Proceedings of Machine Learning Research vol 178:1–48, 2022 35th Annual Conference on Learning Theory

Statistical Estimation and Online Inference via Local SGD

Xiang Li LX10077@PKU.EDU.CN
School of Mathematical Sciences, Peking University

Jiadong Liang JDLIANG@PKU.EDU.CN
School of Mathematical Sciences, Peking University

Xiangyu Chang XIANGYUCHANG@XJTU.EDU.CN
Center for Intelligent Decision-Making and Machine Learning,
School of Management, Xi'an Jiaotong University
Institute for Interdisciplinary Information Core Technology

Zhihua Zhang ZHZHANG@MATH.PKU.EDU.CN
School of Mathematical Sciences, Peking University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We analyze the novel Local SGD in federated Learning, a multi-round estimation procedure that uses intermittent communication to improve communication efficiency. Under a $2+\delta$ moment condition on stochastic gradients, we first establish a *functional central limit theorem* that shows the averaged iterates of Local SGD converge weakly to a rescaled Brownian motion. We next provide two iterative inference methods: the *plug-in* and the *random scaling*. Random scaling constructs an asymptotically pivotal statistic for inference by using the information along the whole Local SGD path. Both the methods are communication efficient and applicable to online data. Our results show that Local SGD simultaneously achieves both statistical efficiency and communication efficiency.

Keywords: Federated Learning, Local SGD, Functional Central Limit Theorem, Statistical Inference

1. Introduction

Federated Learning (FL) is a novel distributed computing paradigm for collaboratively training a global model from data that remote *clients* hold (McMahan et al., 2017). The clients can only cooperate with a central server (e.g., service provider) to train the global model without sharing local datasets. Thus, FL can protect sensitive information that data contain, such as personal identity information and state of health information, from unauthorized access of service providers. The challenge arises when limited data access together with memory constraints, communication budget, and computation restrictions make the traditional statistical estimation and inference methods (Li et al., 2020b; Fan et al., 2021) no longer applicable in the FL scenario. This paper studies how to perform statistical estimation and inference in the FL setting.

A typical FL system considers a pool of K clients, in which the k-th client has a local dataset consisting of i.i.d. samples from some unknown distribution $\mathcal{D}_k$. The central server faces the following distributed optimization problem:

$$\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}) := \sum_{k=1}^{K} p_k \mathbb{E}_{\xi_k \sim \mathcal{D}_k} f_k(\boldsymbol{x}; \xi_k) \right\}, \tag{1}$$

where p_k is the weight of the k-th client and $f_k(\cdot;\xi_k)$ is the user-specified loss with ξ_k being the generated sample from $\mathcal{D}_k$. Thanks to the decentralized nature of data generation, a discrepancy among local data distributions occurs, i.e., $\{\mathcal{D}_k\}_{k=1}^K$ are no longer necessarily identical. In addition, communication is highly restrictive because data with immense volume are scattered across different remote clients.

Many efficient algorithms are proposed to cope with both statistical heterogeneity and expensive communication cost. Perhaps one of the simplest and most celebrated algorithms for FL is *Local SGD* (Stich, 2018). Local SGD runs stochastic gradient descent (SGD) independently in parallel on different clients and averages the sequences only once in a while. Put simple, it learns a shared global model via infrequent communication. It has been shown to have superior performance in training efficiency and scalability (Lin et al., 2018), and converge fast in terms of communication (Li et al., 2019b; Bayoumi et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020a,b; Koloskova et al., 2020). In order to reduce the communication frequency, Local SGD might also be the best choice.

From a statistical viewpoint, it is vital to perform statistical inference in FL because it helps us infer properties of the underlying data distribution. The asymptotic confidence intervals, which becomes more accurate when more samples are observed, help us quantify the uncertainty of our estimator and monitor how our algorithm runs. However, it is still open how to do that and adapt to the peculiarity of FL. In this paper, we would like to address statistical estimation and inference via Local SGD due to its elegant performance mentioned earlier and representativeness in FL. In Local SGD, communication happens at iterations in a prescribed set (denoted $\mathcal{I} = \{t_0, t_1, t_2, \dots\}$). Our goal is to obtain an efficient estimate of $\boldsymbol{x}^* = \mathrm{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ only through the SGD iterates $\{\boldsymbol{x}_{t_m}^k\}_{m\in[T],k\in[K]}$, and provide asymptotic confidence intervals for further inference. Here $[N] = \{1, 2, \dots, N\}$ and $\boldsymbol{x}_t^k$ denotes the parameter hosted by the k-th client at iteration t. Note that we do not have direct access to $\{\boldsymbol{x}_t^k\}_{k\in[K]}$ if $t \notin \mathcal{I}$ due to intermittent communication. It makes the analysis of asymptotic behaviors of Local SGD totally different from that of so-called parallel SGD (Zinkevich et al., 2010), which alternates between one independent step of SGD in parallel and one synchronization. Clearly, the parallel SGD is equivalent to the single-machine SGD, whose asymptotic convergence has been studied extensively (Blum, 1954; Polyak and Juditsky, 1992; Anastasiou et al., 2019; Mou et al., 2020).

Ruppert (1988); Polyak and Juditsky (1992) introduced averaged SGD, a simple modification of SGD where iterates are averaged as the final estimator, and established asymptotic normality via martingale central limit theorem (CLT). It is known that the averaged SGD estimator obtains the optimal asymptotic variance under certain regularity conditions (Duchi and Ruan, 2021). We are motivated to employ the average of Local SGD iterates as the estimator, that is, $\frac{1}{T}\sum_{m=1}^T \bar{\boldsymbol{x}}_{t_m}$ where $\bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^K p_k \boldsymbol{x}_{t_m}^k$. Under common assumptions, we show the proposed estimator $\widehat{\boldsymbol{x}}$ exactly has the optimal asymptotic variance up to a known scale $\nu (\geq 1)$ which is determined by the sequence $\{E_m\}_m$, where $E_m := t_{m+1} - t_m$ is the length of the m-th communication round. And ν barely affects the variance optimality because there exist many diverging sequences $\{E_m\}_m$ satisfying $E_m = o(m)$ and $\nu = 1$. It implies the Local SGD estimator has the optimal asymptotic variance even though it has enlarging communication intermittency. This result somewhat corresponds to the optimization study on Local SGD (Bayoumi et al., 2020; Woodworth et al., 2020a,b, 2021); local updates (i.e., $E_m > 1$) only slow down the L_2 non-asymptotic convergence rate of Local SGD slightly, because the additionally incurred residual error is still dominated by the statistical error. In this case, the averaged communication frequency (ACF, i.e., T/t_T) converges to zero, implying we trade almost all computation for asymptotically zero communication. Therefore, our estimator simultaneously has statistical efficiency and communication efficiency.

To quantify uncertainty, we investigate two online inference methods for statistical inference. One is the plug-in method (Chen et al., 2020), which is available when we have an explicit formula for the covariance matrix of the estimator. The other, a.k.a., random scaling (Lee et al., 2021), borrows insights from time series regression in econometrics (Kiefer et al., 2000; Sun, 2014). It does not attempt to estimate the asymptotic variance but to construct an asymptotically pivotal statistic by normalizing the estimator with its random transformation. To underpins this approach, we establish a functional central limit theorem (FCLT) for the average of Local SGD iterates under much milder conditions than Lee et al. (2021).[1] In particular, we pose a $(2+\delta)$ moment condition on gradient noises (see Assumption 3.2), while Lee et al. (2021) requires a stronger condition: gradient noises should not only be α-mixing but also have at least forth-order moment (see their Assumption 2).[2] Our improvement comes from a specific error decomposition and a careful analysis on a non-asymptotic term with time-varying coefficients (see Lemma B.7). We believe that the advanced proof technique we developed beyond the current work would be of independent interest. We conduct some numerical experiments to illustrate the two inference methods. Due to space limit, they are deferred in Appendix G.

The remainder of this paper is organized as follows. In Section 2 we formulate our problem and introduce Local SGD. In Section 3 we explore the asymptotic properties for the averaged sequence of Local SGD. In Section 4 we introduce two online methods (namely the plug-in method and random scaling) to provide asymptotic confidence intervals and perform hypothesis tests. We provide a proof sketch in Section 5 and review related work in Section 6. We conclude our article in Section 7 with a discussion of our results and future research directions. We illustrate the numerical performance of our methods in synthetic data in Section G. We defer all the proofs to the appendix.

2. Problem Formulation

In this section, we detail some preliminaries to prepare the readers for our results. We are concerned with multi-round distributed learning methods. At iteration t, we use $\boldsymbol{x}_t^k$ to denote the parameter held by the k-th client and ξ_t^k the sample it generates according to $\mathcal{D}_k$. A typical example of multi-round methods is the parallel stochastic gradient descent (P-SGD) (Zinkevich et al., 2010) that runs $\boldsymbol{x}_{t+1}^k = \sum_{k=1}^K p_k \left[\boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)\right]$ for $k \in [K]$ and $t \geq 0$. Other variants have been successively proposed (Jordan et al., 2019; Fan et al., 2019; Chen et al., 2021). It is easy to analyze the statistical property of P-SGD due to its equivalence to the single-machine counterpart. The classical work provides an analysis paradigm for P-SGD, showing it obtains an asymptotically unbiased and efficient estimate (Polyak and Juditsky, 1992). In particular, with $\bar{\boldsymbol{x}}_t = \sum_{k=1}^K p_k \boldsymbol{x}_t^k$, P-SGD achieves the following asymptotic normality with the asymptotic variance satisfying the Cramér-Rao lower bound (Duchi and Ruan, 2021)

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\bar{\boldsymbol{x}}_t - \boldsymbol{x}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\right),$$

1. Note that the standard single-device SGD is a special case of Local SGD by setting $E_m \equiv 1$ and $K = 1$. Thus, our result naturally covers the standard SGD case.
2. The α-mixing assumption forces gradient noises to be asymptotic stationary in a fast rate.

where $\boldsymbol{G} := \nabla^2 f(\boldsymbol{x}^*) = \sum_{k=1}^K p_k \nabla^2 f_k(\boldsymbol{x}^*)$ is the Hessian at the optima $\boldsymbol{x}^*$ and $\boldsymbol{S} = \mathbb{E}(\varepsilon(\boldsymbol{x}^*)\varepsilon(\boldsymbol{x}^*)^\top)$ is the covariance matrix at it. Here $\varepsilon(\boldsymbol{x}^*) = \sum_{k=1}^K p_k \left(\nabla f_k(\boldsymbol{x}^*; \xi_k) - \nabla f_k(\boldsymbol{x}^*)\right)$ is the noise of corresponding aggregated gradients.

2.1. Local SGD

An obvious drawback of P-SGD is its huge communication because it requires synchronization at each iteration. By contrast, Local SGD hopes improve the communication efficiency by lowering the communication frequency (Lin et al., 2018; Stich, 2018; Bayoumi et al., 2020; Woodworth et al., 2020a,b). We now turn to Local SGD and summarize its details. We provide the formal version in Algorithm 1 in Appendix A and related work about Local SGD and its variants in Appendix F. Put simple, it obtains the solution estimate using the following recursive algorithm

$$\boldsymbol{x}_{t+1}^k = \begin{cases} \boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k) & \text{if } t+1 \notin \mathcal{I}, \\ \sum_{k=1}^K p_k \left[\boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)\right] & \text{if } t+1 \in \mathcal{I}, \end{cases} \tag{2}$$

where η_t is the learning rate, ξ_t^k is an independent realization of $\mathcal{D}_k$, and $\mathcal{I}$ denotes the set of communication iterations. At iteration t, each client runs always SGD independently in parallel $\boldsymbol{x}_{t+1}^k = \boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)$. However, when $t+1 \in \mathcal{I}$, the central server aggregates local parameters $\sum_{k=1}^K p_k \boldsymbol{x}_{t+1}^k$ and broadcasts it to all clients, which amounts to the following update rule $\boldsymbol{x}_{t+1}^k = \sum_{k=1}^K p_k \left[\boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)\right]$.

Different choices of $\mathcal{I}$ lead to different communication efficiency for Local SGD. If $\mathcal{I} = \{0, 1, 2, \cdots\}$, then Local SGD is reduced to P-SGD. A famous example in practice is constant communication interval (McMahan et al., 2017), i.e., $\mathcal{I} = \{0, E, 2E, \cdots\}$ for a predefined integer $E(\geq 1)$, which reduces communication frequency from 1 to $1/E$. Local SGD differs from P-SGD if $\mathcal{I}$ has a general form of $\{t_0, t_1, t_2, \cdots\}$ with some $t_m - t_{m-1} > 1$ where t_m is the m-th communication iteration. For example, when $t_m < t < t_{m+1}$ for some m, $\boldsymbol{x}_t^k$ is not likely to equal to $\boldsymbol{x}_t^{k'}$ for $k \neq k'$ due to data heterogeneity, while we always have $\boldsymbol{x}_t^k = \boldsymbol{x}_t^{k'}$ for all k, k' for P-SGD. This difference makes theoretical analysis difficult and different from previous analysis. For seek of simplicity, we assume η_t is a constant when $t_m < t \leq t_{m+1}$ and denote it by η_m with a little abuse of notation (which has been already adopted in Algorithm 1).

3. Statistical Estimation via Local SGD

This section provides asymptotic properties for Local SGD. We start by stating the assumptions needed for the main theoretical results. These assumptions are standard and most of them have been used previously (Polyak and Juditsky, 1992; Su and Zhu, 2018; Chen et al., 2020; Li et al., 2020a).

Assumption 3.1 (Regularity of the objective) *For each $k \in [K]$, we assume the objective function $f_k(\cdot)$ is differentiable and strongly convex with parameter $\mu > 0$, i.e., for any $\boldsymbol{x}, \mathbf{y}$,*

$$f_k(\boldsymbol{x}) \geq f_k(\mathbf{y}) + \langle \nabla f_k(\mathbf{y}), \boldsymbol{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \mathbf{y}\|^2.$$

In addition, each $f_k(\cdot)$ is L-average smooth, i.e.,

$$\sqrt{\mathbb{E}_{\xi_k}\|\nabla f_k(\boldsymbol{x}; \xi_k) - \nabla f_k(\mathbf{y}; \xi_k)\|^2} \leq L\|\boldsymbol{x} - \mathbf{y}\| \tag{3}$$

for some $L > 0$. Finally, the Hessian matrix of the global $f(\cdot)$ exists and is Lipschitz continuous in a neighborhood of the global optimal $\boldsymbol{x}^*$, i.e., there exist some $\delta_1 > 0$ and $L' > 0$ such that

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}^*)\| \le L'\|\boldsymbol{x} - \boldsymbol{x}^*\| \quad \textit{whenever} \quad \|\boldsymbol{x} - \boldsymbol{x}^*\| \le \delta_1.$$

Assumption 3.1 imposes regularity conditions on the objective functions. It requires the global function $f(\cdot)$ to be μ-strongly convex and L-average smooth. The L-average smoothness is stronger than L-smoothness because $\|\nabla f_k(\boldsymbol{x}) - \nabla f_k(\mathbf{y})\| \le \sqrt{\mathbb{E}_{\xi_k}\|\nabla f_k(\boldsymbol{x};\xi_k) - \nabla f_k(\mathbf{y};\xi_k)\|^2} \le L\|\boldsymbol{x} - \mathbf{y}\|$ from Jensen's inequality. The L-average smoothness follows if $\max_{\boldsymbol{x}} \mathbb{E}_{\xi_k}\|\nabla^2 f_k(\boldsymbol{x};\xi_k)\|^2 < \infty$[3] which holds for many statistical learning models such as linear and logistic regression.

Define $\varepsilon_k(\boldsymbol{x}) = \nabla f_k(\boldsymbol{x};\xi_k) - \nabla f_k(\boldsymbol{x})$ as the gradient noise at $\nabla f_k(\boldsymbol{x})$, $\boldsymbol{S}_k = \mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x}^*)\varepsilon_k(\boldsymbol{x}^*)^\top)$, and $\varepsilon(\boldsymbol{x}) = \sum_{k=1}^K p_k\varepsilon_k(\boldsymbol{x})$. Then $\varepsilon_k(\boldsymbol{x})$ (as well as $\varepsilon(\boldsymbol{x})$) has zero mean and its distribution typically depends on $\boldsymbol{x}$. The following assumption regularizes the behavior of each noise ξ_k.

Assumption 3.2 (Regularized gradient noise) *We assume the ξ_k on different devices are independent, though they likely have different distributions. There exists some $C > 0$ such that for each $k \in [K]$,*

$$\left\|\mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x})\varepsilon_k(\boldsymbol{x})^\top) - \boldsymbol{S}_k\right\| \le C\left[\|\boldsymbol{x} - \boldsymbol{x}^*\| + \|\boldsymbol{x} - \boldsymbol{x}^*\|^2\right]. \tag{4}$$

Moreover, we assume there exists a constant $\delta_2 > 0$ such that $\sup_{\boldsymbol{x}} \ \mathbb{E}\|\varepsilon(\boldsymbol{x})\|^{2+\delta_2} < \infty$.

Assumption 3.2 first requites the ξ_k are mutually independent. Note that $\boldsymbol{S} = \sum_{k=1}^K p_k^2\boldsymbol{S}_k$ is the Hessian at the optimum $\boldsymbol{x}^*$ because $\boldsymbol{S} = \sum_{k=1}^K p_k^2\mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x}^*)\varepsilon_k(\boldsymbol{x}^*)^\top) = \mathbb{E}_{\xi}(\varepsilon(\boldsymbol{x}^*)\varepsilon(\boldsymbol{x}^*)^\top)$ from the independence assumption. It then forces the difference between covariance matrices $\mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x})\varepsilon_k(\boldsymbol{x})^\top)$ and $\boldsymbol{S}_k$ controlled by $\|\boldsymbol{x}-\boldsymbol{x}^*\|$. It implies $\left\|\mathbb{E}_{\xi}(\varepsilon(\boldsymbol{x})\varepsilon(\boldsymbol{x})^\top) - \boldsymbol{S}\right\| \le C'\left[\|\boldsymbol{x}-\boldsymbol{x}^*\| + \|\boldsymbol{x}-\boldsymbol{x}^*\|^2\right]$. Finally, the imposed uniformly finite $(2+\delta_2)$ moment of $\varepsilon(\cdot)$ overall $\boldsymbol{x}$ establishes the Lindeberg-Feller condition for martingales, which is much weaker than that used in Lee et al. (2021).

Assumption 3.3 (Slowly decaying effective step sizes) *Define $\gamma_m = E_m\eta_m$ as the effective step size, and assume it is non-increasing in m and satisfies* (i) $\sum_{m=1}^\infty \gamma_m^2 < \infty$*;* (ii) $\sum_{m=1}^\infty \gamma_m = \infty$*; and* (iii) $\frac{\gamma_m - \gamma_{m+1}}{\gamma_m} = o(\gamma_m)$.

In our analysis, $\gamma_m = E_m\eta_m$ serves as the *effective step size*. Indeed, the previous analysis of Li et al. (2019a) shows that the effect of E_m steps of local updates with step-size η_t is similar to one-step update with a larger step-size $E_m\eta_m$. It implies that it is the multiplication of E_m and η_m, rather than either of them alone effecting the convergence. A typical example satisfying the assumption is $\gamma_m = \gamma m^{-\alpha}$ with $\alpha \in (0.5, 1)$, which is also frequently used in previous works (Polyak and Juditsky, 1992; Chen et al., 2020; Su and Zhu, 2018). Because we impose restriction to $\{E_m\}$ latter, in practice, we can first determine the sequence of $\{E_m\}$ and then set $\eta_m = \gamma_m/E_m$ to meet the requirement of $\{\gamma_m\}$.

Assumption 3.4 (Slowly increasing communication intervals) *The sequence $\{E_m\}$ satisfies*

(i) *$\{E_m\}$ is either uniformly bounded or non-decreasing;*

3. This condition is also made by Su and Zhu (2018) to validate (4). See Lemma C.1 therein.

(ii) *There exists some* $\delta_3 > 0$ *such that* $\limsup_{T\to\infty} \frac{1}{T^2}(\sum_{m=0}^{T-1} E_m^{1+\delta_3})(\sum_{m=0}^{T-1} E_m^{-(1+\delta_3)}) < \infty$*;*

(iii) $\lim_{T\to\infty} \frac{1}{T^2}(\sum_{m=0}^{T-1} E_m)(\sum_{m=0}^{T-1} E_m^{-1}) = \nu(\nu \geq 1)$;

(iv) $\lim_{T\to\infty} \frac{\sqrt{t_T}}{T} \cdot \left(\sum_{m=0}^{T} \gamma_m\right) = 0$ *and* $\lim_{T\to\infty} \frac{\sqrt{t_T}}{T} \frac{1}{\sqrt{\gamma_T}} = 0$ *where* $t_T = \sum_{m=0}^{T-1} E_m$.

Assumption 3.4 restricts the growth of $\{E_m\}$. Intuitively, if E_m increases too fast, each $\boldsymbol{x}_t^k$ might converge to their local minimizer $\boldsymbol{x}_k^*$ rapidly before the next communication. Therefore, their average $\bar{\boldsymbol{x}}_t$ is asymptotically biased for $\boldsymbol{x}^*$ with the bias $\sum_{k=1}^K p_k \boldsymbol{x}_k^* - \boldsymbol{x}^*$, which is unlikely zero in FL. Because $\sum_{m=0}^{T-1} \gamma_m \geq \gamma_0$, we have $\sqrt{t_T}/T = \sqrt{\sum_{m=0}^{T-1} E_m}/T \to 0$ from (iv). This, combined with (iii), implies $\sum_{m=0}^{T} E_m^{-1} \to \infty$. It forbids $\{E_m\}$ from growing too fast. In practice, we can choose $E_m \sim \ln m$, $E_m \sim \ln\ln m$ or $E_m \sim m^\beta$ with $\beta \in (0,1)$, all of them satisfying (ii) and (iii). If $\gamma_m \sim m^{-\alpha}$ with $\alpha \in (0.5, 1)$, all the choices of E_m above satisfy (iv).

The following proposition provides another way to check (ii) and (iii) in Assumption 3.4 via investigating the relative difference of E_m and E_{m-1}.

Proposition 3.1 *Assume* $\{E_m\}$ *is non-decreasing. If* $\limsup_{m\to\infty} m(1 - \frac{E_{m-1}}{E_m}) < 1$*, then* (ii) *in Assumption 3.4 holds for some* $\delta_3 > 0$*. Furthermore, if* $\lim_{m\to\infty} m(1 - \frac{E_{m-1}}{E_m})$ *exists (denoted* ρ*), once* $\rho < 1$*, then* (iii) *in Assumption 3.4 holds with* $\nu = \frac{1}{1-\rho^2}$.

According to the aforementioned regularity assumptions, the following asymptotic normality property of the averaged iterates generated by Local SGD is investigated in Theorem 3.1.

Theorem 3.1 (Asymptotic Normality) *Let Assumptions 3.1, 3.2 and 3.3 hold. Then* $\bar{\boldsymbol{x}}_{t_m}$ *converges to* $\boldsymbol{x}^*$ *not only almost surely but also in* L_2 *convergence sense with rate* $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \lesssim \gamma_m$*. Moreover, if Assumption 3.4 holds additionally, the following asymptotic normality follows*

$$\sqrt{t_T}\left(\frac{1}{T}\sum_{m=1}^{T} \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \nu \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\right),$$

where $t_T = \sum_{m=0}^{T-1} E_m$, $\bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^K p_k \boldsymbol{x}_{t_m}^k$, $\boldsymbol{G} = \sum_{k=1}^K p_k \nabla^2 f_k(\boldsymbol{x}^*)$ *is the Hessian matrix at the optima* $\boldsymbol{x}^*$*, and* $\boldsymbol{S}$ *is the covariance matrix of aggregated gradient noise.*

Theorem 3.1 shows that the averaged sequence generated by Local SGD has an asymptotic normal distribution with the asymptotic variance depending on how communication happens (i.e., the sequence $\{E_m\}$) and the problem parameters (i.e., $\boldsymbol{S}$ and $\boldsymbol{G}$). For one thing, the effect of data heterogeneity doesn't show up in the asymptotic normality. The asymptotic variance as well as L_2 convergence rate is the same with that of P-SGD. Technically speaking, this is because the residual error caused by data heterogeneity typically has relatively low order than the statistical error incurred by stochastic gradients (Woodworth et al., 2020b,a). With the choice of γ_m, the residual error vanishes much faster and then seems to disappear. More intuitively, since we set $\gamma_m = E_m \eta_m$ sufficiently small, the effect of E_m steps of local updates using step-size η_m is similar to one-step

Table 1: Statistical efficiency and communication efficiency under different choices of E_m, γ_m and η_m. The statistical efficiency is measured by ν, while the communication efficiency is measured by averaged communication frequency (ACF), i.e., $T/\sum_{m=0}^{T-1} E_m$.

Case	$E_m(\geq 1)$	γ_m	η_m	$\nu(\geq 1)$	ACF
Base	1	$\gamma m^{-\alpha}$ $\alpha \in (0.5, 1)$	$\gamma m^{-\alpha}$	1	1
1	E		$\gamma m^{-\alpha}/E$	1	E^{-1}
2	any $E_m \leq E$		$\gamma m^{-\alpha}/E_m$	1	$[E^{-1}, 1]$
3	$E \ln^\beta m$ $(\beta > 0)$		$\gamma m^{-\alpha}/(E \ln^\beta m)$	1	$E^{-1} \ln^{-\beta} T$
4	$E \ln^\beta \ln m$ $(\beta > 0)$		$\gamma m^{-\alpha}/(E \ln^\beta \ln m)$	1	$E^{-1} \ln^{-\beta} \ln T$
5	Em^β $(\beta \in (0,1))$		$\gamma m^{-(\alpha+\beta)}/E$	$\frac{1}{1-\beta^2}$	$(1+\beta)E^{-1}T^{-\beta}$

update with step-szie γ_m. Hence, Local SGD with step-size η_m actually approximates P-SGD with step-size γ_m. The latter case, as equivalent to single-machine SGD, is unaffected by the statistical heterogeneity and so is Local SGD.

For another thing, it is quite interesting that the whole optimization process affects the asymptotic variance. At the worst case, the way how communication frequency is determined only enlarges the asymptotic variance by a known scale $\nu(\geq 1)$. If $E_m \equiv 1$ for all m (which implies no local update is called), $\nu = 1$ and the result is identical to the typical single-machine central limit theorem (CLT) for SGD (Polyak and Juditsky, 1992). When E_m varies, it is still possible to get communication saved and the asymptotic variance unchanged (i.e., $\nu = 1$) simultaneously (see Table 1). If E_m is uniformly bounded or grows in a rate slower than $E \ln^\beta m(\beta > 0)$, we maintain $\nu = 1$ and obtain a smaller average communication frequency (ACF). In the latter case, the ACF is asymptotic zero, which implies that we trade almost all computation for nearly zero communication without any sacrifice for statistical efficiency. However, if E_m grows like Em^β $(\beta \in (0,1))$, though its ACF decays much more rapidly than that of $E \ln^\beta m$, the asymptotic variance is increased by a factor of $\nu = (1-\beta^2)^{-1}$. It depicts a trade-off between communication efficiency and statistical efficiency when E_m grows too fast. Finally, E_m could not grows like Em^β $(\beta > 1)$ or even exponentially fast, because this will violate the requirement $\sum_{m=0}^{T-1} E_m^{-1} \to \infty$ that is inherent from Assumption 3.4.

4. Statistical Inference via Local SGD

We now conduct statistical inference via Local SGD in the FL setting. As argued in the introduction, the central server only has access to $\{\boldsymbol{x}_t^k\}_{k \in [K]}$ when $t \in \mathcal{I}$. In terms of the established CLT (Theorem 3.1), the average of $\{\bar{\boldsymbol{x}}_{t_m}\}_{m \in [T]}$ achieves an asymptotic normality. Thus it is natural to use $\{\bar{\boldsymbol{x}}_{t_m}\}_{m \in [T]}$ as the main iterate to construct asymptotically valid confidence intervals. We will refer to $\{\bar{\boldsymbol{x}}_{t_m}\}_{m \in [T]}$ as the *path of Local SGD*.

In this section, we assume the data are generated locally in a fully online fashion because it not only can be reduced to the finite-sample setting via bootstrapping, but also covers many realistic FL settings where data are generated sequentially, typical examples including the records of web search, online shopping, and bank credits. In particular, we propose two inference methods depending on whether the second order information of the loss function is available. One is the plug-in method that uses the Hessian information directly and the other is the random scaling method that uses only the information among the path of Local SGD. We also conduct numerical experiments to test the two online inference methods. Due to space limit, we leave them in Appendix G.

4.1. The Plug-in Method

The plug-in method first estimates $\boldsymbol{G}$ and $\boldsymbol{S}$ by $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{S}}$, respectively, and obtains the estimator of the covariance matrix with $\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{G}}^{-\top}$. The key is to obtain consistent estimators $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{S}}$. An intuitive way to construct $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{S}}$ is to use the sample estimate as follows

$$\widehat{\boldsymbol{G}}_T = \frac{1}{T}\sum_{m=1}^{T}\sum_{k=1}^{K} p_k \nabla^2 f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k), \widehat{\boldsymbol{S}}_T = \frac{1}{T}\sum_{m=1}^{T}\left(\sum_{k=1}^{K} p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)\right)\left(\sum_{k=1}^{K} p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)\right)^{\top}.$$

as long as each $\nabla^2 f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)$ is available. Though $\widehat{\boldsymbol{G}}_T$ and $\widehat{\boldsymbol{S}}_T$ are not unbiased for $\boldsymbol{G}$ and $\boldsymbol{S}$, their bias will converge to zero in probability due to $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^*$ almost surely. It is worth noting that with $\bar{\boldsymbol{x}}_{t_m}$, as well as each local Hessian and gradient evaluated at it, communicated to the central server, we can update $\widehat{\boldsymbol{G}}_{m-1}$ to $\widehat{\boldsymbol{G}}_m$ and $\widehat{\boldsymbol{S}}_{m-1}$ to $\widehat{\boldsymbol{S}}_m$. Therefore, they can be computed in an online manner without the need of storing all the data.

Assumption 4.1 *There are some constants $L'' > 0$ such that for any $k \in [K]$,*

$$\mathbb{E}_{\xi_k}\|\nabla^2 f_k(\boldsymbol{x};\xi_k) - \nabla^2 f_k(\boldsymbol{x}^*;\xi_k)\| \le L''\|\boldsymbol{x} - \boldsymbol{x}^*\|.$$

Following Chen et al. (2020), we make Assumption 4.1, which slightly strengthens the Hessian smoothness assumption in Assumption 3.1. Accordingly, we establish the consistency of the sample estimate $\widehat{\boldsymbol{G}}_T$ and $\widehat{\boldsymbol{S}}_T$ in the following theorem.

Theorem 4.1 *Under Assumptions 3.1, 3.2, 3.3 and 4.1, $\widehat{\boldsymbol{G}}_T$ and $\widehat{\boldsymbol{S}}_T$ converge to $\boldsymbol{G}$ and $\boldsymbol{S}$ in probability as $T \to \infty$. As a result of Slutsky's theorem, $\widehat{\boldsymbol{G}}_T^{-1}\widehat{\boldsymbol{S}}_T\widehat{\boldsymbol{G}}_T^{-\top}$ is consistent to $\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$.*

Theorem 4.1 implies that $(\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top})_{jj}$ can be estimated by $\widehat{\sigma}_{T,j}^2 = (\widehat{\boldsymbol{G}}_T^{-1}\widehat{\boldsymbol{S}}_T\widehat{\boldsymbol{G}}_T^{-\top})_{jj}$ for the construction of confidence intervals. Denoting $\bar{\boldsymbol{y}}_T = \frac{1}{T}\sum_{m=1}^{T}\bar{\boldsymbol{x}}_{t_m}$ and $\bar{\boldsymbol{y}}_{T,j}$ its j-th coordinate, we have the following corollary which shows that $\bar{\boldsymbol{y}}_{T,j} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\nu}_T}{t_T}}\widehat{\sigma}_{T,j}$ constructs an asymptotic exact confidence interval for the j-th coordinate of $\boldsymbol{x}^*$. Here $\widehat{\nu}_T$ is any sequence converging to ν.

Corollary 4.1 *Under the assumption of Theorem 4.1,*

$$\mathbb{P}\left(\bar{\boldsymbol{y}}_{T,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\nu}_T}{t_T}}\widehat{\sigma}_{T,j} \le \boldsymbol{x}_j^* \le \bar{\boldsymbol{y}}_{T,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\nu}_T}{t_T}}\widehat{\sigma}_{T,j}\right) \to 1-\alpha,$$

where $\widehat{\nu}_T \to \nu$ and $z_{\frac{\alpha}{2}}$ is $(1-\alpha/2)$-quantile of the standard normal distribution.

We remark that using an estimate $\widehat{\nu}_T$ instead of the true value ν for inference is for the purpose of practice. We find in experiments that directly using the true value ν often results in an unstable confidence interval due to slow convergence of (iii) in Assumption 3.4. As a remedy, we use an estimate $\widehat{\nu}_T = \frac{1}{T^2}(\sum_{m=1}^{T} E_m)(\sum_{m=1}^{T} E_m^{-1})$ which performs better and more stable.

The plug-in method typically works well in practice due to its simplicity and well-established theoretical guarantees. However, it has some drawbacks. The most obvious one is the requirement of the Hessian information, which is not always accessible. Besides, the formulation and sharing of each $\nabla^2 f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)$ requires at least $O(d^2)$ memory and communication cost. Furthermore, it may be computationally expensive when d is large because it involves matrix inversion with computation complexity $O(d^3)$. Finally, the inverse operation is unstable empirically. In practice, we need to set the round T sufficiently large to avoid singularity and ensure stable estimation. The estimator introduced in the next subsection provides a fully online approach, which is cheap in memory, computation, and communication.

4.2. Random Scaling

Random scaling does not attempt to estimate the asymptotic variance, but studentize $\bar{\boldsymbol{y}}_T = \frac{1}{T}\sum_{m=1}^{T} \bar{\boldsymbol{x}}_{t_m}$ with a matrix constructed using iterates along the Local SGD path. In this way, an asymptotically pivotal statistic, though not asymptotically normal, can be obtained. To clarify the method, we should first figure out the asymptotic behavior of the whole Local SGD path rather than its simple average $\bar{\boldsymbol{y}}_T$. In particular, we have the following functional central limit theorem that shows the standardized partial-sum process converges in distribution to a rescaled Brownian motion.

Theorem 4.2 (Functional CLT) *Let Assumptions 3.1, 3.2, 3.3 and 3.4 hold, and define*

$$h(r,T) = \max\left\{ n \in \mathbb{Z}, n > 0 \,\middle|\, r\sum_{m=1}^{T}\frac{1}{E_m} \geq \sum_{m=1}^{n}\frac{1}{E_m} \right\} \quad \textit{for} \quad r \in (0,1].$$

As $T \to \infty$, the following random function weakly converges to a scaled Brownian motion, i.e.,

$$\phi_T(r) := \frac{\sqrt{t_T}}{T}\sum_{m=1}^{h(r,T)} (\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*) \Rightarrow \sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathbf{B}_d(r)$$

where $t_T = \sum_{m=0}^{T-1} E_m$, $\bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^{K} p_k \boldsymbol{x}_{t_m}^k$, and $\mathbf{B}_d(\cdot)$ is the d-dim standard Brownian motion.

Theorem 4.2 has many implications. First, the result is stronger than Theorem 3.1 though under the same assumptions. By applying the continuous mapping theorem to Theorem 4.2 with $\psi : C^d[0,1] \mapsto \psi(1)$, we directly prove Theorem 3.1. Second, the sequence $\{E_m\}$ makes a difference via the time scale $h(r,T)$, which extends previous FCLT results on SGD. For example, if $E_m \equiv E$, then $\nu = 1, t_T = ET$ and $h(r,T) = \lfloor rT \rfloor$, the result turning to be

$$\frac{1}{\sqrt{T}}\sum_{m=1}^{\lfloor rT \rfloor} (\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*) \Rightarrow \sqrt{\frac{1}{E}}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathbf{B}_d(r).$$

When $E = 1$, it reduces to the single-machine result that is recently obtained by Lee et al. (2021). It is worth mentioning that our result requires a much weaker moment condition on gradient noises (i.e., bounded $2 + \delta(\delta > 0)$ moments in Assumption 3.2) than previous Lee et al. (2021). The latter requires that the gradient noises should not only be α-mixing but also have at least forth-order moment (see their Assumption 2). The improvement comes from a specific error decomposition and a careful analysis on a non-asymptotic term with time-varying coefficients (see Lemma B.7). See Section 5 for a sketch of proof ideas. Once $E > 1$, an interesting observation is that local updates reduce the scale of the Brown motion. As an extreme case, the scale vanishes and the Brown motion degenerates when $E = \infty$. It makes sense because when $E = \infty$, $\boldsymbol{x}_{t_m}^k \equiv \boldsymbol{x}_k^*$ and $\bar{\boldsymbol{x}}_{t_m} \equiv \sum_{k=1}^{K} p_k \boldsymbol{x}_{t_m}^k$, the process degenerates. Beyond constant $E_m \equiv E$, Theorem 4.2 also embraces mildly increasing $\{E_m\}$ (see Table 1). Finally, there are some other FCLTs proved via a SDE argument on general stochastic process (Kushner and Yang, 1993) or SGD with constant learning rates (Wang, 2017). By contrast, we consider the particular Local SGD with decaying learning rates in the distributed context and the proof technique (see Section 5 for a short outline) is from a discrete perspective.

With Theorem 4.2, we are ready to describe the inference method. Define $r_0 = 0$ and $r_m = \frac{\sum_{n=1}^{m} \frac{1}{E_n}}{\sum_{n=1}^{T} \frac{1}{E_n}}$ for $m \geq 1$. The choice of r_m satisfies that $\phi_T(r_m) = \frac{\sqrt{t_T}}{T} \sum_{n=1}^{m} (\bar{\boldsymbol{x}}_{t_n} - \boldsymbol{x}^*)$. Note that $\phi_T(1) = \frac{\sqrt{t_T}}{T} \sum_{n=1}^{T} (\bar{\boldsymbol{x}}_{t_n} - \boldsymbol{x}^*) = \sqrt{t_T}(\bar{\boldsymbol{y}}_T - \boldsymbol{x}^*)$. Hence, $\phi_T(r_m) - \frac{m}{T}\phi_T(1) = \frac{\sqrt{t_T}}{T} \sum_{n=1}^{m} (\bar{\boldsymbol{x}}_{t_n} - m\bar{\boldsymbol{y}}_T)$ cancels the dependence on $\boldsymbol{x}^*$. To remove the dependence on the unknown scale $\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}$, we studentize $\phi_T(1)$ via

$$\Pi_T = \sum_{m=1}^{T} \left(\phi_T(r_m) - \frac{m}{T}\phi_T(1)\right) \left(\phi_T(r_m) - \frac{m}{T}\phi_T(1)\right)^\top (r_m - r_{m-1}).$$

Corollary 4.2 *Under the same assumptions of Theorem 4.2 and assuming $g(r_m) \asymp \frac{m}{T}$ for some continuous function g on $[0,1]$, we have that*

$$\phi_T(1)^\top \Pi_T^{-1} \phi_T(1) \xrightarrow{d} \mathbf{B}_d(1)^\top \left[\int_0^1 (\mathbf{B}_d(r) - g(r)\mathbf{B}_d(1)) (\mathbf{B}_d(r) - g(r)\mathbf{B}_d(1))^\top dr\right]^{-1} \mathbf{B}_d(1).$$

This corollary follows immediately from Theorem 4.2 and the continuous mapping theorem. It implies $\phi_T(1)^\top \Pi_T^{-1} \phi_T(1)$ is asymptotically pivotal and thus can be used to construct valid asymptotic confidence intervals. Up to a constant factor, studentizing $\phi_T(1)$ via Π_T is equivalent to studentizing $\bar{\boldsymbol{y}}_T = \frac{1}{T}\sum_{m=1}^{T} \bar{\boldsymbol{x}}_{t_m}$ via $\widehat{\boldsymbol{V}}_T$ where

$$\widehat{\boldsymbol{V}}_T = \frac{1}{T^2 \sum_{m=1}^{T} \frac{1}{E_m}} \sum_{m=1}^{T} \frac{1}{E_m} \left(\sum_{n=1}^{m} \bar{\boldsymbol{x}}_{t_n} - m\bar{\boldsymbol{y}}_T\right) \left(\sum_{n=1}^{m} \bar{\boldsymbol{x}}_{t_n} - m\bar{\boldsymbol{y}}_T\right)^\top.$$

$\widehat{\boldsymbol{V}}_T$ can be updated in an online manner. To state its online updating rule, recall that $\bar{\boldsymbol{y}}_m = \frac{1}{m}\sum_{n=1}^{m} \bar{\boldsymbol{x}}_{t_n}$ and note that

$$\begin{aligned}
\widehat{\boldsymbol{V}}_T &= \frac{1}{T^2 \sum_{m=1}^{T} \frac{1}{E_m}} \sum_{m=1}^{T} \frac{m^2}{E_m} (\bar{\boldsymbol{y}}_m - \bar{\boldsymbol{y}}_T)(\bar{\boldsymbol{y}}_m - \bar{\boldsymbol{y}}_T)^\top \\
&= \frac{1}{T^2 \sum_{m=1}^{T} \frac{1}{E_m}} \left[\sum_{m=1}^{T} \frac{m^2}{E_m} \bar{\boldsymbol{y}}_m \bar{\boldsymbol{y}}_m^\top - \sum_{m=1}^{T} \frac{m^2}{E_m} \bar{\boldsymbol{y}}_T \bar{\boldsymbol{y}}_m^\top - \sum_{m=1}^{T} \frac{m^2}{E_m} \bar{\boldsymbol{y}}_m \bar{\boldsymbol{y}}_T^\top + \sum_{m=1}^{T} \frac{m^2}{E_m} \bar{\boldsymbol{y}}_T \bar{\boldsymbol{y}}_T^\top\right].
\end{aligned}$$

Hence, to update $\widehat{\boldsymbol{V}}_{m-1}$ to $\widehat{\boldsymbol{V}}_m$ when a new observation $\bar{\boldsymbol{x}}_{t_m}$ is available, we only need to keep the following quantities, namely $s_{m-1} = \sum_{n=1}^{m-1} \frac{1}{E_n}$, $q_{m-1} = \sum_{n=1}^{m-1} \frac{n^2}{E_n}$, $\bar{\boldsymbol{y}}_{m-1} = \frac{1}{m-1}\sum_{n=1}^{m-1} \bar{\boldsymbol{x}}_{t_n}$,

$$\boldsymbol{A}_{m-1} = \sum_{n=1}^{m-1} \frac{n^2}{E_n} \bar{\boldsymbol{y}}_n \bar{\boldsymbol{y}}_n^\top \quad \text{and} \quad \boldsymbol{b}_{m-1} = \sum_{n=1}^{m-1} \frac{n^2}{E_n} \bar{\boldsymbol{y}}_n,$$

all of which can be updated in online. In this way, $\widehat{\boldsymbol{V}}_m = \frac{1}{m^2 s_m}\left(\boldsymbol{A}_m - \bar{\boldsymbol{y}}_m \boldsymbol{b}_m^\top - \boldsymbol{b}_m \bar{\boldsymbol{y}}_m^\top + q_m \bar{\boldsymbol{y}}_m \bar{\boldsymbol{y}}_m^\top\right)$. The formal formulation is presented in Algorithm 2 in Appendix A.

Once $\bar{\boldsymbol{y}}_T$ and $\widehat{\boldsymbol{V}}_T$ are obtained, it is straightforward to carry out inference. For example, we construct the $(1-\alpha)$ asymptotic confidence interval for the j-th element $\boldsymbol{x}_j^*$ of $\boldsymbol{x}^*$ as follows

Corollary 4.3 *Under the same conditions of Corollary 4.2, we have that*

$$\mathbb{P}\left(\left[\bar{\boldsymbol{y}}_{T,j} - q_{\frac{\alpha}{2},g}\sqrt{\widehat{\boldsymbol{V}}_{T,jj}} \le \boldsymbol{x}_j^* \le \bar{\boldsymbol{y}}_{T,j} + q_{\frac{\alpha}{2},g}\sqrt{\widehat{\boldsymbol{V}}_{T,jj}}\right]\right) \to 1-\alpha,$$

where $q_{\frac{\alpha}{2},g}$ is $(1-\alpha/2)$-quantile of the following random variable

$$B_1(1) \Big/ \left(\int_0^1 (B_1(r) - g(r)B_1(1))^2 dr\right)^{1/2} \tag{5}$$

with $B_1(\cdot)$ a one-dimensional standard Brownian motion.

If we only care about uncertainty of each coordinate $\boldsymbol{x}_j^*$, for random scaling, we only need to store the diagonal entries of $\widehat{\boldsymbol{V}}_T$ from Corollary 4.3. Both the storage and computation cost are merely $\mathcal{O}(d)$. However, for the plug-in method, the storage cost is $\mathcal{O}(d^2)$ and the computation cost is $\mathcal{O}(d^3)$, since we need to compute and store $\widehat{\boldsymbol{G}}_T$ and $\widehat{\boldsymbol{S}}_T$ and calculate the diagonal entries of $\widehat{\boldsymbol{G}}_T^{-1}\widehat{\boldsymbol{S}}_T\widehat{\boldsymbol{G}}_T^{-\top}$.

The remaining issue is about the specific form of g and the computation of $q_{\alpha,g}$. g actually depends on the growth of $\{E_m\}$. Direct computation reveals that $r_m \asymp \left(\frac{m}{T}\right)^{1-\beta}$ if $E_m \asymp m^\beta$ and $r_m \asymp \frac{m}{T}$ if $E_m \asymp \ln^\beta(m)$. Hence, we are motivated to consider the following family of g: $g_\beta(r) = r^{\frac{1}{1-\beta}}$ indexed by $\beta \in [0,1)$. With this $g_\beta(\cdot)$, we denote the random variable given in (5) by $t^*(\beta)$ and the corresponding critical value by $q_{\alpha,\beta} := \min\{t : \mathbb{P}(t^*(\beta) \le t) \ge 1-\alpha\}$. The limiting distribution $t^*(\beta)$ is mixed normal and symmetric around zero. We compute the critical values of $t^*(\beta)$ via simulation; see Appendix E for more details.

5. Proof Sketch

We provide a short proof sketch for Theorem 4.2 to illustrate our proof technique in this section. A detailed proof is provided in Appendix B.1. Recall $\bar{\boldsymbol{x}}_t = \sum_{k=1}^K p_k \boldsymbol{x}_t^k$. According to the update rule (2), no matter whether communicating or not, we always have $\bar{\boldsymbol{x}}_{t+1} = \bar{\boldsymbol{x}}_t - \eta_m \bar{\boldsymbol{g}}_t$ where $\bar{\boldsymbol{g}}_t = \sum_{k=1}^K p_k \nabla f_k(\boldsymbol{x}_t^k;\xi_t^k)$ for $t_m \le t < t_{m+1}$. Define $\boldsymbol{s}_m = \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*$ and $\gamma_m = \eta_m E_m$ with $E_m = t_{m+1} - t_m$. Iterating over $t = t_m$ to $t_{m+1}-1$ gives $\boldsymbol{s}_{m+1} = \boldsymbol{s}_m - \eta_m \sum_{t=t_m}^{t_{m+1}-1} \bar{\boldsymbol{g}}_t = \boldsymbol{s}_m - \gamma_m \boldsymbol{v}_m$ where $\boldsymbol{v}_m := \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\bar{\boldsymbol{g}}_t$ for short. We decompose $\boldsymbol{v}_m$ as $\boldsymbol{v}_m = \boldsymbol{G}\boldsymbol{s}_m + \boldsymbol{U}_m$ where $\boldsymbol{G} = \nabla^2 f(\boldsymbol{x}^*)$ and $\boldsymbol{U}_m = \boldsymbol{v}_m - \boldsymbol{G}\boldsymbol{s}_m$. Using the notation, we have $\boldsymbol{s}_{m+1} = \boldsymbol{B}_m\boldsymbol{s}_m - \gamma_m \boldsymbol{U}_m$ where $\boldsymbol{B}_m := \boldsymbol{I} - \gamma_m \boldsymbol{G}$ for short. Recursion gives $\boldsymbol{s}_{m+1} = \left(\prod_{j=0}^m \boldsymbol{B}_j\right)\boldsymbol{s}_0 - \sum\limits_{j=0}^m \left(\prod\limits_{i=j+1}^m \boldsymbol{B}_i\right)\gamma_j \boldsymbol{U}_j$. Here we define $\prod\limits_{i=m+1}^m \boldsymbol{B}_i = \boldsymbol{I}$ for any $m \ge 0$. Averaging the last equality over $m=0$ to $h(r,T)$ (define in (12)) and using the notation $\boldsymbol{A}_j^n = \sum\limits_{l=j}^n \left(\prod\limits_{i=j+1}^l \boldsymbol{B}_i\right)$ give $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} \boldsymbol{s}_{m+1} = \frac{\sqrt{t_T}}{T}\left[\frac{1}{\gamma_0}\boldsymbol{A}_0^{h(r,T)}\boldsymbol{B}_0\boldsymbol{s}_0 - \sum_{m=0}^{h(r,T)} \boldsymbol{A}_m^{h(r,T)}\boldsymbol{U}_m\right]$. Roughly speaking, $\boldsymbol{U}_m \approx \boldsymbol{\varepsilon}_m := \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\left[\sum_{k=1}^K p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\right]$. Clearly, by our assumptions, $\{\varepsilon_m\}$ is martingale difference with uniformly bounded $(2+\delta)$ moments. Notice Polyak and Juditsky (1992) implies $\{A_j^n\}_{n\geq j}$ is uniformly bounded and approximates G^{-1} well in the sense that $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n} \|A_j^n - G^{-1}\| = 0$. We are motivated to find $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} s_{m+1}$

$$\approx \frac{\sqrt{t_T}}{T}\left[\frac{1}{\gamma_0}A_0^{h(r,T)}B_0 s_0 - \sum_{m=0}^{h(r,T)} G^{-1}\varepsilon_m - \sum_{m=0}^{h(r,T)} (A_m^T - G^{-1})\varepsilon_m - \sum_{m=0}^{h(r,T)} (A_m^{h(r,T)} - A_m^T)\varepsilon_m\right].$$

The classic result (Hall and Heyde, 2014) implies $\sum_{m=0}^{h(r,T)} G^{-1}\varepsilon_m \implies \sqrt{\nu}G^{-1}S^{1/2}\mathbf{B}_d(r)$. To proceed the proof, we need to show the rest terms converge to zero in probability uniformly in $r\in[0,1]$. It is easy to show $\frac{\sqrt{t_T}}{T\gamma_0}A_0^{h(r,T)}B_0 s_0 = o(1)$ by direct verification and $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(A_m^T - G^{-1})\varepsilon_m = o_{\mathbb{P}}(1)$ by Doob's maximal inequality. Our Lemma B.7 aims to show $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(A_m^{h(r,T)} - A_m^T)\varepsilon_m = o_{\mathbb{P}}(1)$ under the $(2+\delta)$ moment condition, which is the most technical part, because the coefficient $A_m^{h(r,T)} - A_m^T$ varies in r and thus Doob's maximal inequality can't apply directly. To circumvent the issue, Lemma B.7 uses a covering method to show $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(A_m^{h(r,T)} - A_m^T)\varepsilon_m$, as a stochastic process indexed by r, is tight. We divide $[0,1]$ into n equal-width intervals and show the stochastic process has uniformly small expected L_2 norm at each interval with the norm value decreasing as n increases. Thanks to the $(2+\delta)$ moment condition, the sum of norm values on the n intervals is only $\mathcal{O}(n^{-\delta/2})$. Then Lemma B.7 follows from arbitrariness of n. For easy reference, we provide a use-friendly version of Lemma B.7 in Lemma B.8.

6. Related Work

In the context of distributed inference, as we know that no works consider the asymptotic properties of Local SGD or FedAvg, letting alone conduct inference. Most works focus on the optimization properties of Local SGD (or their proposed variants). Woodworth et al. (2020b,a) gave the state-of-the-art convergence analysis for Local SGD in convex settings, showing its convergence rate is dominated by the statistical error incurred by stochastic approximation of gradients. However, it additionally suffers a relatively minor residual error caused by local updates. As a complementary, our work shows that when the *effective step size* is set to $\gamma_m = E_m\eta_m \propto m^{-\alpha}(\alpha\in(0.5,1), m\geq 1)$, Local SGD enjoys the optimal asymptotic variance, even though the communication length increases at a sub-linear rate (i.e., $E_m = o(t_m^{1/2})$). It corresponds to the previous non-asymptotic result (Wang and Joshi, 2018) that shows E_m can be set as large as $O(t_m^{1/2})$ for convergence. Later, Haddadpour et al. (2019) provided a tighter analysis showing E_m can be set as large as $O(t_m^{2/3})$. However, they used a smaller learning rate $\gamma_m \propto m^{-1}$ that cannot guarantee asymptotic normality in our theory. Indeed, the choice of learning rate plays an important role in chasing the non-asymptotic goal of a fast finite-time convergence rate and the asymptotic goal of achieving limiting optimal normality, as noted in Li et al. (2020a) who instead proposed a new SGD variant to achieve both together. In addition, Karimireddy et al. (2020); Liang et al. (2019); Pathak and Wainwright (2020); Zhang et al. (2020) removed the effect of statistical heterogeneity via control variates or primal-dual techniques. From our theory, statistical heterogeneity will not affect the asymptotic variance. Similarly, it has been found that heterogeneity will not alter the minimax optimal bound for the estimation of the commonality parameter (Zhao et al., 2016; Wang et al., 2019).

Statistical estimation and inference via SGD attracts great attention. Ruppert (1988); Polyak and Juditsky (1992) showed averaging iterates along the SGD trajectory has favorable statistical properties in the asymptotic setting, while Anastasiou et al. (2019); Mou et al. (2020) supplemented it with a non-asymptotic analysis. Many papers recently developed iterative algorithms for constructing asymptotically valid confidence intervals (Godichon-Baggioni, 2019). Chen et al. (2020) proposed a consistent plug-in estimator. However, the computation of the Hessian matrix of loss function is not always tractable. Then, Chen et al. (2020) adapted the non-overlapping batch-means method (Glynn and Whitt, 1991) and obtained an offline consistent covariance estimator by using time-increasing batch sizes. Later on, Zhu et al. (2021) extended it to a fully online setting via a recursive counterpart using overlapping batches. In one latest work, Lee et al. (2021) proposed random scaling, which uses nested batches instead. But the analysis in their corrected version requires a stronger condition on the gradient noises that should not only be α-mixing but also have at least forth-order moment (see their Assumption 2). The α-mixing assumption forces gradient noises to be asymptotic stationary in a fast rate. By contrast, we provide a valid analysis for random scaling under only $2 + \delta$ moment assumptions (see Assumption 3.2), which is much weaker and can be of independent interest. We speculate the $(2 + \delta)$ moment condition might not be relaxed any further. In addition, Fang et al. (2018); Fang (2019) proposed online bootstrap procedures for the estimation of confidence intervals via randomly perturbed SGD. Meanwhile, Li et al. (2018); Su and Zhu (2018); Liang and Su (2019) proposed variants of the SGD algorithm to facilitate inference in a non-asymptotic fashion.

7. Conclusion and Future Work

This paper studies how to perform statistical inference via Local SGD in FL. We have established a functional central limit theorem for the averaged iterates of Local SGD and presented two fully online inference methods. We have shown that the Local SGD has statistical efficiency with its asymptotic variance achieving the Cramér–Rao lower bound and communication efficiency with the averaged communication efficiency vanishing asymptotically. It is worth noting that although we considered Local SGD (a distributed variant of SGD), our results also hold for the standard SGD because the latter as a single-device SGD is a special case of Local SGD.

There are many interesting issues for future work. One is to relax the current assumptions and consider Local SGD for more challenging optimization problems (e.g., non-smooth or non-convex problems). Our theory shows that Local SGD enjoys statistical optimality in an asymptotic sense, and it is definitely not also optimal in finite-time convergence (Woodworth et al., 2021). It is then interesting to analyze the statistical properties of other state-of-the-art algorithms in FL. For example, Karimireddy et al. (2020) proposed a new algorithm using control variates to remove the effect of data heterogeneity, which achieves a better non-asymptotic convergence rate. It is also interesting to devise more powerful algorithms as well inference methods to handle the challenge in the decentralized big data era (Fan et al., 2021).

Acknowledgments

Li, Liang and Zhang have been supported by the National Key Research and Development Project of China (No. 2018AAA0101004). Chang was supported in part by the National Natural Science Foundation for Outstanding Young Scholars of China under Grant 72122018 and in part by the Natural Science Foundation of Shaanxi Province under Grant 2021JC-01.

References

Karim M Abadir and Paolo Paruolo. Two mixed normal densities from cointegration analysis. *Econometrica: Journal of the Econometric Society*, pages 671–680, 1997.

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137. PMLR, 2019.

Ahmed Khaled Ragab Bayoumi, 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, 2020.

Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pages 382–386, 1954.

Xi Chen, Jason D Lee, Xin T Tong, Yichen Zhang, et al. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–40, 2021.

YS Chow. A martingale inequality and the law of large numbers. *Proceedings of the American Mathematical Society*, 11(1):107–111, 1960.

SW Dharmadhikari, V Fabian, K Jogdeo, et al. Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, 39(5):1719–1723, 1968.

John C Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *arXiv preprint arXiv:1906.04870*, 2019.

Jianqing Fan, Cong Ma, Kaizheng Wang, and Ziwei Zhu. Modern data modeling: Cross-fertilization of the two cultures. *Observational Studies*, 7(1):65–76, 2021.

Yixin Fang. Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics*, 46(4):987–1002, 2019.

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

Peter W Glynn and Ward Whitt. Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435, 1991.

Antoine Godichon-Baggioni. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19, 2019.

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*, 32:11082–11094, 2019.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Nicholas M Kiefer, Timothy J Vogelsang, and Helle Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, 2000.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

Harold J Kushner and Jichuan Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *arXiv preprint arXiv:2106.03156v3*, 2021.

Chris Junchi Li, Wenlong Mou, Martin J Wainwright, and Michael I Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. *arXiv preprint arXiv:2008.12690*, 2020a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020b.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. In *International Conference on Learning Representations*, 2019a.

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication efficient decentralized training with multiple local updates. *arXiv preprint arXiv:1910.09126*, 2019b.

Tengyuan Liang and Weijie J Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society*, 2019.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Tao Lin, Sebastian U Stich, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

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 (AISTATS)*, 2017.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. *arXiv preprint arXiv:1804.08333*, 2018.

Reese Pathak and Martin J Wainwright. Fedsplit: an algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 7057–7066, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. Federated optimization for heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

Sebastian U Stich. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4447–4458, 2018.

Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

Yixiao Sun. Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics*, 178:659–677, 2014.

Binhuan Wang, Yixin Fang, Heng Lian, and Hua Liang. Additive partially linear models for massive heterogeneous data. *Electronic Journal of Statistics*, 13(1):391–431, 2019.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645. PMLR, 2017.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Yazhen Wang. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *arXiv preprint arXiv:1711.09514*, 2017.

Ward Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media, 2002.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292, 2020b.

Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33, 2020.

Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Qinqing Zheng, Shuxiao Chen, Qi Long, and Weijie Su. Federated f-differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2251–2259. PMLR, 2021.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, (just-accepted):1–30, 2021.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Supplementary Material to "Statistical Estimation and Online Inference via Local SGD"

Appendix A. Formal Version of Algorithms

Algorithm 1 Local SGD

Input: functions $\{f_k\}_{k=1}^K$, initial point $\boldsymbol{x}_0$, step size η_0, communication set $\mathcal{I} = \{t_0, t_1, \cdots\}$.
Initialization: let $\boldsymbol{x}_0^k = \boldsymbol{x}_0$ for all k.
for round $m = 0$ **to** $T-1$ **do**
  for iteration $t = t_m + 1$ **to** t_{m+1} **do**
    for each device $k = 1$ **to** K **do**
      $\boldsymbol{x}_t^k = \boldsymbol{x}_{t-1}^k - \eta_m \nabla f_k(\boldsymbol{x}_{t-1}^k; \xi_{t-1}^k)$. # perform $E_m = t_{m+1} - t_m$ steps of local updates.
    end for
  end for
  The central server aggregates: $\bar{\boldsymbol{x}}_{t_{m+1}} = \sum_{k=1}^K p_k \boldsymbol{x}_{t_{m+1}}^k$.
  Synchronization: $\boldsymbol{x}_{t_{m+1}}^k \leftarrow \bar{\boldsymbol{x}}_{t_{m+1}}$ for all k.
end for
Return: $\widehat{\boldsymbol{x}} = \frac{1}{T}\sum_{m=1}^T \bar{\boldsymbol{x}}_{t_m}$.

Algorithm 2 Online Inference with Local SGD via Random Scaling

Input: functions $\{f_k\}_{k=1}^n$, initial point $\boldsymbol{x}_0$, step size η_t, communication set $\mathcal{I} = \{t_0, t_1, \cdots\}$.
Initialization: set $\boldsymbol{x}_0^{(k)} = \boldsymbol{x}_0$ for all k, let $\boldsymbol{A}_0 = \boldsymbol{0}$ and $\boldsymbol{b}_0 = \boldsymbol{0}$ and $s_0 = q_0 = 0$.
for $m = 1$ **to** T **do**
  Obtain the synchronized variable from Local SGD: $\bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^K p_k \boldsymbol{x}_{t_m}^k$.
  $\bar{\boldsymbol{y}}_m = \frac{m-1}{m}\bar{\boldsymbol{y}}_{m-1} + \frac{1}{m}\bar{\boldsymbol{x}}_{t_m}$,
  $\boldsymbol{A}_m = \boldsymbol{A}_{m-1} + \frac{m^2}{E_m}\bar{\boldsymbol{y}}_m\bar{\boldsymbol{y}}_m^\top$,
  $\boldsymbol{b}_m = \boldsymbol{b}_{m-1} + \frac{m^2}{E_m}\bar{\boldsymbol{y}}_m$,
  $s_m = s_{m-1} + \frac{1}{E_m}$,
  $q_m = q_{m-1} + \frac{m^2}{E_m}$.
  Obtain $\widehat{\boldsymbol{V}}_m$ by

$$\widehat{\boldsymbol{V}}_m = \frac{1}{m^2 s_m}\left(\boldsymbol{A}_m - \bar{\boldsymbol{y}}_m\boldsymbol{b}_m^\top - \boldsymbol{b}_m\bar{\boldsymbol{y}}_m^\top + q_m\bar{\boldsymbol{y}}_m\bar{\boldsymbol{y}}_m^\top\right).$$

  Return: $\bar{\boldsymbol{y}}_m$ **and** $\widehat{\boldsymbol{V}}_m$.
end for

Appendix B. Proofs for the FCLT

This appendix provides a self-contained proof of Theorem 4.2 as well as the first statement of Theorem 3.1.

B.1. Proof Ideas

We follows the perturbed iterate framework that is derived by Mania et al. (2017) and widely used in recent works (Stich, 2018; Stich et al., 2018; Li et al., 2019a; Bayoumi et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020a,b). Then we define a virtual sequence $\bar{\boldsymbol{x}}_t$ in the following way:

$$\bar{\boldsymbol{x}}_t = \sum_{k=1}^K p_k \boldsymbol{x}_t^k.$$

Fix a $m \geq 0$ and consider $t_m \leq t < t_{m+1}$. Local SGD yields that for any device $k \in [K]$,

$$\begin{aligned} \boldsymbol{x}_{t+1}^k &= \boldsymbol{x}_t^k - \eta_m \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k), \\ \boldsymbol{x}_{t_{m+1}}^k &= \sum_{k=1}^K p_k \left(\boldsymbol{x}_{t_{m+1}-1}^k - \eta_m \nabla f_k(\boldsymbol{x}_{t_{m+1}-1}^k; \xi_{t_{m+1}-1}^k) \right), \end{aligned}$$

which implies that we always have

$$\bar{\boldsymbol{x}}_{t+1} = \bar{\boldsymbol{x}}_t - \eta_m \bar{\boldsymbol{g}}_t, \quad \text{where} \quad \bar{\boldsymbol{g}}_t = \sum_{k=1}^K p_k \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k). \tag{6}$$

Define $\boldsymbol{s}_m = \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*$ and recall that $E_m = t_{m+1} - t_m$ and $\gamma_m = \eta_m E_m$. Iterating (6) from $t = t_m$ to $t_{m+1} - 1$ gives

$$\boldsymbol{s}_{m+1} = \boldsymbol{s}_m - \eta_m \sum_{t=t_m}^{t_{m+1}-1} \bar{\boldsymbol{g}}_t = \boldsymbol{s}_m - \gamma_m \boldsymbol{v}_m, \quad \text{where} \quad \boldsymbol{v}_m = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \bar{\boldsymbol{g}}_t. \tag{7}$$

We further decompose $\boldsymbol{v}_m$ into four terms.

$$\begin{aligned} \boldsymbol{v}_m &= \boldsymbol{G}\boldsymbol{s}_m + (\nabla f(\bar{\boldsymbol{x}}_{t_m}) - \boldsymbol{G}\boldsymbol{s}_m) + (\boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m})) + (\boldsymbol{v}_m - \boldsymbol{h}_m) \\ &:= \boldsymbol{G}\boldsymbol{s}_m + \boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m \end{aligned} \tag{8}$$

where $\boldsymbol{G} = \nabla^2 f(\boldsymbol{x}^*)$ is the Hessian at the optimum $\boldsymbol{x}^*$ which is non-singular from our assumption, and

$$\boldsymbol{h}_m = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \sum_{k=1}^K p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m}; \xi_t^k). \tag{9}$$

Note that $\boldsymbol{h}_m$ is almost identical to $\boldsymbol{v}_m$ except that all the stochastic gradients in $\boldsymbol{h}_m$ are evaluated at $\bar{\boldsymbol{x}}_{t_m}$ while those in $\boldsymbol{v}_m$ are evaluated at local variables $\boldsymbol{x}_t^k$'s.

Making use of (7) and (8), we have

$$\boldsymbol{s}_{m+1} = (\boldsymbol{I} - \gamma_m \boldsymbol{G})\boldsymbol{s}_m - \gamma_m(\boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m) := \boldsymbol{B}_m \boldsymbol{s}_m - \gamma_m \boldsymbol{U}_m, \tag{10}$$

where $\boldsymbol{B}_m := \boldsymbol{I} - \gamma_m \boldsymbol{G}$ and $\boldsymbol{U}_m := \boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m$ for short. Recurring (10) gives

$$\boldsymbol{s}_{m+1} = \left(\prod_{j=0}^m \boldsymbol{B}_j \right) \boldsymbol{s}_0 - \sum_{j=0}^m \left(\prod_{i=j+1}^m \boldsymbol{B}_i \right) \gamma_j \boldsymbol{U}_j. \tag{11}$$

Here we use the convention that $\prod_{i=m+1}^{m} \boldsymbol{B}_i = \boldsymbol{I}$ for any $m \geq 0$.

For any $r \in [0,1]$ and $T \geq 1$, define

$$h(r,T) = \max\left\{ n \in \mathbb{Z}_+ \middle| r\sum_{m=1}^{T}\frac{1}{E_m} \geq \sum_{m=1}^{n}\frac{1}{E_m} \right\}. \tag{12}$$

From Assumption 3.4, we know that $\sum_{m=1}^{T}\frac{1}{E_m} \to \infty$ as $T \to \infty$, which implies $h(r,T) \to \infty$ meanwhile. Summing (11) from $m=0$ to $h(r,T)$ gives

$$\begin{aligned}
\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} \boldsymbol{s}_{m+1} &= \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\left[\left(\prod_{j=0}^{m}\boldsymbol{B}_j\right)\boldsymbol{s}_0 - \sum_{j=0}^{m}\left(\prod_{i=j+1}^{m}\boldsymbol{B}_i\right)\gamma_j\boldsymbol{U}_j\right] \\
&= \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\left(\prod_{j=0}^{m}\boldsymbol{B}_j\right)\boldsymbol{s}_0 - \frac{\sqrt{t_T}}{T}\sum_{j=0}^{h(r,T)}\sum_{m=j}^{h(r,T)}\left(\prod_{i=j+1}^{m}\boldsymbol{B}_i\right)\gamma_j\boldsymbol{U}_j.
\end{aligned}\tag{13}$$

Lemma B.1 (Lemma 1 in Polyak and Juditsky (1992)) *Recall that $\boldsymbol{B}_i := \boldsymbol{I} - \gamma_i\boldsymbol{G}$ and $\boldsymbol{G}$ is non-singular. For any $n \geq j$, define $\boldsymbol{A}_j^n$ as*

$$\boldsymbol{A}_j^n = \sum_{l=j}^{n}\left(\prod_{i=j+1}^{l}\boldsymbol{B}_i\right)\gamma_j. \tag{14}$$

Under Assumption 3.3, there exists some universal constant $C_0 > 0$ such that for any $n \geq j \geq 0$, $\|\boldsymbol{A}_j^n\| \leq C_0$. Furthermore, it follows that $\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n}\|\boldsymbol{A}_j^n - \boldsymbol{G}^{-1}\| = 0$.

Using the notation of $\boldsymbol{A}_j^n$, we can further simplify (13) as

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{s}_{m+1} = \frac{\sqrt{t_T}}{T\gamma_0}\boldsymbol{A}_0^{h(r,T)}\boldsymbol{B}_0\boldsymbol{s}_0 - \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{A}_m^{h(r,T)}\boldsymbol{U}_m.$$

Since $\boldsymbol{U}_m = \boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m$, then

$$\begin{aligned}
\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{s}_{m+1} + \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m &= \frac{\sqrt{t_T}}{T\gamma_0}\boldsymbol{A}_0^{h(r,T)}\boldsymbol{B}_0\boldsymbol{s}_0 - \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{A}_m^{h(r,T)}(\boldsymbol{r}_m + \boldsymbol{\delta}_m) \\
&\quad - \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(\boldsymbol{A}_m^T - \boldsymbol{G}^{-1})\boldsymbol{\varepsilon}_m \\
&\quad - \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(\boldsymbol{A}_m^{h(r,T)} - \boldsymbol{A}_m^T)\boldsymbol{\varepsilon}_m \\
&:= \mathcal{T}_0 - \mathcal{T}_1 - \mathcal{T}_2 - \mathcal{T}_3,
\end{aligned}$$

where for simplicity we denote

$$\mathcal{T}_0 = \frac{\sqrt{t_T}}{T\gamma_0}\boldsymbol{A}_0^{h(r,T)}\boldsymbol{B}_0\boldsymbol{s}_0, \quad \mathcal{T}_1 = \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{A}_m^{h(r,T)}(\boldsymbol{r}_m + \boldsymbol{\delta}_m),$$

$$\mathcal{T}_2 = \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(\boldsymbol{A}_m^T - \boldsymbol{G}^{-1})\boldsymbol{\varepsilon}_m, \quad \mathcal{T}_3 = \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}(\boldsymbol{A}_m^{h(r,T)} - \boldsymbol{A}_m^T)\boldsymbol{\varepsilon}_m.$$

With the last equation, we are ready to prove the main theorem which illustrates the partial-sum asymptotic behavior of $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{s}_{m+1}$. The main idea is that we first figure out the partial-sum asymptotic behavior of $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m$ and then show that their difference is uniformly small, i.e.,

$$\sup_{r\in[0,1]}\left\|\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{s}_{m+1} + \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m\right\| = o_{\mathbb{P}}(1).$$

For the second step, it suffices to show that the four separate terms: $\sup_{r\in[0,1]}\|\mathcal{T}_0\|, \sup_{r\in[0,1]}\|\mathcal{T}_1\|$, $\sup_{r\in[0,1]}\|\mathcal{T}_2\|$, and $\sup_{r\in[0,1]}\|\mathcal{T}_4\|$ are $o_{\mathbb{P}}(1)$, respectively. With this idea, our following proof is naturally divided into fives parts.

The establishment of almost sure and L_2 convergence in Lemma B.2 will ease our proof. The following lemma proves the first statement of Theorem 3.1. The second statement of Theorem 3.1 follows directly from Theorem 4.2 which we are going to prove via an argument of the continuous mapping theorem.

Lemma B.2 (Almost sure and L_2 convergence) *Under Assumptions 3.1, 3.2, and 3.3, $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^*$ almost surely when m goes to infinity. In addition, there exists some $\widetilde{C}_0 > 0$ such that*

$$\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \le \widetilde{C}_0\gamma_m.$$

Part 1: Partial-sum asymptotic behavior of $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m$.

Lemma B.3 *Under Assumptions 3.1, 3.2, 3.3 and 3.4, the functional martingale CLT holds, namely, for any $r \in [0,1]$,*

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m \Rightarrow \sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathbf{B}_d(r),$$

where $h(r,T)$ is defined in (12) and $\mathbf{B}_d(r)$ is the d-dimensional standard Brownian motion.

Part 2: Uniform negligibility of $\mathcal{T}_0$. Lemma B.1 characterizes the asymptotic behavior of $\boldsymbol{A}_j^n$. It is uniformly bounded. It implies

$$\sup_{r\in[0,1]}\|\mathcal{T}_0\| = \frac{\sqrt{t_T}}{T\gamma_0}\sup_{r\in[0,1]}\|\boldsymbol{A}_0^{h(r,T)}\boldsymbol{B}_0\boldsymbol{s}_0\| \le \frac{\sqrt{t_T}}{T\gamma_0}C_0\|\boldsymbol{B}_0\boldsymbol{s}_0\| \to 0,$$

as a result of $\frac{\sqrt{t_T}}{T} \to 0$ when $T \to \infty$.

Part 3: Uniform negligibility of $\mathcal{T}_1$. The uniform boundedness of $\boldsymbol{A}_j^n$ implies

$$\sup_{r\in[0,1]}\|\mathcal{T}_1\| = \sup_{r\in[0,1]}\frac{\sqrt{t_T}}{T}\left\|\sum_{m=0}^{h(r,T)}\boldsymbol{A}_m^{h(r,T)}(\boldsymbol{r}_m + \boldsymbol{\delta}_m)\right\|$$

$$
\begin{aligned}
&\leq \sup_{r\in[0,1]} \frac{\sqrt{t_T}}{T} \sum_{m=0}^{h(r,T)} C_0(\|\boldsymbol{r}_m\| + \|\boldsymbol{\delta}_m\|) \\
&= \frac{\sqrt{t_T}}{T} \sum_{m=0}^{T} C_0(\|\boldsymbol{r}_m\| + \|\boldsymbol{\delta}_m\|),
\end{aligned}
$$

where the last inequality uses the fact that $h(r,T)$ increases in r and $h(1,T) = T$. The following two lemmas together imply that $\sup_{r\in[0,1]} \|\mathcal{T}_1\| = o_{\mathbb{P}}(1)$.

Lemma B.4 *Under Assumptions 3.1, 3.2 and 3.3, we have that*

$$
\frac{\sqrt{t_T}}{T} \sum_{m=0}^{T} \|\boldsymbol{r}_m\| = o_{\mathbb{P}}(1).
$$

Lemma B.5 *Under Assumptions 3.1, 3.2 and 3.3, we have that*

$$
\frac{\sqrt{t_T}}{T} \sum_{m=0}^{T} \|\boldsymbol{\delta}_m\| = o_{\mathbb{P}}(1).
$$

Part 4: Uniform negligibility of $\mathcal{T}_2$. By Doob's maximum inequality, it follows that

$$
\begin{aligned}
\mathbb{E} \sup_{r\in[0,1]} \|\mathcal{T}_2\|^2 &= \mathbb{E} \sup_{r\in[0,1]} \frac{t_T}{T^2} \left\| \sum_{m=0}^{h(r,T)} (\boldsymbol{A}_m^T - \boldsymbol{G}^{-1})\boldsymbol{\varepsilon}_m \right\|^2 \\
&\leq \frac{t_T}{T^2} \mathbb{E} \left\| \sum_{m=0}^{T} (\boldsymbol{A}_m^T - \boldsymbol{G}^{-1})\boldsymbol{\varepsilon}_m \right\|^2 \\
&= \frac{t_T}{T^2} \sum_{m=0}^{T} \mathbb{E} \left\| (\boldsymbol{A}_m^T - \boldsymbol{G}^{-1})\boldsymbol{\varepsilon}_m \right\|^2 \\
&\leq \frac{t_T}{T^2} \sum_{m=0}^{T} \left\| \boldsymbol{A}_m^T - \boldsymbol{G}^{-1} \right\|^2 \mathbb{E} \left\| \boldsymbol{\varepsilon}_m \right\|^2.
\end{aligned}
$$

Because $\boldsymbol{\varepsilon}_m = \boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m}) = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} (\nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_t) - \nabla f(\bar{\boldsymbol{x}}_{t_m}))$ is the mean of E_m i.i.d. copies of $\varepsilon(\bar{\boldsymbol{x}}_{t_m}) := \nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})$ at a fixed $\bar{\boldsymbol{x}}_{t_m}$, it implies that

$$
\mathbb{E} \left\| \boldsymbol{\varepsilon}_m \right\|^2 = \frac{1}{E_m} \mathbb{E} \|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^2 \leq \frac{1}{E_m} \left(C_1 + C_2 \mathbb{E} \|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \right) \precsim \frac{1}{E_m}, \tag{15}
$$

where the first inequality is from Lemma B.9 with C_1, C_2 two universal constants defined therein and the second inequality uses Lemma B.2. Using the last result, we have that

$$
\mathbb{E}\mathcal{T}_2 \precsim \frac{t_T}{T^2} \sum_{m=0}^{T} \frac{1}{E_m} \left\| \boldsymbol{A}_m^T - \boldsymbol{G}^{-1} \right\|^2.
$$

By Lemma B.1, it follows that as $T \to \infty$,

$$\frac{1}{T}\sum_{m=0}^{T}\left\|\boldsymbol{A}_m^T - \boldsymbol{G}^{-1}\right\|^2 \le (C_0 + \|\boldsymbol{G}^{-1}\|)\cdot\frac{1}{T}\sum_{m=0}^{T}\left\|\boldsymbol{A}_m^T - \boldsymbol{G}^{-1}\right\| \to 0.$$

Lemma B.6 implies that $\mathbb{E}\sup_{r\in[0,1]}\|\mathcal{T}_2\|^2 = o(1)$.

Lemma B.6 *Let $\{E_m\}$ be the positive-integer-valued sequence that satisfies Assumption 3.4. Let $\{a_{m,T}\}_{m\in[T],T\ge1}$ be a non-negative uniformly bounded sequence satisfying $\lim_{T\to\infty}\frac{1}{T}\sum_{m=0}^{T-1}a_{m,T} = 0$. Then*

$$\lim_{T\to\infty}\frac{(\sum_{m=0}^{T-1}E_m)(\sum_{m=0}^{T-1}E_m^{-1}a_{m,T})}{T^2} = 0.$$

Part 5: Uniform negligibility of $\mathcal{T}_3$. It is subtle to handle $\mathcal{T}_3$ because its coefficient depends on r.

$$\begin{aligned}
\|\mathcal{T}_3\| &= \frac{\sqrt{t_T}}{T}\left\|\sum_{m=0}^{h(r,T)}(\boldsymbol{A}_m^T - \boldsymbol{A}_m^{h(r,T)})\boldsymbol{\varepsilon}_m\right\| \\
&= \frac{\sqrt{t_T}}{T}\left\|\sum_{m=0}^{h(r,T)}\sum_{l=h(r,T)+1}^{T}\left(\prod_{i=m+1}^{l}\boldsymbol{B}_i\right)\gamma_m\boldsymbol{\varepsilon}_m\right\| \\
&= \frac{\sqrt{t_T}}{T}\left\|\sum_{l=h(r,T)+1}^{T}\sum_{m=0}^{h(r,T)}\left(\prod_{i=m+1}^{l}\boldsymbol{B}_i\right)\gamma_m\boldsymbol{\varepsilon}_m\right\| \\
&= \frac{\sqrt{t_T}}{T}\left\|\sum_{l=h(r,T)+1}^{T}\left(\prod_{i=h(r,T)+1}^{l}\boldsymbol{B}_i\right)\sum_{m=0}^{h(r,T)}\left(\prod_{i=m+1}^{h(r,T)}\boldsymbol{B}_i\right)\gamma_m\boldsymbol{\varepsilon}_m\right\| \\
&\precsim \frac{\sqrt{t_T}}{T}\left\|\frac{1}{\gamma_{h(r,T)+1}}\sum_{m=0}^{h(r,T)}\left(\prod_{i=m+1}^{h(r,T)}\boldsymbol{B}_i\right)\gamma_m\boldsymbol{\varepsilon}_m\right\|,
\end{aligned}$$

where the last inequality uses

$$\left\|\sum_{l=h(r,T)+1}^{T}\left(\prod_{i=h(r,T)+1}^{l}\boldsymbol{B}_i\right)\gamma_{h(r,T)+1}\right\| = \left\|\boldsymbol{A}_{h(r,T)+1}^T\boldsymbol{B}_{h(r,T)+1}\right\| \precsim 1.$$

Lemma B.7 shows that $\sup_{r\in[0,1]}\|\mathcal{T}_3\| = o_{\mathbb{P}}(1)$.

Lemma B.7 *Under Assumptions 3.2 and 3.4, it follows that*

$$\sup_{r\in[0,1]}\frac{\sqrt{t_T}}{T}\left\|\frac{1}{\gamma_{h(r,T)+1}}\sum_{m=0}^{h(r,T)}\left(\prod_{i=m+1}^{h(r,T)}\boldsymbol{B}_i\right)\gamma_m\boldsymbol{\varepsilon}_m\right\| = o_{\mathbb{P}}(1).$$

Remark B.1 *There is a more user-friendly version of Lemma B.7 for a plug-and-play use. Define an auxiliary sequence $\{\mathbf{Y}_m\}_{m\ge0}$ as following: $\mathbf{Y}_0 = \mathbf{0}$ and for $m \ge 0$,*

$$\mathbf{Y}_{m+1} = \boldsymbol{B}_m\mathbf{Y}_m + \gamma_m\boldsymbol{\varepsilon}_m = (\boldsymbol{I} - \gamma_m\boldsymbol{G})\mathbf{Y}_m + \gamma_m\boldsymbol{\varepsilon}_m. \tag{16}$$

It is easy to verify that

$$\mathbf{Y}_{t+1} = \sum_{t=0}^{t} \left(\prod_{i=m+1}^{t} \boldsymbol{B}_i \right) \gamma_m \boldsymbol{\varepsilon}_m.$$

Under this notation, Lemma B.7 is equivalent to

$$\sup_{0 \le t \le T} \frac{\sqrt{t_T}}{T} \frac{\|\mathbf{Y}_{t+1}\|}{\gamma_{t+1}} = o_{\mathbb{P}}(1).$$

More formally, we have the following lemma which one can prove from Lemma B.7.

Lemma B.8 *If the martingale difference sequence $\{\boldsymbol{\varepsilon}_m\}_{m\ge0}$ satisfies $\sup_{m\ge0} \mathbb{E}\|\boldsymbol{\varepsilon}_m\|^{2+\delta} < \infty$ for some $\delta > 0$ and Assumption 3.4 holds with $E_m \equiv 1$, for the sequence $\{\mathbf{Y}_m\}_{m\ge0}$ defined in (16) with $\boldsymbol{G}$ positive definite, we have*

$$\sup_{0 \le t \le T} \frac{1}{\sqrt{T}} \frac{\|\mathbf{Y}_{t+1}\|}{\gamma_{t+1}} = o_{\mathbb{P}}(1).$$

B.2. Proof of Lemma B.2

Define $\mathcal{F}_t = \sigma(\{\xi_\tau^k\}_{1\le k\le K, 0\le\tau<t})$ by the natural filtration generated by ξ_τ^k's, so $\{\boldsymbol{x}_t^k\}_t$ is adapted to $\{\mathcal{F}_t\}_t$ and $\{\bar{\boldsymbol{x}}_{t_m}\}_m$ is adapted to $\{\mathcal{F}_{t_m}\}_m$. Notice that $\boldsymbol{v}_m = \boldsymbol{h}_m + \boldsymbol{\delta}_m$ where

$$\boldsymbol{h}_m = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_t) \quad \text{and} \quad \nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_t) = \sum_{k=1}^{K} p_k \nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_t^k),$$

implying $\mathbb{E}[\boldsymbol{h}_m | \mathcal{F}_{t_m}] = \nabla f(\bar{\boldsymbol{x}}_{t_m})$. The L-smoothness of $f(\cdot)$ gives that

$$\begin{aligned} f(\bar{\boldsymbol{x}}_{t_{m+1}}) &\le f(\bar{\boldsymbol{x}}_{t_m}) + \langle \nabla f(\bar{\boldsymbol{x}}_{t_m}), \bar{\boldsymbol{x}}_{t_{m+1}} - \bar{\boldsymbol{x}}_{t_m} \rangle + \frac{L}{2} \|\bar{\boldsymbol{x}}_{t_{m+1}} - \bar{\boldsymbol{x}}_{t_m}\|^2 \\ &= f(\bar{\boldsymbol{x}}_{t_m}) - \gamma_m \langle \nabla f(\bar{\boldsymbol{x}}_{t_m}), \boldsymbol{v}_m \rangle + \frac{\gamma_m^2 L}{2} \|\boldsymbol{v}_m\|^2. \end{aligned}$$

Conditioning on $\mathcal{F}_{t_m}$ in the last inequality gives

$$\begin{aligned} &\mathbb{E}[f(\bar{\boldsymbol{x}}_{t_{m+1}}) | \mathcal{F}_{t_m}] \\ &\quad \le f(\bar{\boldsymbol{x}}_{t_m}) - \gamma_m \langle \nabla f(\bar{\boldsymbol{x}}_{t_m}), \mathbb{E}[\boldsymbol{v}_m | \mathcal{F}_{t_m}] \rangle + \frac{\gamma_m^2 L}{2} \mathbb{E}[\|\boldsymbol{v}_m\|^2 | \mathcal{F}_{t_m}] \\ &\quad = f(\bar{\boldsymbol{x}}_{t_m}) - \gamma_m \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 - \gamma_m \langle \nabla f(\bar{\boldsymbol{x}}_{t_m}), \mathbb{E}[\boldsymbol{\delta}_m | \mathcal{F}_{t_m}] \rangle + \frac{\gamma_m^2 L}{2} \mathbb{E}[\|\boldsymbol{h}_m + \boldsymbol{\delta}_m\|^2 | \mathcal{F}_{t_m}] \\ &\quad \le f(\bar{\boldsymbol{x}}_{t_m}) - \gamma_m \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{\gamma_m}{2} \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{\gamma_m}{2} \|\mathbb{E}[\boldsymbol{\delta}_m | \mathcal{F}_{t_m}]\|^2 \\ &\qquad\quad + \gamma_m^2 L \mathbb{E}[\|\boldsymbol{h}_m\|^2 | \mathcal{F}_{t_m}] + \gamma_m^2 L \mathbb{E}[\|\boldsymbol{\delta}_m\|^2 | \mathcal{F}_{t_m}] \\ &\quad = f(\bar{\boldsymbol{x}}_{t_m}) - \frac{\gamma_m}{2} \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \gamma_m^2 L \mathbb{E}[\|\boldsymbol{h}_m\|^2 | \mathcal{F}_{t_m}] + \left(\frac{\gamma_m}{2} + \gamma_m^2 L \right) \mathbb{E}[\|\boldsymbol{\delta}_m\|^2 | \mathcal{F}_{t_m}], \end{aligned} \tag{17}$$

where we use the conditional Jensen's inequality $\|\mathbb{E}[\boldsymbol{\delta}_m | \mathcal{F}_{t_m}]\|^2 \le \mathbb{E}[\|\boldsymbol{\delta}_m\|^2 | \mathcal{F}_{t_m}]$.

We then bound the last two terms in the right hand side of (17).

Part 1: For $\mathbb{E}[\|\boldsymbol{h}_m\|^2|\mathcal{F}_{t_m}]$, it follows that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{h}_m\|^2|\mathcal{F}_{t_m}] &= \|\mathbb{E}[\boldsymbol{h}_m|\mathcal{F}_{t_m}]\|^2 + \mathbb{E}[\|\boldsymbol{h}_m - \mathbb{E}[\boldsymbol{h}_m|\mathcal{F}_{t_m}]\|^2|\mathcal{F}_{t_m}] \\ &= \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \mathbb{E}[\|\boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2|\mathcal{F}_{t_m}] \\ &= \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{1}{E_m}\mathbb{E}[\|\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2|\mathcal{F}_{t_m}],\end{aligned}$$

where the last equality uses the fact that $\boldsymbol{h}_m$ is the mean of E_m i.i.d. copies of $\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) := \sum_{k=1}^K p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)$ given $\mathcal{F}_{t_m}$, so its conditional variance is E_m times smaller than the latter,

$$\mathbb{E}[\|\boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2|\mathcal{F}_{t_m}] = \frac{1}{E_m}\mathbb{E}[\|\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2|\mathcal{F}_{t_m}]. \tag{18}$$

Lemma B.9 *Recall that $\varepsilon(\bar{\boldsymbol{x}}_{t_m}) := \nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})$ and $\varepsilon_k(\boldsymbol{x}_t^k) := \nabla f(\boldsymbol{x}_t^k;\xi_t^k) - \nabla f(\boldsymbol{x}_t^k)$. Under Assumption 3.2, it follows that*

$$\mathbb{E}_{\xi_t^k}\|\varepsilon_k(\boldsymbol{x}_t^k)\|^2 \le C_1 + C_2\|\boldsymbol{x}_t^k - \boldsymbol{x}^*\|^2 \quad \textit{and} \quad \mathbb{E}_{\xi_{t_m}}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^2 \le C_1 + C_2\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2$$

where $C_1 = d\max_{k\in[K]}\|\boldsymbol{S}_k\| + \frac{dC}{2}$ and $C_2 = \frac{3dC}{2}$ with C defined in Assumption 3.2.

With Lemma B.9, we have

$$\mathbb{E}[\|\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2|\mathcal{F}_{t_m}] \le C_1 + C_2\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2.$$

Then, it follows that

$$\mathbb{E}[\|\boldsymbol{h}_m\|^2|\mathcal{F}_{t_m}] \le \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{C_1}{E_m} + \frac{C_2}{E_m}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2.$$

Part 2: For $\mathbb{E}[\|\boldsymbol{\delta}_m\|^2|\mathcal{F}_{t_m}]$, by Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\delta}_m\|^2|\mathcal{F}_{t_m}] &= \mathbb{E}[\|\boldsymbol{v}_m - \boldsymbol{h}_m\|^2|\mathcal{F}_{t_m}] \\ &= \mathbb{E}\left[\left\|\frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\sum_{k=1}^K p_k\nabla f_k(\boldsymbol{x}_t^k;\xi_t^k) - \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\sum_{k=1}^K p_k\nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)\right\|^2 \middle| \mathcal{F}_{t_m}\right] \\ &\le \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\sum_{k=1}^K p_k\mathbb{E}\left[\left\|\nabla f_k(\boldsymbol{x}_t^k;\xi_t^k) - \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)\right\|^2 |\mathcal{F}_{t_m}\right].\end{aligned}$$

Because $\boldsymbol{x}_t^k, \bar{\boldsymbol{x}}_{t_m} \in \mathcal{F}_t$ and $\mathcal{F}_{t_m} \subseteq \mathcal{F}_t$ for $t_m \le t < t_{m+1}$, we have that

$$\begin{aligned}\mathbb{E}[\|\nabla f_k(\boldsymbol{x}_t^k;\xi_t^k) - \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)\|^2|\mathcal{F}_{t_m}] &= \mathbb{E}[\mathbb{E}[\|\nabla f_k(\boldsymbol{x}_t^k;\xi_t^k) - \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)\|^2|\mathcal{F}_t]|\mathcal{F}_{t_m}] \\ &= \mathbb{E}[\mathbb{E}_{\xi_t^k}\|\nabla f_k(\boldsymbol{x}_t^k;\xi_t^k) - \nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)\|^2|\mathcal{F}_{t_m}] \\ &\le L^2\mathbb{E}[\|\boldsymbol{x}_t^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}],\end{aligned}$$

where the first equality follows from the tower rule of conditional expectation and the second inequality follows from the expected L-smoothness in Assumption 3.1.

Combining the last two results, we have

$$\mathbb{E}[\|\boldsymbol{\delta}_m\|^2|\mathcal{F}_{t_m}] \le \frac{L^2}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\sum_{k=1}^{K} p_k \mathbb{E}[\|\boldsymbol{x}_t^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}] := \frac{L^2}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t,$$

where V_t is the residual error defined by

$$V_t = \sum_{k=1}^{K} p_k \mathbb{E}[\|\boldsymbol{x}_t^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}]. \tag{19}$$

The residual error is incurred by multiple local gradient descents. Intuitively, if no local update is used (i.e., $E_m = 1$), such a residual error would disappear. The following lemma helps us bound $\frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t$ in terms of γ_m and $\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2$.

Lemma B.10 *Under Assumptions 3.1 and 3.2, there exist some universal constants $C_3, C_4 > 0$ such that for any m with $\gamma_m^2 \frac{E_m-1}{E_m} C_4 \le 1$, it follows that*

$$\frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t \le \gamma_m^2 \frac{E_m-1}{E_m}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right).$$

Almost sure convergence: Denote $\Delta_m = f(\bar{\boldsymbol{x}}_{t_m}) - f(\boldsymbol{x}^*)$ for simplicity, then from the μ-strongly convexity and L-smoothness of $f(\cdot)$, it follows that

$$\frac{\mu}{2}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \le \Delta_m \le \frac{1}{2\mu}\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 \quad \text{and} \quad \frac{1}{2L}\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 \le \Delta_m \le \frac{L}{2}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2.$$

Note that $\gamma_m \to 0$ when m goes to infinity, which means there exists some m_0, such that for any $m \ge m_0$, we have $\gamma_m^2 C_4 \le 1$ and $\gamma_m \le \min\{\frac{1}{2L}, 1\}$. It implies that we can apply Lemma B.10 for sufficiently large m. Combining the two parts and plugging them into (17) yield for any $m \ge m_0$,

$$\begin{aligned}
\mathbb{E}[\Delta_{m+1}|\mathcal{F}_{t_m}] &\le \Delta_m - \frac{\gamma_m}{2}\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \gamma_m^2 L \cdot \left[\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{C_1}{E_m} + \frac{C_2}{E_m}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right] \\
&\qquad + \left(\frac{\gamma_m}{2} + \gamma_m^2 L\right)\gamma_m^2 L^2\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right) \\
&\le \Delta_m - \gamma_m\mu\Delta_m + \gamma_m^2 L \cdot \left[\frac{C_1}{E_m} + \left(2L + \frac{2C_2}{\mu E_m}\right)\Delta_m\right] \\
&\qquad + \left(\frac{\gamma_m}{2} + \gamma_m^2 L\right)\gamma_m^2 L^2\left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&\le \Delta_m - \gamma_m\mu\Delta_m + \gamma_m^2 L \cdot \left[C_1 + \left(2L + \frac{2C_2}{\mu}\right)\Delta_m\right] + \gamma_m^3 L^2\left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&\le \Delta_m - \gamma_m\mu\Delta_m + \gamma_m^2 L \cdot \left[C_1 + \left(2L + \frac{2C_2}{\mu}\right)\Delta_m\right] + \gamma_m^2 L^2\left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&= \left(1 + c_1\gamma_m^2\right)\Delta_m + c_2\gamma_m^2 - \mu\gamma_m\Delta_m,
\end{aligned} \tag{20}$$

where

$$c_1 = 2L^2 + \frac{2(LC_2 + L^2C_4)}{\mu} \quad \text{and} \quad c_2 = LC_1 + L^2C_3.$$

To conclude the proof, we need to apply the Robbins-Siegmund theorem (Robbins and Siegmund, 1971).

Lemma B.11 (Robbins-Siegmund theorem) *Let* $\{D_m, \beta_m, \alpha_m, \zeta_m\}_{m=0}^{\infty}$ *be non-negative and adapted to a filtration* $\{\mathcal{G}_m\}_{m=0}^{\infty}$*, satisfying*

$$\mathbb{E}[D_{m+1}|\mathcal{G}_m] \leq (1+\beta_m)D_m + \alpha_m - \zeta_m$$

for all $m \geq 0$ *and both* $\sum_m \beta_m < \infty$ *and* $\sum_m \alpha_m < \infty$ *almost surely. Then, with probability one,* D_m *converges to a non-negative random variable* $D_\infty \in [0, \infty)$ *and* $\sum_m \zeta_m < \infty$.

From Assumption 3.3, we have that $c_1 \sum_{m=m_0}^{\infty} \gamma_m^2 < \infty$ and $c_2 \sum_{m=m_0}^{\infty} \gamma_m^2 < \infty$. Hence, based on (20), Lemma B.11 implies that $\Delta_m = f(\bar{\boldsymbol{x}}_{t_m}) - f(\boldsymbol{x}^*)$ converges to a finite non-negative random variable Δ_∞ almost surely. Moreover, Lemma B.11 also ensures that

$$\mu \sum_{m=m_0}^{\infty} \gamma_m \Delta_m < \infty. \tag{21}$$

If $\mathbb{P}(\Delta_m > 0) > 0$, then the left-hand side of (21) would be infinite with positive probability due to the fact $\sum_{m=m_0}^{\infty} \gamma_m = \infty$. It reveals that $\mathbb{P}(\Delta_m = 0) = 1$ and thus $f(\bar{\boldsymbol{x}}_{t_m}) \to f(\boldsymbol{x}^*)$ as well as $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^*$ with probability one when m goes to infinity.

L_2 **convergence:** We will obtain the L_2 convergence rate from (20). This part follows the same argument of Su and Zhu (2018) (see Page 37-38 therein). For completeness, we conclude this section by presenting the proof of it. Taking expectation on both sides of (20),

$$\frac{\mathbb{E}\Delta_{m+1}}{\gamma_m} \leq \frac{\gamma_{m-1}\left(1 - \mu\gamma_m + c_1\gamma_m^2\right)}{\gamma_m} \frac{\mathbb{E}\Delta_m}{\gamma_{m-1}} + c_2\gamma_m.$$

Because $\gamma_m \to 0$, we have that for sufficiently large m, $c_1\gamma_m^2 \leq 0.5\mu\gamma_m$, and hence,

$$\frac{\mathbb{E}\Delta_{m+1}}{\gamma_m} \leq \frac{\gamma_{m-1}\left(1 - \frac{\mu}{2}\gamma_m\right)}{\gamma_m} \frac{\mathbb{E}\Delta_m}{\gamma_{m-1}} + c_2\gamma_m.$$

Lemma B.12 (Lemma A.10 in Su and Zhu (2018)) *Let* c_1, c_2 *be arbitrary positive constants. Assume* $\gamma_m \to 0$ *and* $\frac{\gamma_{m-1}}{\gamma_m} = 1 + o(\gamma_m)$*. If* $B_m > 0$ *satisfies* $B_m \leq \frac{\gamma_{m-1}(1-c_1\gamma_m)}{\gamma_m} B_{m-1} + c_2\gamma_m$*, then* $\sup_m B_m < \infty$.

With the above lemma, we claim that there exists some $C_5 > 0$ such that

$$\sup_{0<m<\infty} \frac{\mathbb{E}\Delta_m}{\gamma_{m-1}} < C_5, \tag{22}$$

which immediately concludes that

$$\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \leq \frac{2}{\mu}\mathbb{E}\Delta_m \leq \frac{2C_5}{\mu}\gamma_{m-1} = \frac{2C_5}{\mu}(1 + o(\gamma_m))\gamma_m \leq C_0\gamma_m.$$

B.3. Proof of Lemma B.9

Proof By Assumption 3.2, we know that $\varepsilon(\bar{\boldsymbol{x}}_{t_m}) := \nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})$ satisfies

$$\|\mathbb{E}_{\xi_{t_m}} \varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top - \boldsymbol{S}\| \le C\left(\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\| + \|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right).$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{E}[\|\nabla f(\bar{\boldsymbol{x}}_{t_m}; \xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 | \mathcal{F}_{t_m}] &= \mathbb{E}[\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^2 | \mathcal{F}_{t_m}] = \mathbb{E}_{\xi_{t_m}} \|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^2 \\
&= \operatorname{tr}(\mathbb{E}_{\xi_{t_m}} \varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top) \\
&\le d\|\mathbb{E}_{\xi_{t_m}} \varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top\| \\
&\le d\|\boldsymbol{S}\| + dC\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\| + dC\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \\
&\le \left(d\|\boldsymbol{S}\| + \frac{dC}{2}\right) + \frac{3dC}{2}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \\
&\le C_1 + C_2\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2
\end{aligned}$$

with $C_1 = d\max_k \|\boldsymbol{S}_k\| + \frac{dC}{2}$ and $C_2 = \frac{3dC}{2}$. Here we use the fact that $\boldsymbol{S} = \sum_{k=1}^K p_k^2 \boldsymbol{S}_k$ and thus $\|\boldsymbol{S}\| \le \sum_{k=1}^K p_k^2 \|\boldsymbol{S}_k\| \le \sum_{k=1}^K p_k \|\boldsymbol{S}_k\| \le \max_{k\in[K]} \|\boldsymbol{S}_k\|$.

With a similar argument, it follows that

$$\mathbb{E}_{\xi_t^k} \|\varepsilon_k(\boldsymbol{x}_t^k)\|^2 \le d\|\boldsymbol{S}_k\| + \frac{dC}{2} + \frac{3dC}{2}\|\boldsymbol{x}_t^k - \boldsymbol{x}^*\|^2 \le C_1 + C_2\|\boldsymbol{x}_t^k - \boldsymbol{x}^*\|^2.$$

■

B.4. Proof of Lemma B.10

For a fixed $m \ge 0$, let us consider the case where $t_{m+1} > t_m + 1$, otherwise the result follows directly due to $V_{t_m} = 0$. For $t_m \le t < t_{m+1} - 1$ and $k \in [K]$, we have $\boldsymbol{x}_{t_m}^k = \bar{\boldsymbol{x}}_{t_m}$ and

$$\boldsymbol{x}_{t+1}^k = \boldsymbol{x}_t^k - \eta_m \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k) \quad \Rightarrow \quad \boldsymbol{x}_{t+1}^k = \bar{\boldsymbol{x}}_{t_m} - \eta_m \sum_{\tau=t_m}^{t} \nabla f_k(\boldsymbol{x}_\tau^k; \xi_\tau^k).$$

Using the last iteration relation, we obtain that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_{t+1}^k - \bar{\boldsymbol{x}}_{t_m}\|^2 | \mathcal{F}_{t_m}] &= \eta_m^2 \mathbb{E}\left[\left\|\sum_{\tau=t_m}^{t} \nabla f_k(\boldsymbol{x}_\tau^k; \xi_\tau^k)\right\|^2 \middle| \mathcal{F}_{t_m}\right] \\
&\le \eta_m^2 (t + 1 - t_m) \sum_{\tau=t_m}^{t} \mathbb{E}[\|\nabla f_k(\boldsymbol{x}_\tau^k; \xi_\tau^k)\|^2 | \mathcal{F}_{t_m}] \\
&\le \eta_m^2 E_m \sum_{\tau=t_m}^{t} \mathbb{E}[\|\nabla f_k(\boldsymbol{x}_\tau^k; \xi_\tau^k)\|^2 | \mathcal{F}_{t_m}] \\
&= \eta_m^2 E_m \sum_{\tau=t_m}^{t} \mathbb{E}\left[\mathbb{E}(\|\nabla f_k(\boldsymbol{x}_\tau^k; \xi_\tau^k)\|^2 | \mathcal{F}_\tau) \middle| \mathcal{F}_{t_m}\right].
\end{aligned}$$

We then turn to bound $\mathbb{E}[\|\nabla f_k(\boldsymbol{x}_\tau^k;\xi_\tau^k)\|^2 \big| \mathcal{F}_\tau]$ as follows:

$$\begin{aligned}
\mathbb{E}[\|\nabla f_k(\boldsymbol{x}_\tau^k;\xi_\tau^k)\|^2 \big| \mathcal{F}_\tau] &= \mathbb{E}[\|\nabla f_k(\boldsymbol{x}_\tau^k;\xi_\tau^k) - \nabla f_k(\boldsymbol{x}_\tau^k)\|^2 \big| \mathcal{F}_\tau] + \|\nabla f_k(\boldsymbol{x}_\tau^k)\|^2 \\
&\le \mathbb{E}_{\xi_\tau^k}\|\varepsilon_k(\boldsymbol{x}_\tau^k)\|^2 + 2\|\nabla f_k(\boldsymbol{x}_\tau^k) - \nabla f_k(\boldsymbol{x}^*)\|^2 + 2\|\nabla f_k(\boldsymbol{x}^*)\|^2 \\
&\le \left(C_1 + 2\|\nabla f_k(\boldsymbol{x}^*)\|^2\right) + \left(C_2 + 2L^2\right)\|\boldsymbol{x}_\tau^k - \boldsymbol{x}^*\|^2 \\
&\le C_3 + \frac{C_4}{2}\|\boldsymbol{x}_\tau^k - \boldsymbol{x}^*\|^2 \\
&\le C_3 + C_4\|\boldsymbol{x}_\tau^k - \bar{\boldsymbol{x}}_{t_m}\|^2 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2,
\end{aligned}$$

where $C_3 = C_1 + 2\max_{k\in[K]}\|\nabla f_k(\boldsymbol{x}^*)\|^2$ and $C_4 = 2C_2 + 4L^2$. The second inequality uses the L-smoothness to bound $\|\nabla f_k(\boldsymbol{x}_\tau^k) - \nabla f_k(\boldsymbol{x}^*)\|$ and Lemma B.9 to bound $\mathbb{E}_{\xi_\tau^k}\|\varepsilon_k(\boldsymbol{x}_\tau^k)\|^2$ which yields

$$\mathbb{E}_{\xi_\tau^k}\|\varepsilon_k(\boldsymbol{x}_\tau^k)\|^2 \le C_1 + C_2\|\boldsymbol{x}_\tau^k - \boldsymbol{x}^*\|^2.$$

Therefore, by combing the last two results, we have

$$\mathbb{E}[\|\boldsymbol{x}_{t+1}^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}] \le \eta_m^2 E_m \sum_{\tau=t_m}^{t}\left[C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + C_4\mathbb{E}[\|\boldsymbol{x}_\tau^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}]\right].$$

Hence, for $t_m \le t < t_{m+1} - 1$, we have

$$V_{t+1} = \sum_{k=1}^{K} p_k\mathbb{E}(\|\boldsymbol{x}_{t+1}^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathcal{F}_{t_m}) \le \eta_m^2 E_m \sum_{\tau=t_m}^{t}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + C_4 V_\tau\right). \tag{23}$$

Because $V_{t_m} = 0$, it then follows that

$$\begin{aligned}
\frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t &= \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-2} V_{t+1} \\
&\le \eta_m^2 \sum_{t=t_m}^{t_{m+1}-2}\sum_{\tau=t_m}^{t}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + C_4 V_\tau\right) \\
&= \eta_m^2 \sum_{t=t_m}^{t_{m+1}-2}\left(t_{m+1} - t - 1\right)\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + C_4 V_t\right) \\
&\le \eta_m^2(E_m - 1)\sum_{t=t_m}^{t_{m+1}-1}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + C_4 V_t\right) \\
&\le \gamma_m^2\frac{E_m - 1}{E_m}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 + \frac{C_4}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t\right),
\end{aligned}$$

where we use the definition of $E_m = t_{m+1} - t_m$ and $\gamma_m = \eta_m E_m$.

Hence, rearranging the last inequality and using the condition $\gamma_m^2\frac{E_m-1}{E_m}C_4 \le \frac{1}{2}$ gives

$$\frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t \le 2\gamma_m^2\frac{E_m - 1}{E_m}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right).$$

Finally redefining $C_3 := 2C_3$ and $C_4 := 2C_4$ completes the proof and the restriction on γ_m becomes $\gamma_m^2\frac{E_m-1}{E_m}C_4 \le 1$ under the new notation of C_4.

B.5. Proof of Lemma B.3

Recall that

$$\boldsymbol{\varepsilon}_m = \boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m}) = \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\left(\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_t) - \nabla f(\bar{\boldsymbol{x}}_{t_m})\right)$$

where $\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_t) = \sum_{k=1}^K p_k \nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_t^k)$ and $\xi_t = \{\xi_t^k\}_{k\in[K]}$, and recall that $\varepsilon(\bar{\boldsymbol{x}}_{t_m}) = \nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\bar{\boldsymbol{x}}_{t_m})$. Hence $\boldsymbol{\varepsilon}_m$ is the mean of E_m i.i.d. copies of $\varepsilon(\bar{\boldsymbol{x}}_{t_m})$ at a fixed $\bar{\boldsymbol{x}}_{t_m}$.

Define $\mathcal{F}_t = \sigma(\{\xi_\tau^k\}_{1\le k\le K, 0\le\tau<t})$ by the natural filtration generated by ξ_τ^k's and $\mathcal{G}_{m-1} = \mathcal{F}_{t_m}$. Then $\{\boldsymbol{\varepsilon}_m\}_{m=1}^\infty$ is a martingale difference with respect to $\{\mathcal{G}_m\}_{m=0}^\infty$ (for convention $\mathcal{G}_0 = \{\emptyset,\Omega\}$ if $\bar{\boldsymbol{x}}_0$ is deterministic, otherwise $\mathcal{G}_0 = \sigma(\bar{\boldsymbol{x}}_0)$): $\mathbb{E}[\boldsymbol{\varepsilon}_m|\mathcal{G}_{m-1}] = \mathbf{0}$.

The following lemma establishes an invariance principle which allows us to extend traditional martingale CLT. Interesting readers can find its proof in Hall and Heyde (2014) (see Theorems 4.1, 4.2 and 4.4 therein).

Lemma B.13 (Invariance principles in the martingale CLT) *Let $\{S_n,\mathcal{G}_n\}_{n\ge1}$ be a zero-mean, square-integrable martingale with difference $X_n = S_n - S_{n-1} (S_0 = 0)$. Let $U_n^2 = \sum_{m=1}^n \mathbb{E}[X_m^2|\mathcal{G}_{m-1}]$ and $s_n^2 = \mathbb{E}U_n^2 = \mathbb{E}S_n^2$. Define $\zeta_n(t)$ as the linear interpolation among the points $(0,0)$, $(U_n^{-2}U_1^2, U_n^{-1}S_1)$, $(U_n^{-2}U_2^2, U_n^{-1}S_2)$, …, $(1, U_n^{-1}S_n)$, namely, for $t\in[0,1]$ and $0\le i\le n-1$,*

$$\zeta_n(t) := U_n^{-1}\left[S_i + (U_{i+1}^2 - U_i^2)^{-1}(tU_n^2 - U_i^2)X_{i+1}\right] \quad \text{if} \quad U_i^2 \le tU_n^2 < U_{i+1}^2.$$

As $n\to\infty$, if (i) the Linderberg conditions holds, namely for any $\varepsilon>0$,

$$s_n^{-2}\sum_{m=1}^n \mathbb{E}[X_m^2 I(|X_m|\ge\varepsilon s_n)] \to 0, \tag{24}$$

and (ii) $s_n^{-2}U_n^2 \to 1$ almost surely and $s_n^2\to\infty$, then

$$\zeta_n(t) \Rightarrow B(t) \quad \text{in the sense of} \quad (C,\rho).$$

Here $B(t)$ is the standard Brownian motion on $[0,1]$ and $C = C[0,1]$ is the space of real, continuous functions on $[0,1]$ with the uniform metric $\rho: C[0,1]\to[0,\infty)$, $\rho(\omega) = \max_{t\in[0,1]}|\omega(t)|$.

Lemma B.13 is for univariate martingales. We will use the Cramér-Wold device to reduce the issue of convergence of multivariate martingales to univariate ones. To that end, we fix any uni-norm vector $\boldsymbol{a}$ and define $X_m = \boldsymbol{a}^\top\boldsymbol{\varepsilon}_m$. We then check the two conditions in Lemma B.13 hold for such $\{X_m,\mathcal{G}_m\}_{m\ge1}$.

The Linderberg condition: For one thing, since $\bar{\boldsymbol{x}}_{t_m}\to\boldsymbol{x}^*$ almost surely from Lemma B.2, we have $\mathbb{E}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^{2+\delta_2} \precsim 1$ from Assumption 3.2 when m is sufficiently large.

Lemma B.14 (Marcinkiewicz–Zygmund inequality and Burkholder inequality) *If $Z_1,\dots,Z_n$ are independent random vectors such that $\mathbb{E}Z_m = 0$ and $\mathbb{E}|Z_m|^p<\infty$ for $1\le p<\infty$, then*

$$\mathbb{E}\left|\frac{1}{n}\sum_{m=1}^n Z_m\right|^p \le \frac{C_p}{n^{\frac{p}{2}}}\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^n |Z_m|^2\right)^{\frac{p}{2}},$$

where the C_p are positive constants which depend only on p and not on the underlying distribution of the random variables involved. If $Z_1,\dots,Z_n$ are martingale difference sequence, the above inequality still holds. It is named as Burkholder's inequality (Dharmadhikari et al., 1968).

Notice that we can rewrite X_m as the mean of E_m i.i.d. random variables which have the same distribution as $Z_1 = \boldsymbol{a}^\top \varepsilon(\bar{\boldsymbol{x}}_{t_m})$: $X_m = \frac{1}{E_m}\sum_{i=1}^{E_m} Z_i$. With the Marcinkiewicz–Zygmund inequality and Jensen inequality, it follows that

$$\begin{aligned}\mathbb{E}|X_m|^{2+\delta_2} &\precsim E_m^{-(1+\frac{\delta_2}{2})}\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^{n}|Z_m|^2\right)^{1+\frac{\delta_2}{2}} \precsim E_m^{-(1+\frac{\delta_2}{2})}\mathbb{E}\left|Z_1\right|^{2+\delta_2} \\ &\precsim E_m^{-(1+\frac{\delta_2}{2})}\|\boldsymbol{a}\|^{2+\delta_2}\mathbb{E}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^{2+\delta_2} \precsim E_m^{-1}. \end{aligned} \tag{25}$$

Moreover, from Assumption 3.2 and Lemma B.2, we have that

$$\begin{aligned}\left|\boldsymbol{a}^\top\left[\mathbb{E}\varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top - \boldsymbol{S}\right]\boldsymbol{a}\right| &\le C\left[\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - x^*\| + \mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - x^*\|^2\right] \\ &\le C(\sqrt{\gamma_m} + \gamma_m) \to 0.\end{aligned}$$

Recall that $\sum_{m=1}^T {E_m}^{-1} \to \infty$ as $T \to \infty$. The Stolz–Cesàro theorem (Lemma B.15) implies that

$$\lim_{T\to\infty}\frac{s_T^2}{\sum_{m=1}^T \frac{1}{E_m}\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = \lim_{T\to\infty}\frac{\sum_{m=1}^T \frac{\boldsymbol{a}^\top\mathbb{E}\varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top\boldsymbol{a}}{E_m}}{\sum_{m=1}^T \frac{1}{E_m}\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = \lim_{T\to\infty}\frac{\boldsymbol{a}^\top\mathbb{E}\varepsilon(\bar{\boldsymbol{x}}_{t_T})\varepsilon(\bar{\boldsymbol{x}}_{t_T})^\top\boldsymbol{a}}{\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = 1. \tag{26}$$

Hence, for any $\varepsilon > 0$, as $T \to \infty$, we have that

$$\begin{aligned} s_T^{-2}\sum_{m=1}^T \mathbb{E}[X_m^2 I(|X_m| \ge \varepsilon s_T)] &\le \varepsilon^{-\delta_2}s_T^{-(2+\delta_2)}\sum_{m=1}^T \mathbb{E}[|X_m|^{2+\delta_2}I(|X_m|\ge\varepsilon s_T)] \\ &\le \varepsilon^{-\frac{\delta_2}{2}}s_T^{-(2+\delta_2)}\sum_{m=1}^T\mathbb{E}|X_m|^{2+\delta_2} \\ &\precsim \varepsilon^{-\delta_2}s_T^{-(2+\delta_2)}\sum_{m=1}^T\frac{1}{E_m} \\ &\asymp \varepsilon^{-\delta_2}s_T^{-\delta_2} \to 0.\end{aligned}$$

The second condition: We have established the divergence of $\{\boldsymbol{s}_T^2\}_T$ in (26). Notice that

$$\begin{aligned} U_T^2 = \sum_{m=1}^T \mathbb{E}[X_m^2|\mathcal{G}_{m-1}] &= \sum_{m=1}^T\frac{1}{E_m}\boldsymbol{a}^\top\mathbb{E}[\varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top|\mathcal{G}_{m-1}]\boldsymbol{a} \\ &= \sum_{m=1}^T\frac{1}{E_m}\boldsymbol{a}^\top\mathbb{E}_{\xi_{t_m}}\varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top\boldsymbol{a}.\end{aligned}$$

Therefore, from (26) and the Stolz–Cesàro theorem (Lemma B.15), it follows almost surely that

$$\begin{aligned}\lim_{T\to\infty}\left|\frac{U_T^2}{s_T^2} - 1\right| &\le \lim_{T\to\infty}\frac{C}{s_T^2}\sum_{m=1}^T\frac{1}{E_m}\left[\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\| + \|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right] \\ &= \lim_{T\to\infty}\frac{C}{\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}}\left[\|\bar{\boldsymbol{x}}_{t_T} - \boldsymbol{x}^*\| + \|\bar{\boldsymbol{x}}_{t_T} - \boldsymbol{x}^*\|^2\right] \to 0.\end{aligned}$$

Lemma B.15 (Stolz–Cesàro theorem) *Let $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ be two sequences of real numbers. Assume that $\{b_n\}_{n\geq 1}$ is a strictly monotone and divergent sequence. We have that*

$$\textit{if } \lim_{n\to\infty} \frac{a_{n+1}-a_n}{b_{n+1}-b_n} = l, \textit{ then } \lim_{n\to\infty} \frac{a_n}{b_n} = l.$$

We have shown that the two conditions in Lemma B.13 hold. Hence, by definition, $\zeta_T(r) \Rightarrow B(r)$ where

$$\zeta_T(r) := U_T^{-1}\left[S_i + (U_{i+1}^2 - U_i^2)^{-1}(rU_T^2 - U_i^2)X_{i+1}\right] \quad \text{if} \quad U_i^2 \leq rU_T^2 < U_{i+1}^2$$

and $S_i = \sum_{m=1}^{i} X_m$. Since $s_T/U_T \to 1$ almost surely and (26), it follows that

$$\frac{\sqrt{t_T}}{T} U_T \zeta_T(r) \Rightarrow \sqrt{\nu}\sqrt{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}} B(r) \overset{d.}{=} \sqrt{\nu}\boldsymbol{a}^\top \boldsymbol{S}^{1/2}\mathbf{B}_d(r),$$

where $\mathbf{B}_d(r)$ is the d-dimensional standard Brownian motion. Recall that

$$h(r,T) = \max\left\{ n \in \mathbb{Z}_+ \middle| r\sum_{m=1}^{T}\frac{1}{E_m} \geq \sum_{m=1}^{n}\frac{1}{E_m}\right\}.$$

Lemma B.16 *Under the same condition of Lemma B.3, it follows that*

$$\sup_{r\in[0,1]}\left| \frac{\sqrt{t_T}}{T} U_T \zeta_T\left(\frac{U_{h(r,T)}^2}{U_T^2}\right) - \frac{\sqrt{t_T}}{T} U_T \zeta_T(r) \right| \to 0 \quad \textit{in probability.}$$

Hence,

$$\frac{\sqrt{t_T}}{T}\sum_{m=1}^{h(r,T)} \boldsymbol{a}^\top \boldsymbol{\varepsilon}_m = \frac{\sqrt{t_T}}{T} S_{h(r,T)} = \frac{\sqrt{t_T}}{T} U_T \zeta_T\left(\frac{U_{h(r,T)}^2}{U_T^2}\right) \Rightarrow \sqrt{\nu}\boldsymbol{a}^\top \boldsymbol{S}^{1/2}\mathbf{B}_d(r).$$

By the arbitrariness of $\boldsymbol{a}$, it follows that[4]

$$\frac{\sqrt{t_T}}{T}\sum_{m=1}^{h(r,T)} \boldsymbol{\varepsilon}_m \Rightarrow \sqrt{\nu}\boldsymbol{S}^{1/2}\mathbf{B}_d(r).$$

Applying the continuous mapping theorem to the linear function $\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \mapsto \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}$, we have

$$\frac{\sqrt{t_T}}{T}\sum_{m=1}^{h(r,T)} \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m \Rightarrow \sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathbf{B}_d(r).$$

Finally, since $\mathbb{E}\frac{\sqrt{t_T}}{T}\|\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_0\| \to 0$, it implies that $\frac{\sqrt{t_T}}{T}\boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_0 = o_{\mathbb{P}}(1)$. Then it is clear that $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m \Rightarrow \sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathbf{B}_d(r)$.

4. See the proof of Theorem 4.3.5. in Whitt (2002) for more detail about how to argue multivariate weak convergence from univariate weak convergence along any direction.

B.6. Proof of Lemma B.16

From the Theorem A.2 of Hall and Heyde (2014), if some random function $\phi_n \Rightarrow \phi$ in the sense of (C, ρ), $\{\phi_n\}$ must be tight in the sense that for any $\varepsilon > 0$, $\mathbb{P}(\sup_{|s-t|\leq\delta} |\phi_n(s) - \phi_n(t)| \geq \varepsilon) \to 0$ uniformly in n as $\delta \to 0$. Since $\frac{\sqrt{t_T}}{T} U_T \zeta_T(r) \Rightarrow \sqrt{\nu} \boldsymbol{a}^\top \boldsymbol{S}^{1/2} \mathbf{B}_d(r)$, $\{\frac{\sqrt{t_T}}{T} U_T \zeta_T\}_T$ is tight. We denote the following notation for simplicity

$$\phi_T(r) = \frac{\sqrt{t_T}}{T} U_T \zeta_T(r) \quad \text{and} \quad p_T(r) = \frac{U^2_{h(r,T)}}{U^2_T}.$$

Since $p_T(r)$ satisfies $p_T(0) = 1 - p_T(1) = 0$ and $p_T(r)$ is non-decreasing and right-continuous in r, we can view $p_T(r)$ as the cumulative distribution function of some random variable on $[0, 1]$ and $p(r) : r \mapsto r$ is the cumulative distribution function of uniform distribution on $[0, 1]$. It is clearly that $p_T(r) \to p(r)$ for every $r \in [0, 1]$ almost surely, because

$$\lim_{T\to\infty} p_T(r) = \lim_{T\to\infty} \frac{U^2_{h(r,T)}}{U^2_T} = \lim_{T\to\infty} \frac{s^2_{h(r,T)}}{s^2_T} = \lim_{T\to\infty} \frac{\sum_{m=1}^{h(r,T)} \frac{1}{E_m}}{\sum_{m=1}^{T} \frac{1}{E_m}} = r = p(r).$$

Here we use $h(r, T) \to \infty$ for any $r \in [0, 1]$ as $T \to \infty$. Since $p(\cdot)$ is additionally continuous, weak convergence implies uniform convergence in cumulative distribution functions, i.e.,

$$\lim_{T\to\infty} \sup_{r\in[0,1]} |p_T(r) - r| = 0. \tag{27}$$

By the tightness of $\{\phi_n\}$, for any $\varepsilon, \eta > 0$, we can find a sufficiently small δ such that

$$\limsup_{T\to\infty} \mathbb{P}\left(\sup_{|s-t|\leq\delta} |\phi_T(s) - \phi_T(t)| \geq \varepsilon \right) \leq \eta.$$

With (27), for this δ, $\mathbb{P}(\sup_{r\in[0,1]} |p_T(r) - r| > \delta) \to 0$ as $T \to \infty$. It implies that

$$\begin{aligned}
&\limsup_{T\to\infty} \mathbb{P}\left(\sup_{r\in[0,1]} |\phi_T(p_T(r)) - \phi_T(r)| \geq \varepsilon \right) \\
&\leq \limsup_{T\to\infty} \mathbb{P}\left(\sup_{r\in[0,1]} |\phi_T(p_T(r)) - \phi_T(r)| \geq \varepsilon, \sup_{r\in[0,1]} |p_T(r) - r| \leq \delta \right) \\
&\qquad + \lim_{T\to\infty} \mathbb{P}\left(\sup_{r\in[0,1]} |p_T(r) - r| > \delta \right) \\
&\leq \limsup_{T\to\infty} \mathbb{P}\left(\sup_{|s-t|\leq\delta} |\phi_T(s) - \phi_T(t)| \geq \varepsilon \right) \leq \eta.
\end{aligned}$$

Because η is arbitrary, we have shown that

$$\sup_{r\in[0,1]} |\phi_T(p_T(r)) - \phi_T(r)| \to 0 \quad \text{in probability.}$$

B.7. Proof of Lemma B.4

Recall that $\boldsymbol{G} = \nabla^2 f(\boldsymbol{x}^*)$, $\boldsymbol{s}_m = \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*$ and

$$\boldsymbol{r}_m = \nabla f(\bar{\boldsymbol{x}}_{t_m}) - \boldsymbol{G}\boldsymbol{s}_m.$$

When $\|\boldsymbol{s}_m\| \leq \delta_1$, by Assumption 3.1, $\|\nabla^2 f(s\boldsymbol{s}_m + \boldsymbol{x}^*) - \nabla^2 f(\boldsymbol{x}^*)\| \leq sL'\|\boldsymbol{s}_m\|$, then

$$\begin{aligned}
\|\boldsymbol{r}_m\| &= \|\nabla f(\boldsymbol{s}_m + \boldsymbol{x}^*) - \nabla f(\boldsymbol{x}^*) - \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{s}_m\| \\
&= \left\| \int_0^1 \nabla^2 f(s\boldsymbol{s}_m + \boldsymbol{x}^*)\boldsymbol{s}_m ds - \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{s}_m \right\| \\
&\leq \int_0^1 \left\|\nabla^2 f(s\boldsymbol{s}_m + \boldsymbol{x}^*) - \nabla^2 f(\boldsymbol{x}^*)\right\| \|\boldsymbol{s}_m\| ds \\
&\leq \frac{L'}{2}\|\boldsymbol{s}_m\|^2.
\end{aligned}$$

When $\|\boldsymbol{s}_m\| > \delta_1$, $\|\boldsymbol{r}_m\| \leq \|\nabla f(\bar{\boldsymbol{x}}_{t_m})\| + \|\boldsymbol{G}\boldsymbol{s}_m\| \leq L\|\boldsymbol{s}_m\| + L\|\boldsymbol{s}_m\| = 2L\|\boldsymbol{s}_m\|$. Applying the results above yields

$$\|\boldsymbol{r}_m\| \leq L'\|\boldsymbol{s}_m\|^2 1_{\{\|\boldsymbol{s}_m\| \leq \delta_1\}} + 2L\|\boldsymbol{s}_m\| 1_{\{\|\boldsymbol{s}_m\| > \delta_1\}}.$$

Hence,

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \|\boldsymbol{r}_m\| \leq \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \left[L'\|\boldsymbol{s}_m\|^2 1_{\{\|\boldsymbol{s}_m\| \leq \delta_1\}} + 2L\|\boldsymbol{s}_m\| 1_{\{\|\boldsymbol{s}_m\| > \delta_1\}}\right].$$

By Lemma B.2, $\boldsymbol{s}_m \to 0$ almost surely, which implies

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \|\boldsymbol{s}_m\| 1_{\{\|\boldsymbol{s}_m\| > \delta_1\}} \to 0 \quad \text{almost surely.}$$

It then suffices to show that $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \|\boldsymbol{s}_m\|^2 1_{\{\|\boldsymbol{s}_m\| \leq \delta_1\}} = o_{\mathbb{P}}(1)$, which is implied by

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \mathbb{E}\|\boldsymbol{s}_m\|^2 = o(1).$$

It holds because $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \mathbb{E}\|\boldsymbol{s}_m\|^2 \precsim \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} \gamma_m \to 0$ from Lemma B.2 and Assumption 3.4.

B.8. Proof of Lemma B.5

In the proof of Lemma B.2 (see the Part 2 therein), we have established for sufficiently large m,

$$\mathbb{E}[\|\boldsymbol{\delta}_m\|^2 | \mathcal{F}_{t_m}] \leq \frac{L^2}{E_m}\sum_{t=t_m}^{t_{m+1}-1} V_t \leq L^2\gamma_m^2 \frac{E_m - 1}{E_m}\left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right),$$

where V_t is the residual error defined in (19) and $C_3, C_4 > 0$ are universal constants defined in Lemma B.10. Besides, Lemma B.2 implies that $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \precsim \gamma_m \precsim 1$. It follows that

$$\mathbb{E}\|\boldsymbol{\delta}_m\|^2 \leq L^2\gamma_m^2\left(C_3 + C_4\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2\right) \precsim \gamma_m^2.$$

In order to prove the conclusion, it suffices to show that $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\mathbb{E}\left\|\boldsymbol{\delta}_m\right\| \to 0$, which is satisfied because

$$\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\mathbb{E}\left\|\boldsymbol{\delta}_m\right\| \leq \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\sqrt{\mathbb{E}\left\|\boldsymbol{\delta}_m\right\|^2} \precsim \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\gamma_m \to 0$$

from Lemma B.2 and Assumption 3.4.

B.9. Proof of Lemma B.6

If $\{E_m\}$ is uniformly bounded (i.e., there exists some C such that $1 \leq E_m \leq C$ for all m), the conclusion follows because

$$0 \leq \frac{(\sum_{m=0}^{T-1} E_m)(\sum_{m=0}^{T-1} E_m^{-1} a_{m,T})}{T^2} \leq \frac{CT(\sum_{m=0}^{T-1} a_{m,T})}{T^2} = \frac{1}{T}\sum_{m=0}^{T-1} a_{m,T} \to 0 \quad \text{when} \quad T \to \infty.$$

In the following, we instead assume E_m is non-decreasing in m (i.e., $1 \leq E_m \leq E_{m+1}$ for all m). Let $H_k = \sum_{m=0}^{k} a_{m,T}$. For any ε, there exist some $N = N(\varepsilon)$, such that for any $m \geq N$, $0 \leq H_m \leq m\varepsilon$. Then

$$\begin{aligned}
\sum_{n=N}^{T}\frac{a_{m,T}}{E_m} &= \sum_{n=N}^{T}\frac{H_m - H_{m-1}}{E_m} = \frac{H_T}{E_T} + \sum_{n=N}^{T-1}\left(\frac{1}{E_m} - \frac{1}{E_{m+1}}\right)H_m - \frac{H_{N-1}}{E_N} \\
&\leq \frac{H_T}{E_T} + \sum_{n=N}^{T-1}\left(\frac{1}{E_m} - \frac{1}{E_{m+1}}\right)m\varepsilon - \frac{H_{N-1}}{E_N} \\
&= \frac{H_T - T\varepsilon}{E_T} + \left[\frac{T\varepsilon}{E_T} + \sum_{n=N}^{T-1}\left(\frac{1}{E_m} - \frac{1}{E_{m+1}}\right)m\varepsilon - \frac{(N-1)\varepsilon}{E_N}\right] - \frac{H_{N-1} - (N-1)\varepsilon}{E_N} \\
&= \varepsilon \cdot \sum_{n=N}^{T}\frac{1}{E_m} + \frac{H_T - T\varepsilon}{E_T} - \frac{H_{N-1} - (N-1)\varepsilon}{E_N} \\
&\leq \varepsilon \cdot \sum_{n=N}^{T}\frac{1}{E_m} + \frac{N\varepsilon}{E_N}
\end{aligned}$$

Recall $t_T = \sum_{m=0}^{T-1} E_m$. Therefore,

$$\begin{aligned}
\frac{t_T(\sum_{m=0}^{T-1} E_m^{-1} a_{m,T})}{T^2} =& \frac{t_T(\sum_{m=0}^{N-1} E_m^{-1} a_{m,T})}{T^2} + \frac{t_T(\sum_{m=N}^{T-1} E_m^{-1} a_{m,T})}{T^2} \\
\leq& \frac{t_T(\sum_{m=0}^{N-1} E_m^{-1} a_{m,T})}{T^2} + \varepsilon\frac{t_T(\sum_{m=N}^{T} E_m^{-1})}{T^2} + \frac{t_T N\varepsilon}{T^2 E_N}.
\end{aligned}$$

Taking superior limit on both sides and noting $a_{m,T} \precsim 1$ uniformly and $\lim\limits_{T\to\infty}\frac{t_T}{T^2} = 0$, we have

$$0 \leq \limsup_{T\to\infty}\frac{t_T(\sum_{m=0}^{T-1} E_m^{-1} a_{m,T})}{T^2} \leq \varepsilon\nu.$$

By the arbitrariness of ε, we complete the proof.

B.10. Proof of Lemma B.7

Without loss of generality, we assume $\boldsymbol{G}^{-1}$ is a positive diagonal matrix. Otherwise, we apply the spectrum decomposition to $\boldsymbol{G} = \boldsymbol{V}\mathbf{D}\boldsymbol{V}^\top$ and focus on the coordinates of each $\boldsymbol{\varepsilon}_m$ with respect to the orthogonal base $\boldsymbol{V}$. This simplification reduces our multivariate case to a univariate one. Hence, it is enough to show that the result holds for one-dimensional $\boldsymbol{\varepsilon}_m$ and $\boldsymbol{G}$. In the following argument, we focus on an eigenvalue λ of $\boldsymbol{G}$ and its eigenvector $\boldsymbol{v}$, and denote $\varepsilon_m = \boldsymbol{v}^\top \boldsymbol{\varepsilon}_m$ and $B_m = 1 - \gamma_m \lambda \in \mathbb{R}$ for simplicity. Clearly, $\lambda \geq 0$ and $0 < B_m \leq 1$ for sufficiently large m.

Given a positive integer n, we separate the time interval $[0, T]$ uniformly into n portions with $h_i = \left[\frac{iT}{n}\right]$ $(i = 0, 1, \ldots, n)$ the i-th endpoint. The choice of n is independent of T, which implies that $\lim_{T\to\infty} h_i = \infty$ for any i. Define an event $\mathcal{A}$ whose complement is

$$\mathcal{A}^c = \left\{ \exists h_i \text{ s.t. } \left\| \frac{\sqrt{t_T}}{T\gamma_{h_i+1}} \sum_{m=0}^{h_i} \left(\prod_{i=m+1}^{h_i} B_i \right) \gamma_m \varepsilon_m \right\| \geq \varepsilon \right\}.$$

We claim that $\limsup_{T\to\infty} \mathbb{P}(\mathcal{A}^c) = 0$. Indeed, by the union bound and Markov's inequality,

$$\begin{aligned}
\mathbb{P}(\mathcal{A}^c) &\leq \sum_{i=0}^{n} \mathbb{P} \left\{ \left\| \frac{\sqrt{t_T}}{T\gamma_{h_i+1}} \sum_{m=0}^{h_i} \prod_{j=m+1}^{h_i} B_j \gamma_m \varepsilon_m \right\| \geq \varepsilon \right\} \\
&\precsim \sum_{i=0}^{n} \frac{t_T}{\varepsilon^2 T^2 \gamma_{h_i+1}^2} \sum_{m=0}^{h_i} \left(\prod_{j=m+1}^{h_i} B_j \right)^2 \gamma_m^2 \\
&\precsim \frac{t_T}{\varepsilon^2 T^2} \sum_{i=0}^{n} \frac{1}{\gamma_{h_i+1}} \\
&\leq \frac{t_T(n+1)}{\varepsilon^2 T^2 \gamma_{T+1}} \to 0 \quad \text{as} \quad T \to \infty.
\end{aligned}$$

Here the last two inequality uses for any $i \in [n]$,

$$\frac{1}{\gamma_{h_i+1}} \sum_{m=0}^{h_i} \left(\prod_{j=m+1}^{h_i} B_j \right)^2 \gamma_m^2 \precsim 1,$$

which is implied by

$$\begin{aligned}
&\lim_{h_i\to\infty} \left\{ \sum_{m=0}^{h_i} \gamma_m^2 \left(\prod_{j=0}^{m} B_j \right)^{-2} \right\} \Big/ \left\{ \gamma_{h_i} \left(\prod_{j=0}^{h_i} B_j \right)^{-2} \right\} \\
&= \lim_{h_i\to\infty} \left\{ \gamma_{h_i}^2 \left(\prod_{j=0}^{h_i} B_j^{-2} \right) \right\} \Big/ \left\{ o(\gamma_{h_i-1})\gamma_{h_i-1} \prod_{j=0}^{h_i} B_j^{-2} + \gamma_{h_i} \prod_{j=0}^{h_i} B_j^{-2}(1 - B_{h_i}^2) \right\} \\
&= \lim_{h_i\to\infty} \frac{\gamma_{h_i}^2}{o(1)\gamma_{h_i-1}^2 + 2\lambda\gamma_{h_i}^2 - \lambda^2\gamma_{h_i}^3}
\end{aligned}$$

$$= \frac{1}{2\lambda} < \infty$$

as a result of the Stolz–Cesàro theorem (Lemma B.15). Here we observe that the denominator $\gamma_{h_i}\left(\prod_{j=0}^{h_i} B_j\right)^{-2}$ increases in h_i and diverges when h_i is sufficiently large.

Since the event $\mathcal{A}^c$ has diminishing probability, we focus on the event $\mathcal{A}$. We will prove that on the event $\mathcal{A}$ our target random sequence is uniformly tight. For notation simplicity, we define

$$X_m^h = \prod_{i=m}^{h} B_i.$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{0\le h\le T}\left|\frac{1}{\gamma_{h+1}}\sum_{m=0}^{h}\left(\prod_{i=m+1}^{h} B_i\right)\gamma_m\varepsilon_m\right| \ge 2\varepsilon\ ;\ \mathcal{A}\right\}\\
&= \mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{0\le h\le T}\left|\frac{1}{\gamma_{h+1}X_{h+1}^T}\sum_{m=0}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right| \ge 2\varepsilon\ ;\ \mathcal{A}\right\}\\
&\le \sum_{i=0}^{n-1}\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{h\in[h_i,h_{i+1})}\left|\frac{1}{\gamma_{h+1}X_{h+1}^T}\left(\sum_{m=0}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right)\right| \ge 2\varepsilon\ ;\ \mathcal{A}\right\}\\
&\le \sum_{i=0}^{n-1}\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{h\in[h_i,h_{i+1})}\frac{1}{\gamma_{h+1}X_{h+1}^T}\left|\sum_{m=0}^{h_i} X_{m+1}^T\gamma_m\varepsilon_m + \sum_{m=h_i+1}^{h} X_m^T\gamma_m\varepsilon_m\right| \ge 2\varepsilon\ ;\ \mathcal{A}\right\}\\
&\le \sum_{i=0}^{n-1}\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{h\in[h_i,h_{i+1})}\left[\frac{1}{\gamma_{h+1}X_{h+1}^T}\left|\sum_{m=0}^{h_i} X_{m+1}^T\gamma_m\varepsilon_m\right| + \left|\frac{1}{\gamma_h X_{h+1}^T}\sum_{m=h_i+1}^{h} X_m^T\gamma_m\varepsilon_m\right|\right] \ge 2\varepsilon\ ;\ \mathcal{A}\right\}\\
&\le \sum_{i=0}^{n-1}\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{h\in[h_i,h_{i+1})}\left|\frac{1}{\gamma_{h+1}X_{h+1}^T}\sum_{m=h_i+1}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right| \ge \varepsilon\ ;\ \mathcal{A}\right\}\\
&\le \sum_{i=0}^{n-1}\mathbb{P}\left\{\frac{\sqrt{t_T}}{T}\sup_{h\in[h_i,h_{i+1})}\left|\frac{1}{\gamma_{h+1}X_{h+1}^T}\sum_{m=h_i+1}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right| \ge \varepsilon\right\}\\
&= \sum_{i=0}^{n-1}\mathbb{P}\left\{\left(\frac{\sqrt{t_T}}{T}\right)^{2+\delta}\sup_{h\in[h_i,h_{i+1})}\left(\frac{1}{\gamma_{h+1}X_{h+1}^T}\right)^{2+\delta}\left|\sum_{m=h_i+1}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right|^{2+\delta} \ge \varepsilon^{2+\delta}\right\}\\
&:= \sum_{i=0}^{n-1}\mathcal{P}_i,
\end{aligned}$$

where δ is any positive real number less than $\min\{\delta_2,\delta_3\}$.

Let $Y_h = \left|\sum_{m=h_i+1}^{h} X_{m+1}^T\gamma_m\varepsilon_m\right|^{2+\delta}$. It is clear that Y_h is a sub-martingale adapted to the natural filtration. Let $c_h = \frac{1}{(\gamma_h X_h^T)^{2+\delta}}$. Then $\{c_h\}$ is a non-increasing sequence when h is sufficiently large because

$$\gamma_h X_h^T = \frac{\gamma_h}{\gamma_{h+1}}(1-\lambda\gamma_h)\gamma_{h+1}X_{h+1}^T = (1+o(\gamma_h))(1-\lambda\gamma_h)\gamma_{h+1}X_{h+1}^T \le \gamma_{h+1}X_{h+1}^T$$

for sufficiently large h. Indeed, since $h \geq h_i = \left[\frac{iT}{n}\right] \to \infty$ as $T \to \infty$, $(1+o(\gamma_h))(1-\lambda\gamma_h) \leq 1$ is solid and X_h^T is non-negative when T goes to infinity. Hence, each $\mathcal{P}_i$ is the probability of the event where the maximum of a sub-martingale multiplied by a non-increasing sequence is larger than a threshold. To bound each $\mathcal{P}_i$, we use Chow's inequality which is a generalization of Doob's inequality (Chow, 1960). It follows that

$$\begin{aligned}\mathcal{P}_i &= \mathbb{P}\left\{\frac{t_T^{1+\delta/2}}{T^{2+\delta}} \sup_{h\in[h_i,h_{i+1})} c_h Y_h \geq \varepsilon^{2+\delta}\right\} \\ &\leq \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left\{c_{h_{i+1}-1}\mathbb{E}Y_{h_{i+1}-1} + \sum_{j=h_i+1}^{h_{i+1}-2}(c_i - c_{i+1})\mathbb{E}Y_j\right\}. \end{aligned} \tag{28}$$

We then apply Burkholder's inequality to bound each $\mathbb{E}Y_j$. Burkholder's inequality is a generalization of the Marcinkiewicz–Zygmund inequality (Lemma B.14) to martingale differences (Dharmadhikari et al., 1968). That is,

$$\begin{aligned}\mathbb{E}Y_j &= \mathbb{E}\left|\sum_{m=h_i+1}^{j} X_{m+1}^T\gamma_m\varepsilon_m\right|^{2+\delta} \\ &\lesssim (j-h_i)^{\delta/2}\sum_{m=h_i+1}^{j}\mathbb{E}\left|X_{m+1}^T\gamma_m\varepsilon_m\right|^{2+\delta} \\ &\lesssim (j-h_i)^{\delta/2}\sum_{m=h_i+1}^{j}(X_{m+1}^T\gamma_m)^{2+\delta}/E_m^{1+\delta/2} \\ &\lesssim (j-h_i)^{\delta/2}\sum_{m=h_i+1}^{j} c_m^{-1}/E_m^{1+\delta/2},\end{aligned}$$

where we use $\mathbb{E}\left|\varepsilon_m\right|^{2+\delta} \lesssim 1/E_m^{1+\delta/2}$ for sufficiently large m that is already derived in (25).

Plugging it into (28) yields that $\mathcal{P}_i$ is bounded by

$$\begin{aligned}&\frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left\{c_{h_{i+1}-1}\mathbb{E}Y_{h_{i+1}-1} + \sum_{j=h_i+1}^{h_{i+1}-2}(c_i - c_{i+1})\mathbb{E}Y_j\right\} \\ &\lesssim \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left\{c_{h_{i+1}-1}(h_{i+1}-h_i)^{\frac{\delta}{2}}\sum_{m=h_i+1}^{h_{i+1}-1}\frac{c_m^{-1}}{E_m^{1+\delta/2}} + \sum_{j=h_i+1}^{h_{i+1}-2}(c_j-c_{j+1})(j-h_i)^{\frac{\delta}{2}}\sum_{m=h_i+1}^{j}\frac{c_m^{-1}}{E_m^{1+\delta/2}}\right\} \\ &\leq \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left(\frac{T}{n}\right)^{\delta/2}\left\{c_{h_{i+1}-1}\sum_{m=h_i+1}^{h_{i+1}-1}\frac{c_m^{-1}}{E_m^{1+\delta/2}} + \sum_{j=h_i+1}^{h_{i+1}-2}(c_j - c_{j+1})\sum_{m=h_i+1}^{j}\frac{c_m^{-1}}{E_m^{1+\delta/2}}\right\} \\ &= \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left(\frac{T}{n}\right)^{\delta/2}\left\{c_{h_{i+1}-1}\sum_{m=h_i+1}^{h_{i+1}-1}\frac{c_m^{-1}}{E_m^{1+\delta/2}} + \sum_{m=h_i+1}^{h_{i+1}-2}(c_m - c_{h_{i+1}-1})\frac{c_m^{-1}}{E_m^{1+\delta/2}}\right\} \\ &= \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left(\frac{T}{n}\right)^{\delta/2}\left\{\sum_{m=h_i+1}^{h_{i+1}-1} c_m\frac{c_m^{-1}}{E_m^{1+\delta/2}}\right\}\end{aligned}$$

$$= \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta} T^{2+\delta}} \left(\frac{T}{n}\right)^{\delta/2} \sum_{m=h_i+1}^{h_{i+1}-1} \frac{1}{E_m^{1+\delta/2}}.$$

Recall $t_T = \sum_{m=0}^{T-1} E_m$. Summing the last bound over $i = 0, 1, \ldots, n-1$ gives

$$\begin{aligned}
\sum_{i=0}^{n-1} \mathcal{P}_i &\precsim \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta} T^{2+\delta}} \left(\frac{T}{n}\right)^{\delta/2} \sum_{m=0}^{T-1} \frac{1}{E_m^{1+\delta/2}} \\
&= \frac{1}{\varepsilon^{2+\delta} n^{\delta/2}} \frac{(\frac{1}{T}\sum_{m=0}^{T-1} E_m)^{1+\delta/2}}{\frac{1}{T}\sum_{m=0}^{T-1} E_m^{1+\delta/2}} \frac{\sum_{m=0}^{T-1} E_m^{1+\delta/2} \sum_{m=0}^{T-1} 1/E_m^{1+\delta/2}}{T^2} \\
&\precsim \frac{1}{n^{\delta/2}},
\end{aligned}$$

where we use (ii) in Assumption 3.4 which implies

$$\sup_T \frac{\sum_{m=0}^{T-1} E_m^{1+\delta/2} \sum_{m=0}^{T-1} 1/E_m^{1+\delta/2}}{T^2} \le \sup_T \frac{\sum_{m=0}^{T-1} E_m^{1+\delta_3} \sum_{m=0}^{T-1} 1/E_m^{1+\delta_3}}{T^2} < \infty$$

as a result of $\delta < \delta_3$.

Summing them all, we have

$$\begin{aligned}
&\limsup_{T\to\infty} \mathbb{P}\left\{ \frac{\sqrt{t_T}}{T} \sup_{0\le h\le T} \left| \frac{1}{\gamma_{h+1}} \sum_{m=0}^{h} \left(\prod_{i=m+1}^{h} B_i \right) \gamma_m \varepsilon_m \right| \ge 2\varepsilon \right\} \\
&\le \limsup_{T\to\infty} \mathbb{P}\left\{ \frac{\sqrt{t_T}}{T} \sup_{0\le h\le T} \left| \frac{1}{\gamma_{h+1}} \sum_{m=0}^{h} \left(\prod_{i=m+1}^{h} B_i \right) \gamma_m \varepsilon_m \right| \ge 2\varepsilon \;;\, \mathcal{A} \right\} + \limsup_{T\to\infty} \mathbb{P}(\mathcal{A}^c) \\
&\le \limsup_{T\to\infty} \sum_{i=0}^{n-1} \mathcal{P}_i \\
&\precsim \frac{1}{n^{\delta/2}}.
\end{aligned}$$

Since the probability of the left hand side has nothing to do with n, letting $n \to \infty$ concludes the proof.

Appendix C. Proofs of Proposition 3.1

To prove the proposition, we make two following claims.

Claim 1: For any positive sequences $\{a_n\}$ and $\{b_n\}$ with $\sum_{n=1}^{T} b_n \to \infty$, we have

$$\limsup_{T\to\infty} \frac{\sum_{n=1}^{T} a_n}{\sum_{n=1}^{T} b_n} \le \limsup_{T\to\infty} \frac{a_T}{b_T}. \tag{29}$$

Without loss of generality, we assume the right hand side is finite, otherwise (29) follows obviously. We denote that $\limsup_{T\to\infty} \frac{a_T}{b_T} = \lambda$ for simplicity. Based on the definition of limit superior, for any $\varepsilon > 0$, there exists N_ε subject to $a_n < (\lambda+\varepsilon)b_n$ for $\forall n \geq N_\varepsilon$. As a result,

$$\sum_{n=1}^{T} a_n = \sum_{n=1}^{N_\varepsilon} a_n + \sum_{n=N_\varepsilon+1}^{T} a_n \leq \sum_{n=1}^{N_\varepsilon} a_n + (\lambda+\varepsilon) \sum_{n=N_\varepsilon+1}^{T} b_n,$$

which implies

$$\frac{\sum_{n=1}^T a_n}{\sum_{n=1}^T b_n} \leq \frac{\sum_{n=1}^{N_\varepsilon} a_n + (\lambda+\varepsilon)\sum_{n=N_\varepsilon+1}^T b_n}{\sum_{n=1}^T b_n}.$$

Taking limit superior on both sides and noting that $\sum_{n=1}^{T} b_n \to \infty$, we have $\frac{\sum_{n=1}^T a_n}{\sum_{n=1}^T b_n} \leq \lambda + 2\varepsilon$. By the arbitrariness of ε, (29) follows.

Claim 2: For any non-decreasing sequence $\{E_m\}$ satisfying $\limsup_{T\to\infty} T(1-\frac{E_{T-1}}{E_T}) < 1$, we can find $\delta > 0$ such that

$$T\left(\frac{1}{E_T}\right)^{1+\delta} - (T-1)\left(\frac{1}{E_{T-1}}\right)^{1+\delta} > 0.$$

In fact, we can choose any $\delta < 1 - \limsup_{T\to\infty} T(1-\frac{E_{T-1}}{E_T})$. In this way, for sufficiently large T, we have

$$\begin{aligned} T\left(\frac{1}{E_T}\right)^{1+\delta} - (T-1)\left(\frac{1}{E_{T-1}}\right)^{1+\delta} &= \left(\frac{1}{E_{T-1}}\right)^{1+\delta}\left(T\left(\frac{E_{T-1}}{E_T}\right)^{1+\delta} - T + 1\right) \\ &\geq T\left(\frac{1}{E_{T-1}}\right)^{1+\delta}\left[\left(1-\frac{1-\delta}{T}\right)^{1+\delta} - 1 + \frac{1}{T}\right]. \end{aligned}$$

To lower bound the right hand side, we consider the auxiliary function $h(x) = (1-(1-\delta)x)^{1+\delta} + x$ where $x \in (0,1)$. We claim that $h(x) > 1$ for any $x \in (0,1)$. We check it by investigating the derivative of $h(\cdot)$,

$$\dot{h}(x) = -(1+\delta)(1-(1-\delta)x)^{1+\delta}(1-\delta) + 1 > -(1+\delta)(1-\delta) + 1 = \delta^2 > 0.$$

Therefore, by mean value theorem, $h(x) > h(0) = 1$ which proves the claim.

Now we are well prepared to prove the proposition. It follows that

$$\begin{aligned} \limsup_{T\to\infty} T\left[1-\left(\frac{E_{T-1}}{E_T}\right)^{1+\delta}\right] &= \limsup_{T\to\infty} T\frac{(1+\delta)(\theta_T E_T + (1-\theta_T)E_{T-1})^{\delta}(E_T - E_{T-1})}{E_T^{1+\delta}} \\ &\leq (1+\delta)\limsup_{T\to\infty}\left(\frac{\theta_T E_T + (1-\theta_T)E_{T-1}}{E_T}\right)^{\delta} \limsup_{T\to\infty} T\frac{E_T - E_{T-1}}{E_T} \\ &\leq (1+\delta)(1-\delta)\limsup_{T\to\infty}\left(\frac{\theta_T E_T + (1-\theta_T)E_{T-1}}{E_T}\right)^{\delta} \\ &\leq 1-\delta^2, \end{aligned}$$

where the first equality uses mean value theorem with some $\theta_T \in [0,1]$.

Therefore,

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{(\sum_{m=1}^T E_m^{1+\delta})(\sum_{m=1}^T (1/E_m)^{1+\delta})}{T^2} \\
\overset{(a)}{\le}\; & \limsup_{T\to\infty} \frac{E_T^{1+\delta}\sum_{m=1}^T (1/E_m)^{1+\delta} + (\sum_{m=1}^T E_m^{1+\delta})/(E_T)^{1+\delta}}{2T-1} \\
\le\; & \limsup_{T\to\infty} \frac{\sum_{m=1}^T (1/E_m)^{1+\delta}}{(2T-1)/E_T^{1+\delta}} + \frac{1}{2} \\
<\; & \limsup_{T\to\infty} \frac{\sum_{m=1}^T (1/E_m)^{1+\delta}}{T(1/E_T)^{1+\delta}} \\
\overset{(b)}{\le}\; & \limsup_{T\to\infty} \frac{(1/E_T)^{1+\delta}}{T(1/E_T)^{1+\delta} - (T-1)(1/E_{T-1})^{1+\delta}} \\
\le\; & \limsup_{T\to\infty} \frac{1}{1 - T\left[1 - \left(\frac{E_{T-1}}{E_T}\right)^{1+\delta}\right]} \\
\le\; & \left\{1 - \limsup_{T\to\infty} T\left[1 - \left(\frac{E_{T-1}}{E_T}\right)^{1+\delta}\right]\right\}^{-1} \le \delta^{-2} < \infty,
\end{aligned}$$

where (a) uses Claim 1 and (b) uses Claim 1 and Claim 2 together.

Furthermore, if the sequence $\{E_m\}$ satisfies $\lim_{T\to\infty} T\left(1 - \frac{E_{T-1}}{E_T}\right) = \rho < 1$, then by the Stolz–Cesàro theorem (Lemma B.15), we have

$$\begin{aligned}
&\lim_{T\to\infty} \frac{(\sum_{m=1}^T E_m)(\sum_{m=1}^T 1/E_m)}{T^2} \\
=\; & \lim_{T\to\infty} \frac{E_T(\sum_{n=1}^T 1/E_n) + (\sum_{n=1}^{T-1} E_n)/E_T}{2T-1} \\
=\; & \frac{1}{2}\left\{\lim_{T\to\infty} \frac{\sum_{n=1}^T 1/E_n}{T/E_T} + \lim_{T\to\infty} \frac{\sum_{n=1}^T E_n}{TE_T}\right\} \\
=\; & \frac{1}{2}\left\{\lim_{T\to\infty} \frac{1/E_T}{T/E_T - (T-1)/E_{T-1}} + \lim_{T\to\infty} \frac{E_T}{TE_T - (T-1)E_{T-1}}\right\} \\
=\; & \frac{1}{2}\left\{\lim_{T\to\infty} \frac{E_{T-1}}{E_T} \times \frac{1}{1 - T(1 - E_{T-1}/E_T)} + \lim_{T\to\infty} \frac{1}{1 + (T-1)(1 - E_{T-1}/E_T)}\right\} \\
=\; & \frac{1}{2}\left\{\frac{1}{1-\rho} + \frac{1}{1+\rho}\right\} = \frac{1}{1-\rho^2},
\end{aligned}$$

which completes the proof.

Appendix D. Proof for the Plug-in Method, Theorem 4.1

For simplicity, we denote $\nabla f(\boldsymbol{x};\xi_t) = \sum_{k=1}^K p_k \nabla f_k(\boldsymbol{x};\xi_t^k)$ and $\nabla^2 f(\boldsymbol{x};\xi_t) = \sum_{k=1}^K p_k \nabla^2 f_k(\boldsymbol{x};\xi_t^k)$ where $\xi_t = \{\xi_t^k\}_{k\in[K]}$. We decompose $\widehat{\boldsymbol{G}}_T - \boldsymbol{G}$ into the following terms:

$$\begin{aligned}\widehat{\boldsymbol{G}}_T - \boldsymbol{G} &= \frac{1}{T}\sum_{m=1}^T \nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \boldsymbol{G} \\ &= \left[\frac{1}{T}\sum_{m=1}^T \nabla^2 f(\boldsymbol{x}^*;\xi_{t_m}) - \boldsymbol{G}\right] + \frac{1}{T}\sum_{m=1}^T \left[\nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla^2 f(\boldsymbol{x}^*;\xi_{t_m})\right]. \qquad (30)\end{aligned}$$

The first term in (30) is asymptotically zero due to the strong law of large number. With Theorem 3.1, we have known that under the condition, $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\| \le \sqrt{\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2} \lesssim \sqrt{\gamma_m}$. Then the second term in (30) can be bounded via Assumption 4.1

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{T}\sum_{m=1}^T \left[\nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla^2 f(\boldsymbol{x}^*;\xi_{t_m})\right]\right\| &\le \frac{1}{T}\sum_{m=1}^T \mathbb{E}\left\|\nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla^2 f(\boldsymbol{x}^*;\xi_{t_m})\right\| \\ &\le \frac{L''}{T}\sum_{m=1}^T \mathbb{E}\left\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\right\| \\ &\lesssim \frac{1}{T}\sum_{m=1}^T \sqrt{\gamma_m} \to 0\end{aligned}$$

as $T\to\infty$. Hence, $\widehat{\boldsymbol{G}}_T$ converges to $\boldsymbol{G}$ in probability.

For $\widehat{\boldsymbol{S}}_T$, note that

$$\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) = \nabla f(\boldsymbol{x}^*;\xi_{t_m}) + [\nabla f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla f(\boldsymbol{x}^*;\xi_{t_m})] := \boldsymbol{C}_m + \boldsymbol{D}_m.$$

We decompose $\widehat{\boldsymbol{S}}_T - \boldsymbol{S}$ into the following terms:

$$\widehat{\boldsymbol{S}}_T - \boldsymbol{S} = \left(\frac{1}{T}\sum_{m=1}^T \boldsymbol{C}_m\boldsymbol{C}_m^\top - \boldsymbol{S}\right) + \frac{1}{T}\sum_{m=1}^T \boldsymbol{C}_m\boldsymbol{D}_m^\top + \frac{1}{T}\sum_{m=1}^T \boldsymbol{D}_m\boldsymbol{C}_m^\top + \frac{1}{T}\sum_{m=1}^T \boldsymbol{D}_m\boldsymbol{D}_m^\top.$$

Because $\{\boldsymbol{C}_m\}_m$ are i.i.d. and $\mathbb{E}\boldsymbol{C}_m\boldsymbol{C}_m^\top = \boldsymbol{S}$, the first term is asymptotically zero due to the strong law of large number. Note that $\mathbb{E}\|\boldsymbol{C}_m\|^2 = \mathbb{E}\|\boldsymbol{C}_m\boldsymbol{C}_m^\top\| \le \mathrm{tr}(\mathbb{E}\boldsymbol{C}_m\boldsymbol{C}_m^\top) = \mathrm{tr}(\boldsymbol{S})$ and

$$\begin{aligned}\mathbb{E}\|\boldsymbol{D}_m\|^2 &= \mathbb{E}\left\|\sum_{k=1}^K p_k\left(\nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k) - \nabla f(\boldsymbol{x}^*;\xi_{t_m}^k)\right)\right\|^2 \\ &\le \sum_{k=1}^k p_k\mathbb{E}\left\|\nabla f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k) - \nabla f(\boldsymbol{x}^*;\xi_{t_m}^k)\right\|^2 \\ &\le L^2\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*\|^2 \lesssim \gamma_m.\end{aligned}$$

Then, the second and third terms can be bounded via

$$\mathbb{E}\left\|\frac{1}{T}\sum_{m=1}^T \boldsymbol{C}_m\boldsymbol{D}_m^\top\right\| \le \frac{1}{T}\sum_{m=1}^T \mathbb{E}\|\boldsymbol{C}_m\|\|\boldsymbol{D}_m\|$$

Table 2: Asymptotic critic values $q_{\alpha,\beta}$ of $t^*(\beta)$ defined by $q_{\alpha,\beta} = \min\{t : \mathbb{P}(t^*(\beta) \le t) \ge 1-\alpha\}$.

β \ $1-\alpha$	1%	2.5%	5%	10%	50%	90%	95%	97.5%	99%
0	-8.634	-6.753	-5.324	-3.877	0.000	3.877	5.324	6.753	8.634
$1/3$	-8.0945	-6.339	-5.048	-3.712	0.000	3.712	5.048	6.339	8.0945
$1/2$	-7.386	-5.851	-4.621	-3.446	0.000	3.446	4.621	5.851	7.386
$2/3$	-6.292	-4.993	-4.012	-3.027	0.000	3.027	4.012	4.993	6.292

$$
\begin{aligned}
&\le \frac{1}{T}\sum_{m=1}^{T}\sqrt{\mathbb{E}\|\boldsymbol{C}_m\|^2\mathbb{E}\|\boldsymbol{D}_m\|^2} \\
&\lesssim \frac{1}{T}\sum_{m=1}^{T}\sqrt{\gamma_m} \to 0.
\end{aligned}
$$

Finally, for the last term, we have that

$$
\mathbb{E}\left\|\frac{1}{T}\sum_{m=1}^{T}\boldsymbol{D}_m\boldsymbol{D}_m^\top\right\| \le \frac{1}{T}\sum_{m=1}^{T}\mathbb{E}\left\|\boldsymbol{D}_m\right\|^2 \lesssim \frac{1}{T}\sum_{m=1}^{T}\gamma_m \to 0.
$$

Hence, $\widehat{\boldsymbol{S}}_T$ converges to $\boldsymbol{S}$ in probability.

Appendix E. Computation of Critical Values

For easy reference, critical values of $t^*(\beta)$ are computed via simulations and listed in Table 2. In particular, the Brownian motion $B_1(\cdot)$ is approximated by normalized sums of i.i.d. $\mathcal{N}(0,1)$ pseudo random deviates using 1,000 steps and 50,000 replications. We then smooth the 50,000 realizations by standard Gaussian-kernels techniques with the bandwidth selected according to Scott's rule (Scott, 2015). Kernel density estimation is a way to estimate the probability density function of a random variable in a non-parametric way. Because we smooth the data, our critical values of the case $\beta = 0$ are slightly different from previous computations by Kiefer et al. (2000). In particular, when $1-\alpha = 97.5\%$ and $\beta = 0$, our critical value 6.753 is smaller than previous 6.811, which shrinks the length of our confidence intervals. Our critical value 6.753 is also close to 6.747 computed in Abadir and Paruolo (1997).

Appendix F. Related Work on Local SGD

Federated learning enables a large amount of edge computing devices to jointly learn a global model without data sharing (Kairouz et al., 2019). The seminal paper McMahan et al. (2017) proposed *Federated Average* (FedAvg) for FL, which is slightly different from Local SGD that we focus on in this work. The main difference is that FedAvg randomly samples a small portion of clients at the beginning of each communication round to alleviate the straggler effect caused by massively distributed clients. When all clients are forced to participate, FedAvg is reduced to Local SGD. Their theoretical convergence does not vary too much with an additional statistical error incurred when clients participate partially (Li et al., 2019a). There has been a rapidly growing line of work concerning various aspects of FedAvg and its variants recently (Zhao et al., 2018; Sahu et al., 2018; Nishio and Yonetani, 2018; Koloskova et al., 2020; Yuan and Ma, 2020; Yuan et al., 2021; Zheng et al., 2021). Local SGD or Fedavg is an iterative and multi-round distributed algorithm that communicates only gradient information at each communication round. Other algorithms of this type have been proposed and analyzed previously (Shamir et al., 2014; Wang et al., 2017; Jordan et al., 2018; Fan et al., 2019). The biggest difference is that Local SGD lowers the communication frequency, while others do not. This simple change improves communication efficiency greatly (Lin et al., 2018).

Appendix G. Numerical Simulations

This section investigates the empirical performance of the plug-in and random scaling methods via Monte Carlo experiments. We consider both the linear and logistic regression models. At iteration t, the k-th client observes the pair $(\boldsymbol{a}_t^k, b_t^k)$ with $\boldsymbol{a}_t^k$ the d-dimensional covariates generated from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and b_t^k the response generated according to the model. We detail the data generation process as follows:

- In linear regression, $b_t^k = (\boldsymbol{a}_t^k)^\top \boldsymbol{x}_k^* + \varepsilon_t^k$ where the ε_t^k are i.i.d. according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\boldsymbol{x}_k^*$ is the true local parameter which we also generate from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. In this case, the global parameter $\boldsymbol{x}^*$ is the average of $\boldsymbol{x}_k^*$'s.
- In logistic regression, $b_t^k \in \{0, 1\}$ is generated to be 1 with probability $\sigma((\boldsymbol{a}_t^k)^\top \boldsymbol{x}^*)$ and 0 with probability $1 - \sigma((\boldsymbol{a}_t^k)^\top \boldsymbol{x}^*)$. Here $\sigma(\theta) = 1/(1 + \exp(-\theta))$ is the sigmoid function. We do not impose data heterogeneity for logistic regression in order to avoid numerical error in the calculation of $\boldsymbol{x}^*$. Here $\boldsymbol{x}^*$ is equi-spaced on the interval $[0, 1]$ following previous works (Chen et al., 2020; Lee et al., 2021).

We set $\gamma_m = \gamma_0 / m^{0.505}$ with $\gamma_0 = 0.5$ for linear regression and $\gamma_0 = 2$ for logistic regression. The initial value $\bar{\boldsymbol{x}}_0$ is set as zero. We fix $K = 10$ in all our experiments and vary the number of rounds T. In all cases, we set $E_m = 1$ for the first 5% observations as a warm-up and then increase E_m from scratch, i.e., $E_m = E'_{m-5\%*T}$ for another sequence $\{E'_m\}$. We consider six choices of $\{E'_m\}_m$, namely (i) `C1`: constant $E'_m \equiv 1$, (ii) `C5`: constant $E'_m \equiv 5$, (iii) `Log`: logarithmic $E'_m = \lceil \log_2(m+1) \rceil$, (iv) `P(1/3)`: power $E'_m = \lceil m^{1/3} \rceil$, (v) `P(1/2)`: power $E'_m = \lceil m^{1/2} \rceil$, and (vi) `P(2/3)`: power $E'_m = \lceil m^{2/3} \rceil$. The nominal coverage probability is set at 95%. The performance is measured by three statistics: the coverage rate, the average length of the 95% confidence interval, and the average communication frequency. For brevity, we focus on the first coefficient $\boldsymbol{x}_1^*$ hereafter. All the reported results are obtained by taking the average of 1000 independent runs.

We first turn to study the communication efficiency for Local SGD. From Figure 1, we find the faster E_m grows, the faster the L_2 convergence in terms of communication, which is consistent with previous studies from optimization perspective (McMahan et al., 2017; Lin et al., 2018). Figure 2 shows the empirical coverage rates and confidence interval lengths in linear regression, both obtained by averaging over 1000 Local SGD paths. The result of logistic regression is depicted in Figure 3. For plug-in, though wandering above 90%, the faster E_m family (namely, `Log`, `P(1/3)` and `P(1/2)`) has relatively inferior coverage rate than the slower E_m family (namely, `C1` and `C5`). The coverage rate of `P(2/3)` can't even cross 90%. For random scaling, it is clear that the coverage rate of all

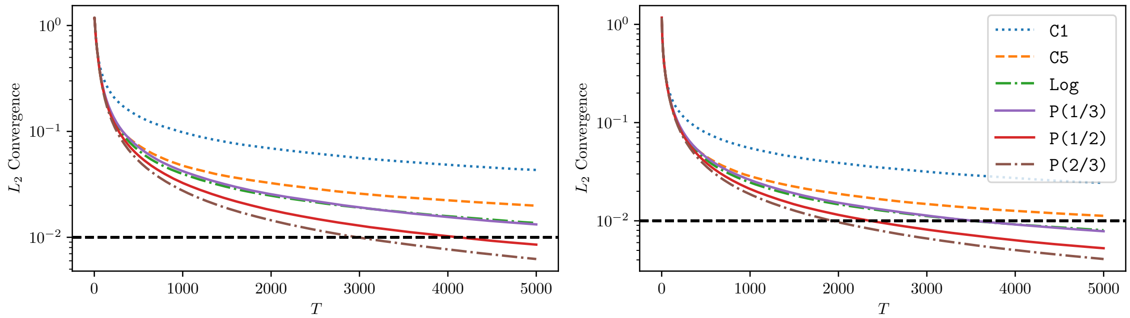

Figure 1: L_2 convergence $\|\bar{\boldsymbol{y}}_T - \boldsymbol{x}^*\|$ in terms of communication T. Left: Results of linear regression. Right: Results of logistic regression. Black dashed line denotes the nominal coverage rate of 95%.

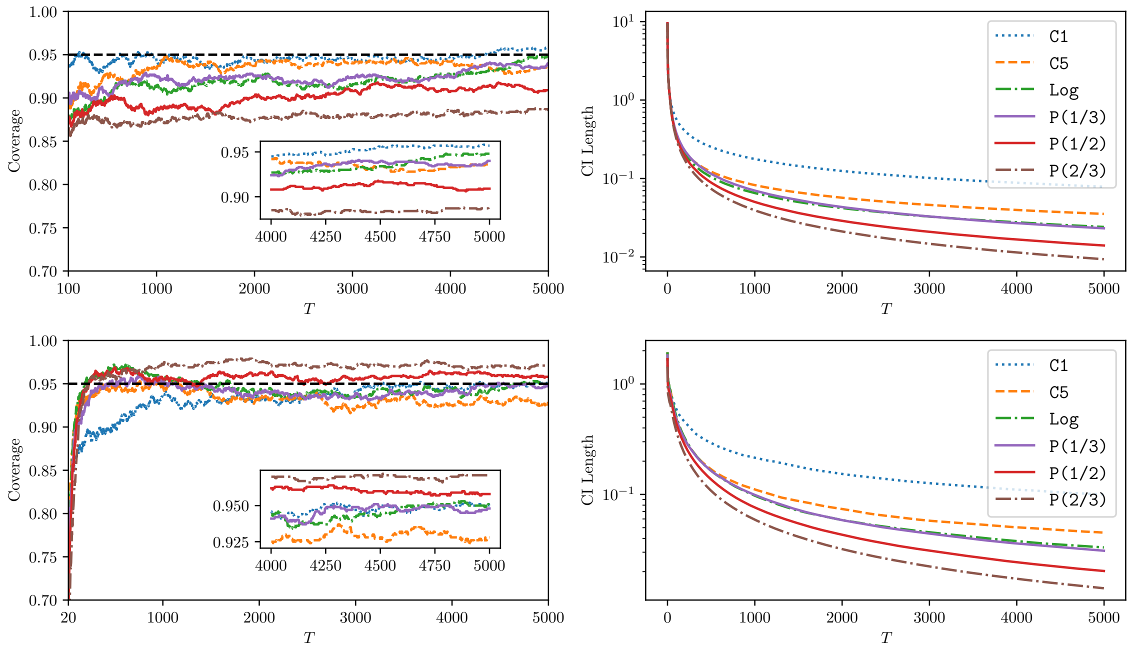

Figure 2: Comparison of the plug-in (the top row) and random scaling (the bottom row) in linear regression. Left: Empirical coverage rate against the number of communication. Black dashed line denotes the nominal coverage rate of 95%. Right: Length of confidence intervals.

the methods fluctuates around 95%. Though with a much smaller deviation from 95%, the slow E_m family has the slower shrinkage rate for its confidence interval. By contrast, the faster E_m family achieves comparable coverage with faster shrinkage of confidence intervals. It implies that Local SGD has high efficiency of communication and maintains a good statistic efficiency via random scaling.

We then turn to the empirical performance of Local SGD with limited computation or finite samples. Table 3 shows the empirical performance of the six methods under linear models with four different t_T's. t_T is actually the total iteration each client runs through T rounds or equivalently the number of observations they receive. From the table, almost all the methods achieve good

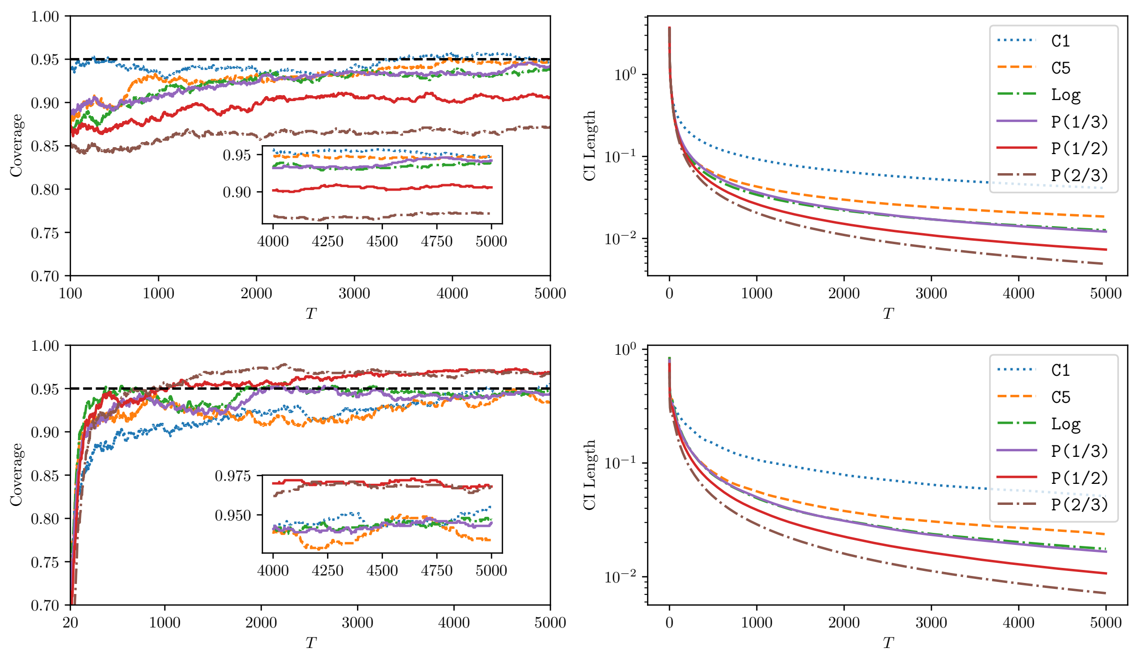

Figure 3: Comparison of the plug-in (the top row) and random scaling (the bottom row) estimators in logistic regression. Left: Empirical coverage rate against the number of communication. Black dashed line denotes the nominal coverage rate of 95%. Right: Length of confidence intervals.

performance. Except `P(2/3)`, random scaling gives better average coverage rates than the plug-in method, because its average coverage rates of all different communication intervals are near (or even exceed) 95%. However, its average length is usually larger than that of plug-in. Furthermore, its average length usually has a much larger deviation than that of plug-in. For example, when $t_T = 5000$, for `C5`, the standard deviation of average lengths for plug-in is 0.807×10^{-2}, while it increases to 3.714×10^{-2} for random scaling. Such a wider average length might account for the unexpected advantage on the average coverage rates. We speculate the reason for the poor performance of `P(2/3)` is because less frequent communication enlarges asymptotic variance and decrease the sample efficiency. It might require more samples to reach a counterpart level of coverage rates. However, as the communication round increases and more observations are available, the average length decreases and the coverage rate increases, with both deviations reduced. The poor performance of `P(2/3)` implies that when E_m grows too faster (e.g., $E_m = \lceil m^2 \rceil$), its performance might deteriorate, accordant to our Theorem 4.2.

In addition, comparing the results of `Log`, `P(1/3)`, and `P(1/2)`, we can find that the faster E_m increases, the larger average length as well as its standard deviations. However, they all have satisfactory performance when observations are sufficient. Indeed, Local SGD trades more computation for less communication, resulting in a residual error gradually accumulated when communication is off, slowing down the convergence rate and enlarging asymptotic variance (e.g., the existence of ν). However, the benefit is also attractive: the averaged communication frequency is substantially reduced and the convergence in terms of communication largely increases. It implies that Local SGD obtains both statistical efficiency and communication efficiency as expected. We further consider the logistic regression, which is a standard non-linear model. The result is given in

Table 3: Simulation results of linear regression with $d = 5$. The standard errors of coverage rates $\widehat{p}$ are computed via $\sqrt{\widehat{p}(1-\widehat{p})/1000} \times 100\%$ and reported inside the parentheses.

Methods	Items		$t_T = 5000$	$t_T = 10000$	$t_T = 20000$	$t_T = 40000$
Plug-in	Cov Rate (%)	`C1`	95.70(0.641)	94.20(0.739)	94.20(0.739)	93.80(0.763)
		`C5`	93.70(0.768)	94.00(0.751)	94.30(0.733)	93.10(0.801)
		`Log`	91.70(0.872)	93.20(0.796)	93.80(0.763)	93.80(0.763)
		`P(1/3)`	91.90(0.863)	92.70(0.823)	93.90(0.757)	93.60(0.774)
		`P(1/2)`	91.10(0.900)	92.60(0.828)	93.90(0.757)	93.80(0.763)
		`P(2/3)`	91.00(0.905)	92.60(0.828)	93.40(0.785)	93.60(0.774)
	Avg Len (10^{-2})	`C1`	7.857(0.099)	5.547(0.050)	3.917(0.025)	2.768(0.013)
		`C5`	9.737(0.242)	6.868(0.121)	4.847(0.061)	3.423(0.031)
		`Log`	12.168(0.371)	8.953(0.204)	6.602(0.106)	4.864(0.058)
		`P(1/3)`	11.372(0.336)	8.656(0.195)	6.613(0.110)	5.059(0.063)
		`P(1/2)`	15.431(0.559)	12.100(0.327)	9.433(0.188)	7.300(0.112)
		`P(2/3)`	19.593(0.791)	15.375(0.491)	11.896(0.274)	9.083(0.156)
Random Scaling	Cov Rate (%)	`C1`	95.00(0.689)	93.90(0.757)	93.70(0.768)	94.80(0.702)
		`C5`	97.70(0.474)	96.90(0.548)	97.20(0.522)	96.90(0.548)
		`Log`	98.20(0.420)	98.70(0.358)	98.90(0.330)	98.80(0.344)
		`P(1/3)`	97.60(0.484)	98.20(0.420)	98.50(0.384)	98.00(0.443)
		`P(1/2)`	96.00(0.620)	97.20(0.522)	96.40(0.589)	96.60(0.573)
		`P(2/3)`	88.70(1.001)	89.90(0.953)	90.70(0.918)	90.00(0.949)
	Avg Len (10^{-2})	`C1`	10.011(4.343)	7.081(3.106)	5.010(2.092)	3.605(1.511)
		`C5`	14.434(6.950)	10.043(4.923)	7.078(3.389)	4.946(2.448)
		`Log`	19.187(9.763)	14.120(7.154)	10.430(5.219)	7.611(3.895)
		`P(1/3)`	16.781(8.397)	12.810(6.460)	9.821(4.906)	7.440(3.777)
		`P(1/2)`	20.888(10.842)	16.127(8.004)	12.379(6.027)	9.314(4.460)
		`P(2/3)`	21.495(11.324)	16.463(7.991)	12.509(5.924)	9.276(4.325)

Table 4. A similar pattern is observed: random scaling has higher average coverage rates at the price of wider average lengths which typically shrink as more observations are generated.

Table 4: Simulation results of logistic regression with $d = 5$. The standard errors of coverage rates $\widehat{p}$ are computed via $\sqrt{\widehat{p}(1-\widehat{p})/1000} \times 100\%$ and reported inside the parentheses.

Methods	Items		$t_T = 5000$	$t_T = 10000$	$t_T = 20000$	$t_T = 40000$
Plug-in	Cov Rate (%)	C1	94.70(0.708)	93.50(0.780)	94.60(0.715)	95.40(0.662)
		C5	93.00(0.807)	92.30(0.843)	93.50(0.780)	94.10(0.745)
		Log	92.30(0.843)	92.10(0.853)	92.60(0.828)	92.90(0.812)
		P(1/3)	92.70(0.823)	92.00(0.858)	92.50(0.833)	92.90(0.812)
		P(1/2)	90.80(0.914)	92.20(0.848)	91.70(0.872)	92.10(0.853)
		P(2/3)	90.90(0.909)	92.80(0.817)	91.30(0.891)	92.20(0.848)
	Avg Len (10^{-2})	C1	4.113(0.046)	2.903(0.022)	2.049(0.011)	1.448(0.005)
		C5	5.081(0.118)	3.587(0.057)	2.534(0.029)	1.790(0.014)
		Log	6.347(0.175)	4.681(0.093)	3.453(0.049)	2.544(0.027)
		P(1/3)	5.949(0.146)	4.526(0.091)	3.456(0.049)	2.647(0.027)
		P(1/2)	8.062(0.256)	6.320(0.149)	4.927(0.088)	3.821(0.052)
		P(2/3)	10.254(0.380)	8.036(0.218)	6.223(0.127)	4.752(0.070)
Random Scaling	Cov Rate (%)	C1	95.50(0.656)	92.40(0.838)	94.10(0.745)	94.70(0.708)
		C5	96.00(0.620)	95.90(0.627)	96.80(0.557)	95.80(0.634)
		Log	97.60(0.484)	97.40(0.503)	97.80(0.464)	98.20(0.420)
		P(1/3)	96.10(0.612)	96.60(0.573)	97.50(0.494)	97.90(0.453)
		P(1/2)	94.40(0.727)	94.30(0.733)	94.50(0.721)	95.10(0.683)
		P(2/3)	88.30(1.016)	88.00(1.028)	86.80(1.070)	88.80(0.997)
	Avg Len (10^{-2})	C1	5.112(2.302)	3.612(1.502)	2.646(1.162)	1.877(0.816)
		C5	7.296(3.714)	5.166(2.535)	3.687(1.836)	2.637(1.316)
		Log	9.703(5.176)	7.241(3.713)	5.383(2.787)	4.023(2.063)
		P(1/3)	8.499(4.465)	6.569(3.345)	5.071(2.621)	3.924(1.999)
		P(1/2)	10.574(5.688)	8.278(4.193)	6.340(3.194)	4.880(2.366)
		P(2/3)	10.915(5.876)	8.497(4.244)	6.373(3.147)	4.850(2.293)