruption probability $p = 0.1$ for 500 epochs (6000 iterations). We initialize the model by the same strategy introduced in the first paragraph pf Appendix L.1. Applying the linear warmup scheme proposed by Goyal et al. [15], we gradually ramp up the learning rate η from 0.1 to 3.2 for the first 50 epochs and multiply the learning rate by 0.1 at epoch 250. All subsequent experiments in Figure 5 (a) use learning rate 0.1. The weight decay λ is set as 5×10^{-4} . Note that adding weight decay in the presence of normalization accelerates the limiting dynamics and will not affect the implicit regularization on the original loss function [35].

VGG-16. We follow Yang's implementation of the model architecture except that we replace maximum with average pooling and use Swish activation [49] to make the training loss smooth. We initialize all weight parameters by Kaiming Normal and all bias parameters as zero. The pre-trained model is obtained by running label noise SGD with total batch size 4096 and corruption probability $p = 0.1$ for 6000 iterations. We use a linear learning rate warmup from 0.1 to 0.5 in the first 500 iterations. Resuming from the model obtained by SGD, we use learning rate $\eta = 0.1$. The weight decay λ is set as zero.