北京大学

博士研究生学位论文

题目：关于联邦学习以及非线性随机近似的在线统计推断

姓　　名：李翔

学　　号：1801110058

院　　系：数学科学学院

专　　业：统计学

研究方向：机器学习

导　　师：张志华教授

二〇二三 年 六 月

版权声明

摘要

本论文研究如何使用在机器学习和统计学中流行的随机迭代算法来进行在线统计推断，并分为两部分。

在第一部分，我们着重研究联邦学习。这是一种相对较新的分布式机器学习领域。它允许端设备（例如智能手机和便携设备）协同学习一个公共的统计模型，同时禁止共享本地生成的数据。我们分析了当前流行的局部随机梯度下降方法（Local SGD）。这是一种多轮通讯的估计方法。它企图通过降低通讯频率来提高通讯效率。我们探讨了如何使用同步了的平均序列构造渐近有效的置信区间。我们提出了两种方法来构造置信区间：插值方法和随机缩放。前者利用我们建立起的渐近正态性质以及在线估计的渐近方差矩阵来构建置信区间，而后者则利用整个 Local SGD 轨迹的信息来构造渐近有效的枢轴量。为了从理论上支持后一种方法，我们建立了一个函数中心极限定理。在当前最弱的关于随机梯度的 $2+\delta$ 矩假设下，我们证明了平均了的 Local SGD 序列的部分和过程会弱收敛于一个缩放的布朗运动。我们的结果表明这两种方法不仅通讯有效，而且适用于在线数据。此外，当以适当的频率通讯时，Local SGD 可以在逐渐减少至零的通讯频率下达到最优的统计估计效率。

在第二部分，我们研究了仅使用单条马尔科夫数据的非线性随机逼近算法的统计推断。主要的应用场景包括自回归数据上的随机梯度下降和强化学习中的异步 Q 学习方法。我们使用标准随机逼近框架估计目标参数，并为其部分和过程（记为 $\boldsymbol{\phi}_T$）建立了一个函数中心极限定理。为了进一步支持这个理论，我们一方面提供了一个半参数的渐近方差矩阵下界来展示该过程的方差最优性，另一方面刻画了在 Lévy-Prokhorov 度量下的该随机过程弱收敛的收敛速度，量化部分影响因素。我们的推断方法基于建立的函数中心极限定理——通过选择任何连续的尺度不变的泛函 f，我们可以构造出渐近有效的枢轴量 $f(\boldsymbol{\phi}_T)$。基于此，我们能够构造一个渐近有效的置信区间。我们提出了一个由 $m\in\mathbb{N}$ 索引的函数族 f_m，并通过理论和数值手段分析了相应的拒绝概率。最后通过数值试验验证了我们的理论的正确性以及方法的有效性。

关键词：联邦学习，统计推断，非线性随机近似，函数中心极限定理

Online Statistical Inference for Federated Learning and Nonlinear Stochastic Approximation

Xiang Li (Statistics)
Directed by: Prof. Zhihua Zhang

ABSTRACT

This dissertation investigates the implementation of popular stochastic iterative algorithms in machine learning and statistics for online statistical inference.

In the first part, we focus on Federated Learning (FL), a relatively new field of distributed machine learning that allows end devices (such as smartphones and portable devices) to collaboratively learn a shared model without sharing locally generated data points. We analyze Local SGD, a multi-round estimation procedure that uses intermittent communication to improve communication efficiency, and explore how to construct asymptotically valid confidence intervals using synchronized iterates. We present two methods for constructing these intervals: the plug-in method, which estimates the asymptotic variance matrix and constructs confidence intervals via the established asymptotic normality, and random scaling, which uses information from the entire Local SGD trajectory to construct an asymptotically pivotal statistic. To support the second method, we establish a functional central limit theorem that shows the partial-sum process of averaged Local SGD iterates weakly converges to a scaled Brownian motion under the weakest bounded $2 + \delta$-moment assumption on stochastic gradients. Our results demonstrate that both methods are communication-efficient and applicable to online data. Furthermore, once communicating at an appropriate frequency, Local SGD achieves both statistical and communication efficiency simultaneously.

In the second part, we investigate the statistical inference of nonlinear stochastic approximation algorithms using a single trajectory of Markovian data. Our approach has practical applications in various scenarios, such as Stochastic gradient descent (SGD) on autoregressive data and asynchronous Q-Learning. We estimate the target parameter using the standard stochastic approximation (SA) framework and establish a functional central limit theorem for its partial-sum process, denoted by $\boldsymbol{\phi}_T$. To further support this theory, we provide, on one hand, a semi-parametric efficient asymptotic variance matrix lower bound to demonstrate the variance optimality of the process. On the other hand, we characterize the convergence rate of weak convergence of this stochastic process under the Lévy-Prokhorov metric, quantifying some influencing factors. The functional central limit theorem serves as the foundation for our inference method. By selecting any continuous scale-invariant functional f, the asymptotic pivotal statistic $f(\boldsymbol{\phi}_T)$ becomes accessible, enabling us to construct an asymptotically valid confidence interval. We propose a family of functionals f_m, indexed by $m \in \mathbb{N}$, and analyze the corresponding rejection probability through theoretical and numerical means. Our simulation results demonstrate the validity and efficiency of our method.

KEY WORDS: Federated learning, Statsitical inference, Nonlinear stochastic approximation, Functional central limit theorem

Contents

Chapter 1 Introduction

Statistical estimation and statistical inference are two fundamental concepts in statistics that are used to make sense of data and draw meaningful conclusions from it[1-2]. Statistical estimation is the process of using sample data to estimate unknown parameters of a population, such as its mean or standard deviation. This is typically done using point estimates. On the other hand, statistical inference is the process of using sample data to draw conclusions about the population from which the sample was drawn. This can involve hypothesis testing, where we test a claim about a population parameter, or constructing confidence intervals to estimate the uncertainty around our estimates. Typically, statistical inference is more difficult than statistical estimation because the former needs to figure out the inherent uncertainty and variability in real-world data while the later only cares about a point estimation.

In nowadays modern statistics, both of them are facing great challenges due to the incredibly large data volume generated by ubiquitous man-machine interaction[3]. Browsers gather shopping history, cellphones collect keyboard inputs, potable devices record sporting activity, and, institutions (e.g., banks and hospitals) copy service information. In these real applications, two difficulties arises, namely the large data volume and its online generation.

The first challenge is related to the size of the dataset, which may be too large to fit into memory or process on a single machine. In these cases, a distributed setting must be considered where data points are generated on different devices. Once data points are generated locally, it would be privacy-destructive to upload these raw data points to an unauthenticated central server (that often locates in privates companies or governmental institutes). Furthermore, it violates privacy and policies because many rigorous data protection regulations are launched to regulate personal data usage such as EU/UK General Data Protection Regulation (GDPR)[4]. To address this difficulty, *Federated Learning* (FL) has been proposed as a solution, which allows multiple devices to collaboratively learn a shared statistical model without sharing local datasets directly[5]. In this way, FL can protect sensitive information that data contain, such as personal identity information and state of health information, from unauthorized access of service providers. However, limited data access, together with memory constraints, communication budget, and computation restrictions, makes traditional statistical estimation and inference methods inapplicable in the FL scenario[6-7].

The second challenge arises when data points are generated continuously, making it physically impossible to collect them all in one dataset pool. Selecting only a small portion of these points can result in loss of information and statistical power. Additionally, as new data points arrive, it becomes computationally inefficient to redo the estimation from the expanded dataset, which can cause significant delays in real-time applications such as online recommendation and autonomous driving. To address this challenge, we need an estimation procedure that is adaptive to the streaming data setting, allowing us to do incremental updates when new data points arrive without requiring a full re-estimation of the model. This is the essence of online learning[8] and reinforcement learning[9], which enable agents to make decisions in real-time based on incoming data while adapting to changes in the underlying environment.

To handle these two challenges posed by a big volume of streaming data, efficient online algorithms have been the developed for statistical estimation. For example, in federated learning, Local SGD is proposed to improve communication efficiency by decreasing communication frequency[10]. More specifically, Local SGD runs stochastic gradient descent (SGD) independently in parallel on different clients and averages the sequences only once in a while. It has been shown to have superior performance in training efficiency and scalability[11], and converge fast in terms of communication[12-16]. In reinforcement learning, Q-Learning is perhaps the most popular model-free approach to estimate the optimal value function, which is the optimal expected accumulated rewards of taking a given action in a given state[17]. In practice, one key feature of Q-learning is that its observed data is generated from a trajectory of Markov chain. More specifically, by performing an action at the current state, only an instant reward variable and the next state are observed, which implies only incomplete and noisy data are available. It has been show that once assuming a bounded mixing time for the underlying Markov chain, non-asymptotic convergence rates are still accessible[18-21].

Despite significant progress in achieving fast and even optimal non-asymptotic convergence guarantees for both FL and RL, conducting statistical inference in these contexts remains an open problem. The main difficulty is to quantify the randomness of a proposed estimate and further to propose an effective online procedure to construct confidence intervals.

In the case of FL, an effective statistical inference method must balance communication efficiency (that is less communication starving), statistical heterogeneity (that is adaptive to the different local data distributions), and statistical efficiency, with the goal of achieving the Cramér–Rao lower bound. A classic approach is one-shot averaging or divide-and-conquer method that performs only one communication to average the output of each devices for distributed tasks[22-28]. As equivalent to the single-agent setting, one-shot averaging is simple and easy to analyze, whose asymptotic convergence has been studied extensively in the early years[29-32]. Typically, one-shot averaging works well when the bias of the predictor returned by each client is much smaller than the variance[33]. When each device has a sufficient number of data that are generated independently from (nearly) the same distribution, a small estimation error is well guaranteed in theory[24, 34]. However, the statistical heterogeneity in FL is likely to render the predictor outputted by each device a larger bias and makes the one-shot average invalid. Another approach focuses on a multi-communication procedure due to its stable performance and weak requirement on local dataset size[35-36]. An extreme example is parallel SGD[37] that alternates between one independent step of SGD in parallel and one synchronization. However, the feature that parallel SGD (as well as its many variants) performs one communication per round would incur huge communication costs in an online setting where the data arrive sequentially. By contrast, Local SGD performs one communication after several (even an increasing number of) rounds and intuitively improve the communication efficiency. Though Local SGD and its many variants improve communication efficiency for many federated estimation tasks, no works consider to quantify the randomness of obtained estimates, let alone perform statistical inference. To the best of our knowledge, no inference method for FL has met the criteria mentioned early.

In RL, a satisfactory statistical inference method must not only be statistically efficient but also be able to handle trajectory Markovian data. Early works in RL often rely on the generator assumption, which assumes independent rewards and independent next states for all state-action pairs. However, even with this assumption of independence and completeness, quantifying the randomness for RL algorithms in an online manner is challenging, as there is no direct access to the curvature information for estimating the asymptotic variance. In contrast, in the case of stochastic gradient descent (SGD), once the Hessian matrix is available, one can use a plug-in estimator or a batch-mean estimator for the asymptotic variance under i.i.d. data[38-39]. To address this issue in RL, most existing works on statistical inference mainly rely on bootstrap-based methods, where multiple perturbed iterates are maintained to approximate the asymptotic variance matrix when the number of perturbed iterates is sufficiently large[40-43]. However, this line of study has several limitations. First, most existing works focus on the off-policy evaluation (OPE) problem, where the agent evaluates the performance of a hypothetical policy using only offline i.i.d. log data. As OPE is essentially a linear problem, it is unclear whether this approach can be extended to nonlinear problems. Second, few works consider the Markovian data setting. One exception is[43], which, however, still focuses on OPEs. Lastly, bootstrap-based methods require multiple oracles that the agent is able to evaluate the values of stochastic incremental updates at different parameters while keeping the source of randomness unchanged. This oracle is only feasible in environments where the agent can fully control the environment. Without complete control of the environment, there is currently no known method to estimate the asymptotic variance in the presence of Markovian data. To summarize, conducting statistical inference under the existence of Markovian data and without multiple oracles remains an open research problem.

In this thesis, we are motivated to tackle the challenge of conducting statistical inference in these two challenging settings. The contributions of the thesis related both to theoretical and practical sides of our findings are listed in the next section along with a brief explanation.

1.1 Contributions of the Thesis

Before we give a detailed overview with precise definitions and explanations of the concepts briefly introduced above, we present the list of the main contributions of the thesis from a high-level perspective, and we cite the related publications. Detailed descriptions for each point will be given in Sections 2.1.1 and 3.1.1 respectively.

Statistical inference for FL In order to perform statistical inference in federated learning (FL), an optimization algorithm must be chosen first to estimate the target parameter. Local SGD is selected here due to its simplicity and representativeness in FL. The reason is that as the key feature of Local SGD, local updates or intermittent communication has motivated a lot federated algorithms to improve communication efficiency in application such as device sampling[5, 44], distributed PCA[45-46]., non-convex optimization[12], and minimax optimization[47-49]. Local SGD is the simplest in the sense that it is the very combination of local updates and SGD. It runs SGD independently in parallel on different devices and averages the sequences only occasionally[5].

Chapter 2 focuses on Local SGD and establishes its asymptotic normality. We find that local updates, or intermittent communication, introduce an interesting trade-off between statistical and communication efficiency (see Section 2.3.1 in Chapter 2). By decaying the communication frequency at an appropriate rate, Local SGD can achieve both efficiency measures. Based on the asymptotic normality, an online plug-in estimator for the asymptotic variance is proposed. When the second-order information, such as the Hessian matrix, is not available, a non-parametric inference method is then proposed. The key idea is to construct an asymptotically pivotal statistic by using information along the whole Local SGD trajectory. To support this method, a functional central limit theorem is established, which shows that the averaged iterates of Local SGD converge weakly to a scaled Brownian motion under the currently weakest $2+\delta$ moment condition on stochastic gradients. This method is more computationally efficient and memory-friendly than the plug-in method. Numerical experiments are conducted to illustrate both inference methods.

This contribution is based on the following publication[50]. Before preparing this work in statistical inference for Local SGD, the author has already applied local updates for other applications, such as providing the non-asymptotic convergence rate for a variant of Local SGD[51], and proposing efficient algorithms for decentralized optimization[12] or distributed PCA[46]. To maintain simplicity and relevance, these papers are not included in this thesis.

- [50]Li X, Liang J, Chang X, Zhang Z. Statistical estimation and online inference via Local SGD[C]// Conference on Learning Theory: vol. 178. [S.l.]: PMLR, 2022: 1613-1661.

Statistical inference for nonlinear stochastic approximation In Chapter 3, we adopt a more general approach to study Q-Learning by examining it through the lens of nonlinear stochastic approximation (SA). Q-Learning can be seen as a recursive stochastic procedure that aims to find the root of a given nonlinear equation. Nonlinear stochastic approximation is a class of methods that has been studied for the past two decades[52-54]. Since Q-Learning is an important case of the single trajectory case, the absence of multiple oracles makes it difficult to apply previous online bootstrap methods[43]. To address this issue, we propose to utilize trajectory information to construct an asymptotically pivotal statistic that allows us to obtain confidence intervals by inverting it. To support this theory, we establish a functional central limit theorem for the partial-sum process of nonlinear SA methods, denoted by $\boldsymbol{\phi}_T$, that shows weak convergence to a scaled Brownian motion, even if the data is generated along a Markov chain. To further support our findings, we provide a semiparametric efficient lower bound and a non-asymptotic upper bound on weak convergence, measured in the Lévy-Prokhorov metric. By selecting any continuous scale-invariant functional f, we can make the asymptotic pivotal statistic $f(\boldsymbol{\phi}_T)$ accessible, which allows us to construct an asymptotically valid confidence interval. We analyze the rejection probability of a family of functionals f_m indexed by $m \in \mathbb{N}$ through theoretical and numerical means, and the simulation results demonstrate the validity and efficiency of our method.

This contribution is based on the following paper[55]. It is a follow-up work to our previous conference publication[21], where we analyze (synchronous) Q-Learning under a weaker condition, namely the synchronous setting where a generative model produces independent samples in every iteration[56]. To overcome the difficulty brought by Markov data and iterative algorithms, we made several technical extensions. See Section 3.1.1 for more details.

- [55]Li X, Liang J, Zhang Z. Online statistical inference for nonlinear stochastic approximation with Markovian data[J]. ArXiv preprint arXiv:2302.07690, 2023.
- [21]Li X, Yang W, Jiadong L, Zhang Z, Jordan M I. A statistical analysis of Polyak-Ruppert averaged Q-learning[C]//International Conference on Artificial Intelligence and Statistics: vol. 206. [S.l. : s.n.], 2023.

Other paper organization In Chapter 4, we provide a brief comparison of different inference methods and highlight potential future work. To facilitate comprehension of our inference methods, we introduce some preliminaries on weak convergence in metric spaces. These concepts are essential to understanding our proposed methods. All notations will be introduced in the corresponding chapters.

1.2 Preliminaries on Weak Convergence in Metric Spaces

We will introduce some basic knowledge of weak convergence in metric spaces. See Section 12-15 in the book of Billingsley [57] for a detailed introduction.

A Polish space is a topological space that is separable, complete, and metrizable. Let

$$\mathsf{D}_{[0,1],\mathbb{R}^d} = \{\boldsymbol{\phi} : \text{càdlàg function } \boldsymbol{\phi}(r) \in \mathbb{R}^d, r \in [0,1]\}$$

collect all d-dimensional functions which are right continuous with left limits. These functions are also known as *càdlàg* functions. The J_1 *Skorokhod topology* equips $\mathsf{D}_{[0,1],\mathbb{R}^d}$ with the *Skorokhod metric* d_{S} such that $(\mathsf{D}_{[0,1],\mathbb{R}^d}, d_{\mathrm{S}})$ is a Polish space and $\mathscr{D}_{[0,1],\mathbb{R}^d}$ is its Borel σ-field (the σ-field generated by all open subsets) in the Skorokhod metric. In particular, denoting by Λ the class of strictly increasing continuous mappings $\lambda : [0,1] \to [0,1]$ with $\lambda(0) = 0$ and $\lambda(1) = 1$, we have for any $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathsf{D}_{[0,1],\mathbb{R}^d}$,

$$d_{\mathrm{S}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \inf_{\lambda\in\Lambda} \max\left\{ \sup_{0\le s<t\le 1} \left| \ln \frac{\lambda(t)-\lambda(s)}{t-s} \right|, \sup_{t\in[0,1]} \|\boldsymbol{\phi}_1(\lambda(t)) - \boldsymbol{\phi}_2(t)\| \right\}. \tag{1.1}$$

An important closed subset of $\mathsf{D}_{[0,1],\mathbb{R}^d}$ is

$$\mathsf{C}_{[0,1],\mathbb{R}^d} := \{\boldsymbol{\phi} : \text{continuous } \boldsymbol{\phi}(r) \in \mathbb{R}^d, r \in [0,1]\},$$

which collects all d-dimensional continuous functions defined on $[0,1]$. The *uniform topology* equips $\mathsf{C}_{[0,1],\mathbb{R}^d}$ with the uniform metric $|||\boldsymbol{\phi}||| := \sup_{r\in[0,1]} \|\boldsymbol{\phi}(r)\|$ such that $(\mathsf{C}_{[0,1],\mathbb{R}^d}, |||\cdot|||)$ is a Polish space. Furthermore, we have $d_{\mathrm{S}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \leq |||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2|||$ for any $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathsf{D}_{[0,1],\mathbb{R}^d}$, implying the J_1 Skorokhod topology is weaker than the uniform topology. Unfortunately, $(\mathsf{D}_{[0,1],\mathbb{R}^d}, |||\cdot|||)$ is not separable and we have $\mathscr{D}_{[0,1],\mathbb{R}^d} \subsetneq \mathscr{C}_{[0,1],\mathbb{R}^d}$ with $\mathscr{C}_{[0,1],\mathbb{R}^d}$ the Borel σ-field in the uniform metric.

Any random element $\boldsymbol{\phi}_t \in \mathsf{D}_{[0,1],\mathbb{R}^d}$ introduces a probability measure on $\mathsf{D}_{[0,1],\mathbb{R}^d}$ denoted by $\mathscr{L}(\boldsymbol{\phi}_t)$ such that $(\mathsf{D}_{[0,1],\mathbb{R}^d}, \mathscr{D}_{[0,1],\mathbb{R}^d}, \mathscr{L}(\boldsymbol{\phi}_t))$ becomes a probability space. We say a sequence of random elements $\{\boldsymbol{\phi}_t\}_{t\geq 0} \subseteq \mathsf{D}_{[0,1],\mathbb{R}^d}$ *weakly converges* to $\boldsymbol{\phi}$, if for any bounded, continuous, $\mathscr{D}_{[0,1],\mathbb{R}^d}$-measurable functional $f : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$, we have $\mathbb{E}f(\boldsymbol{\phi}_T) \to \mathbb{E}f(\boldsymbol{\phi})$ as $T \to \infty$. The condition is equivalent to that any finite-dimensional projections of $\boldsymbol{\phi}_T$ converge in distribution in the sense that for any given integer $n \geq 1$ and any $0 \leq t_1 < \cdots < t_n \leq 1$, when T goes to infinity,

$$(\boldsymbol{\phi}_T(t_1), \boldsymbol{\phi}_T(t_2), \cdots, \boldsymbol{\phi}_T(t_n)) \overset{d}{\to} (\boldsymbol{\phi}(t_1), \boldsymbol{\phi}(t_2), \cdots, \boldsymbol{\phi}(t_n)). \tag{1.2}$$

We denote weak convergence by $\boldsymbol{\phi}_T \overset{w}{\to} \boldsymbol{\phi}$. If further $\boldsymbol{\phi} \in \mathsf{C}_{[0,1],\mathbb{R}^d}$, we have $\boldsymbol{\phi}_T \overset{w}{\to} \boldsymbol{\phi}$ if and only if $\mathbb{E}f(\boldsymbol{\phi}_T) \to \mathbb{E}f(\boldsymbol{\phi})$ for any bounded, continuous, $\mathscr{C}_{[0,1],\mathbb{R}^d}$-measurable functional $f : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$. Therefore, if $\boldsymbol{\phi}_T \overset{w}{\to} \boldsymbol{\phi} \in \mathsf{C}_{[0,1],\mathbb{R}^d}$, we then have $f(\boldsymbol{\phi}_T) \overset{d}{\to} f(\boldsymbol{\phi})$ for any $|||\cdot|||$-continuous functional f. The Slutsky theorem also holds here; for $\boldsymbol{\phi}_T^{(1)}, \boldsymbol{\phi}_T^{(2)} \in \mathsf{D}_{[0,1],\mathbb{R}^d}$ satisfying $\boldsymbol{\phi}_T^{(1)} \overset{w}{\to} \boldsymbol{\phi}$ and $d_{\mathrm{S}}(\boldsymbol{\phi}_T^{(2)}, \boldsymbol{\phi}_T^{(1)}) \overset{d}{\to} 0$, we have $\boldsymbol{\phi}_T^{(2)} \overset{w}{\to} \boldsymbol{\phi}$.

Chapter 2 Statistical Estimation and Online Inference via Local SGD

2.1 Introduction

Federated Learning (FL) is a distributed computing paradigm that allows for collaborative training of a global model using data held by remote clients, as described by McMahan, Moore, Ramage, Hampson, y Arcas [5]. FL ensures the protection of sensitive information contained in local datasets by only allowing cooperation with a central server, without sharing the data. However, limited data access, memory constraints, communication budget, and computation restrictions make traditional statistical estimation and inference methods inapplicable in the FL scenario[3, 6-7]. This chapter aims to address this challenge by studying how to perform statistical estimation and inference in the FL setting.

In a typical FL system, a pool of K clients each has a local dataset consisting of independently and identically distributed (i.i.d.) data from some unknown distribution $\mathscr{D}_k$. The central server faces a distributed optimization problem, where the goal is to minimize a user-specified loss function over all clients, that is

$$\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}) := \sum_{k=1}^{K} p_k \mathbb{E}_{\xi_k \sim \mathscr{D}_k} f_k(\boldsymbol{x}; \xi_k) \right\}, \tag{2.1}$$

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 data point from $\mathscr{D}_k$. The FL scenario poses a challenge due to the decentralized nature of data generation, resulting in statistical heterogeneity among local data distributions,[①] as well as the restrictive communication cost due to immense data volumes scattered across remote clients. To cope with these challenges, efficient algorithms have been proposed, with Local SGD being one of the simplest and most effective algorithms. Local SGD runs SGD independently in parallel on different clients and averages the sequences only once in a while to learn a shared global model via infrequent communication McMahan, Moore, Ramage, Hampson, y Arcas [5]. It has been shown to have superior performance in training efficiency and scalability[11], and converge fast in terms of communication[12-16]. The idea of lowering communication frequency for improving communication efficiency also motivates

[①] That is $\{\mathscr{D}_k\}_{k=1}^{K}$ are no longer necessarily identical.

Algorithm 1 Local SGD

```
Input: functions $\{f_k\}_{k=1}^K$, initial point $\boldsymbol{x}_0$, step size $\eta_m$, 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: $\hat{\boldsymbol{x}} = \frac{1}{T} \sum_{m=1}^T \bar{\boldsymbol{x}}_{t_m}$.
```

algorithms for other federated learning problems, including minimax problems[47-49] and distributed PCA[45-46].

From a statistical viewpoint, it is crucial to perform statistical inference in FL in order to infer properties of the underlying data distribution, quantify the uncertainty of the estimator, and monitor the algorithm's performance. However, it is still an open question of how to perform statistical inference in FL and adapt to its peculiarities. This paper aims to address statistical estimation and inference in FL via Local SGD, given its superior performance in training efficiency and scalability and representativeness in FL. Our goal is to obtain an efficient estimate of the optimal parameter value $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ and provide asymptotic confidence intervals for further inference, using only the Local SGD iterates $\{\boldsymbol{x}_{t_m}^k\}_{m \in [T], k \in [K]}$, obtained through communication at specific iterations. Here $[N] = \{1, 2, \ldots, N\}$ 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[37], 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[29-32].

2.1.1 Contribution

The following questions emerge which we study in the following:

1. how one constructs the estimator from Local SGD iterates $\{\boldsymbol{x}_{t_m}^k\}_{m \in [T], k \in [K]}$;

2. how local updates (or intermittent communication) affect its asymptotic behavior;
3. how one quantifies the variability and randomness of the estimator.

For the first question, Polyak, Juditsky [30], Ruppert [58] 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[59].

$$\hat{\boldsymbol{x}} = \frac{1}{T}\sum_{m=1}^{T} \bar{\boldsymbol{x}}_{t_m} \quad \text{where} \quad \bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^{K} p_k \boldsymbol{x}_{t_m}^k.$$

For the second question, under common assumptions, we show the proposed estimator $\hat{\boldsymbol{x}}$ 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[13, 15-16, 60]; 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.

For the third question of uncertainty quantification, we investigate two online inference methods for statistical inference. One is the plug-in method[61], which is available when we have an explicit formula for the covariance matrix of the estimator. The other, a.k.a., random scaling[62], borrows insights from time series regression in econometrics[63-64]. 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, Liao, Seo, Shin [62].① In particular, we pose a (2+δ) moment condition on gradient noises (see Assumption 2.3.2), while Lee, Liao, Seo, Shin [62] 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

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.

decomposition and a careful analysis on a non-asymptotic term with time-varying coefficients (see Lemma 2.5.7). 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.

2.1.2 Chapter Organization

The remainder of this chapter is organized as follows. In Section 2.2 we formulate our problem and introduce Local SGD. In Section 2.3 we explore the asymptotic properties for the averaged sequence of Local SGD. In Section 2.4 we introduce two online methods, namely the plug-in method and random scaling, to provide asymptotic confidence intervals and perform hypothesis tests. We provide the proof idea in Section 2.5 and review related work in Section 2.6. We illustrate the numerical performance of our methods in synthetic data in Section 2.7. We conclude our article in Section 2.8 with a discussion of our results and future research directions.

2.2 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 or data point it generates according to $\mathscr{D}_k$. A typical example of multi-round methods is the parallel stochastic gradient descent (P-SGD)[37] 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[36, 65-66]. 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[30]. 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[59]

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

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

2.2.1 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[10-11, 13, 15-16]. We now turn to Local SGD and summarize its details. We provide the formal version in Algorithm 1. Put simple, it obtains the solution estimate using the following recursive algorithm

$$\boldsymbol{x}_{t+1}^k = \begin{cases} \boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k) & \text{if } t+1 \notin \mathscr{I}, \\ \sum_{k=1}^K p_k \left[\boldsymbol{x}_t^k - \eta_t \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)\right] & \text{if } t+1 \in \mathscr{I}, \end{cases} \tag{2.2}$$

where η_t is the learning rate, ξ_t^k is an independent realization of $\mathscr{D}_k$, and $\mathscr{I}$ 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 $t+1 \in \mathscr{I}$, 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 $\mathscr{I}$ lead to different communication efficiency for Local SGD. If $\mathscr{I} = \{0, 1, 2, \cdots\}$, then Local SGD is reduced to P-SGD. A famous example in practice is constant communication interval[5], i.e., $\mathscr{I} = \{0, E, 2E, \cdots\}$ for a predefined integer $E(\geq 1)$, which reduces communication frequency from 1 to $1/E$. Local SGD differs from P-SGD if $\mathscr{I}$ 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.

2.3 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[30, 61, 67-68].

Assumption 2.3.1 (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 $\boldsymbol{x}, \boldsymbol{y}$,*

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

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

$$\sqrt{\mathbb{E}_{\xi_k}\|\nabla f_k(\boldsymbol{x};\xi_k) - \nabla f_k(\boldsymbol{y};\xi_k)\|^2} \le L\|\boldsymbol{x} - \boldsymbol{y}\| \tag{2.3}$$

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 $\boldsymbol{x}^\star$, i.e., there exist some $\delta_1 > 0$ and $L' > 0$ such that

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

Assumption 2.3.1 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 the Jensen's inequality implies that

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

The L-average smoothness follows if $\max_{\boldsymbol{x}} \mathbb{E}_{\xi_k}\|\nabla^2 f_k(\boldsymbol{x};\xi_k)\|^2 < \infty$ which holds for many statistical learning models such as linear and logistic regression.[1]

Define $\varepsilon_k(\boldsymbol{x}) = \nabla f_k(\boldsymbol{x};\xi_k) - \nabla f_k(\boldsymbol{x})$ as the gradient noise at $\nabla f_k(\boldsymbol{x})$, $\boldsymbol{S}_k = \mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x}^\star)\varepsilon_k(\boldsymbol{x}^\star)^\top)$, 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 2.3.2 (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 $k \in [K]$,*

$$\left\|\mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x})\varepsilon_k(\boldsymbol{x})^\top) - \boldsymbol{S}_k\right\| \le C\left[\|\boldsymbol{x} - \boldsymbol{x}^\star\| + \|\boldsymbol{x} - \boldsymbol{x}^\star\|^2\right]. \tag{2.4}$$

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 2.3.2 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 $\boldsymbol{x}^\star$ because $\boldsymbol{S} = \sum_{k=1}^K p_k^2\mathbb{E}_{\xi_k}(\varepsilon_k(\boldsymbol{x}^\star)\varepsilon_k(\boldsymbol{x}^\star)^\top) = \mathbb{E}_{\xi}(\varepsilon(\boldsymbol{x}^\star)\varepsilon(\boldsymbol{x}^\star)^\top)$ 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 $\|\boldsymbol{x} - \boldsymbol{x}^\star\|$. It implies $\left\|\mathbb{E}_{\xi}(\varepsilon(\boldsymbol{x})\varepsilon(\boldsymbol{x})^\top) - \boldsymbol{S}\right\| \le C'\left[\|\boldsymbol{x}-\boldsymbol{x}^\star\| + \|\boldsymbol{x}-\boldsymbol{x}^\star\|^2\right]$. 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, Liao, Seo, Shin [62].

Assumption 2.3.3 (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) $\sum_{m=1}^\infty \gamma_m =$

[1] This condition is also made by[67] to validate (2.4). See Lemma C.1 therein.

∞*, 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, Huang, Yang, Wang, Zhang [51] 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[30, 61, 67]. 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 2.3.4 (Slowly increasing communication intervals)**.** *The sequence* $\{E_m\}$ *satisfies*

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

(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 2.3.4 restricts the growth of $\{E_m\}$. Intuitively, if E_m increases too fast, each $\boldsymbol{x}_t^k$ might converge to their local minimizer $\boldsymbol{x}_k^\star$ rapidly before the next communication. Therefore, their average $\bar{\boldsymbol{x}}_t$ is asymptotically biased for $\boldsymbol{x}^\star$ with the bias $\sum_{k=1}^K p_k \boldsymbol{x}_k^\star - \boldsymbol{x}^\star$, 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 2.3.4 via investigating the relative difference of E_m and E_{m-1}.

Proposition 2.3.1. *Assume* $\{E_m\}$ *is non-decreasing. If* $\limsup_{m\to\infty} m(1 - \frac{E_{m-1}}{E_m}) < 1$*, then* (*ii*) *in Assumption 2.3.4 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 2.3.4 holds with* $\nu = \frac{1}{1-\rho^2}$.

Proofs of Proposition 2.3.1. To prove the proposition, we make two additional lemmas.

Lemma 2.3.1. *For any positive sequences* $\{a_n\}$ *and* $\{b_n\}$ *with* $\sum\limits_{n=1}^{T} b_n \to \infty$*, we have*

$$\limsup_{T\to\infty} \frac{\sum_{n=1}^{T} a_n}{\sum_{n=1}^{T} b_n} \leq \limsup_{T\to\infty} \frac{a_T}{b_T}. \tag{2.5}$$

Proof of Lemma 2.3.1. Without loss of generality, we assume the right hand side is finite, otherwise (2.5) follows obviously. We denote that $\limsup\limits_{T\to\infty} \frac{a_T}{b_T} = \lambda$ 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 $\sum\limits_{n=1}^{T} b_n \to \infty$, we have $\frac{\sum_{n=1}^{T} a_n}{\sum_{n=1}^{T} b_n} \leq \lambda + 2\varepsilon$. By the arbitrariness of ε, (2.5) follows. □

Lemma 2.3.2. *For any non-decreasing sequence* $\{E_m\}$ *satisfying* $\limsup\limits_{T\to\infty} T(1-\frac{E_{T-1}}{E_T}) < 1$*, we can find* $\delta > 0$ *such that*

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

Proof Lemma 2.3.2. In fact, we can choose any $\delta < 1 - \limsup\limits_{T\to\infty} T(1-\frac{E_{T-1}}{E_T})$. 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

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

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 Lemma 2.3.1 and (b) uses Lemma 2.3.1 and 2.3.2 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 A.2.3), 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. □

2.3.1 Asymptotic Results

According to the aforementioned regularity assumptions, the following asymptotic normality property of the averaged iterates generated by Local SGD is investigated in Theorem 2.3.1.

Theorem 2.3.1 (Asymptotic Normality). *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. Then* $\bar{\boldsymbol{x}}_{t_m}$ *converges to* $\boldsymbol{x}^\star$ *not only almost surely but also in* L_2 *convergence sense with rate* $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2 \lesssim \gamma_m$. *Moreover, if Assumption 2.3.4 holds additionally, the following asymptotic normality follows*

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

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

Theorem 2.3.1 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

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 $(\beta \in (0, 1))$		$\gamma m^{-(\alpha+\beta)}/E$	$\frac{1}{1-\beta^2}$	$(1+\beta)E^{-1}T^{-\beta}$

Table 2.1 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., $T/\sum_{m=0}^{T-1} E_m$.

order than the statistical error incurred by stochastic gradients[15-16]. 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 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[30]. When E_m varies, it is still possible to get communication saved and the asymptotic variance unchanged (i.e., $\nu = 1$) simultaneously (see Table 2.1). 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 2.3.4.

2.4 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 2.3.1), 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.

2.4.1 The plug-in Method

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

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

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

as long as each $\nabla^2 f_k(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}^k)$ is available. Though $\hat{\boldsymbol{G}}_T$ and $\hat{\boldsymbol{S}}_T$ are not unbiased for $\boldsymbol{G}$ and $\boldsymbol{S}$, their bias will converge to zero in probability due to $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^\star$ 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 $\hat{\boldsymbol{G}}_{m-1}$ to $\hat{\boldsymbol{G}}_m$ and $\hat{\boldsymbol{S}}_{m-1}$ to $\hat{\boldsymbol{S}}_m$. Therefore, they can be computed in an online manner without the need of storing all the data.

Assumption 2.4.1. *There are some constants $L'' > 0$ such that for any $k \in [K]$,*

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

Following[61], we make Assumption 2.4.1, which slightly strengthens the Hessian smoothness assumption in Assumption 2.3.1. Accordingly, we establish the consistency of the sample estimate $\hat{\boldsymbol{G}}_T$ and $\hat{\boldsymbol{S}}_T$ in the following theorem.

Theorem 2.4.1. *Under Assumptions 2.3.1, 2.3.2, 2.3.3 and 2.4.1, $\hat{\boldsymbol{G}}_T$ and $\hat{\boldsymbol{S}}_T$ converge to $\boldsymbol{G}$ and $\boldsymbol{S}$ in probability as $T \to \infty$. As a result of Slutsky's theorem, $\hat{\boldsymbol{G}}_T^{-1}\hat{\boldsymbol{S}}_T\hat{\boldsymbol{G}}_T^{-\top}$ is consistent to $\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$.*

Proof of Theorem 2.4.1. For simplicity, we denote

$$\nabla f(\boldsymbol{x};\xi_t) = \sum_{k=1}^{K} p_k \nabla f_k(\boldsymbol{x};\xi_t^k) \text{ and } \nabla^2 f(\boldsymbol{x};\xi_t) = \sum_{k=1}^{K} p_k \nabla^2 f_k(\boldsymbol{x};\xi_t^k),$$

where $\xi_t = \{\xi_t^k\}_{k\in[K]}$. We decompose $\hat{\boldsymbol{G}}_T - \boldsymbol{G}$ into the following terms:

$$\begin{aligned}\hat{\boldsymbol{G}}_T - \boldsymbol{G} &= \frac{1}{T}\sum_{m=1}^{T} \nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \boldsymbol{G} \\ &= \left[\frac{1}{T}\sum_{m=1}^{T} \nabla^2 f(\boldsymbol{x}^\star;\xi_{t_m}) - \boldsymbol{G}\right] + \frac{1}{T}\sum_{m=1}^{T}\left[\nabla^2 f(\bar{\boldsymbol{x}}_{t_m};\xi_{t_m}) - \nabla^2 f(\boldsymbol{x}^\star;\xi_{t_m})\right]. \qquad (2.6)\end{aligned}$$

The first term in (2.6) is asymptotically zero due to the strong law of large number. With Theorem 2.3.1, we have known that under the condition, $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\| \le \sqrt{\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2} \lesssim \sqrt{\gamma_m}$. Then the second term in (2.6) can be bounded via Assumption 3.2.3:

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

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

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

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

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

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

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 $\mathbb{E}\|\boldsymbol{C}_m\|^2 = \mathbb{E}\|\boldsymbol{C}_m\boldsymbol{C}_m^\top\| \le \operatorname{tr}(\mathbb{E}\boldsymbol{C}_m\boldsymbol{C}_m^\top) = \operatorname{tr}(\boldsymbol{S})$ and

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

Then, the second and third terms can be bounded via

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

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, $\hat{\boldsymbol{S}}_T$ converges to $\boldsymbol{S}$ in probability. □

Theorem 2.4.1 implies that $(\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top})_{jj}$ can be estimated by $\hat{\sigma}_{T,j}^2 = (\hat{\boldsymbol{G}}_T^{-1}\hat{\boldsymbol{S}}_T\hat{\boldsymbol{G}}_T^{-\top})_{jj}$ 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 $\bar{\boldsymbol{y}}_{T,j} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\nu}_T}{t_T}}\hat{\sigma}_{T,j}$ constructs an asymptotic exact confidence interval for the j-th coordinate of $\boldsymbol{x}^\star$. Here $\hat{\nu}_T$ is any sequence converging to ν.

Corollary 2.4.1. *Under the assumption of Theorem 2.4.1,*

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

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

We remark that using an estimate $\hat{\nu}_T$ 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 2.3.4. As a remedy, we use an estimate $\hat{\nu}_T = \frac{1}{T^2}(\sum_{m=1}^{T} E_m)(\sum_{m=1}^{T} E_m^{-1})$ 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.

2.4.2 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 2.4.2 (Functional CLT)**.** *Let Assumptions 2.3.1, 2.3.2, 2.3.3 and 2.3.4 hold, and define*

$$h(r, T) = \max\left\{ n \in \mathbb{Z}, n > 0 \,\middle|\, r\sum_{m=1}^{T} \frac{1}{E_m} \geq \sum_{m=1}^{n} \frac{1}{E_m} \right\} \quad \textit{for any fraction } r \in (0, 1]. \qquad (2.7)$$

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

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

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 $\boldsymbol{W}(\cdot)$ is the d-dim standard Brownian motion.

Theorem 2.4.2 has many implications. First, the result is stronger than Theorem 2.3.1 though under the same assumptions. By applying the continuous mapping theorem to Theorem 2.4.2 with $\psi : \mathsf{D}_{[0,1],\mathbb{R}^d} \mapsto \psi(1)$, we directly prove Theorem 2.3.1. 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

$$\frac{1}{\sqrt{T}} \sum_{m=1}^{\lfloor rT \rfloor} \left(\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star \right) \Rightarrow \sqrt{\frac{1}{E}} \boldsymbol{G}^{-1} \boldsymbol{S}^{1/2} \boldsymbol{W}(r).$$

When $E = 1$, it reduces to the single-machine result that is recently obtained by Lee, Liao, Seo, Shin [62]. 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 2.3.2) than previous Lee, Liao, Seo, Shin [62]. 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 2.5.7). See Section 2.5 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 $E = \infty$, $\boldsymbol{x}_{t_m}^k \equiv \boldsymbol{x}_k^\star$ 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 2.4.2 also embraces mildly increasing $\{E_m\}$ (see Table 2.1). Finally, there are some other FCLTs proved via a SDE argument on general stochastic process[69] or SGD with constant learning rates[70]. By contrast, we consider the particular Local SGD with decaying learning rates in the distributed context and the proof technique in Section 2.5 is from a discrete perspective.

With Theorem 2.4.2, 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 $\phi_T(r_m) = \frac{\sqrt{t_T}}{T} \sum_{n=1}^m (\bar{\boldsymbol{x}}_{t_n} - \boldsymbol{x}^\star)$. Note that $\phi_T(1) = \frac{\sqrt{t_T}}{T} \sum_{n=1}^T (\bar{\boldsymbol{x}}_{t_n} - \boldsymbol{x}^\star) = \sqrt{t_T}(\bar{\boldsymbol{y}}_T - \boldsymbol{x}^\star)$. 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 $\boldsymbol{x}^\star$. 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 2.4.2. *Under the same assumptions of Theorem 2.4.2 and assuming $g(r_m) \asymp \frac{m}{T}$ for some continuous function g on $[0,1]$, we have that*

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

This corollary follows immediately from Theorem 2.4.2 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 $\hat{\boldsymbol{V}}_T$ where

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

$\hat{\boldsymbol{V}}_T$ 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

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

Hence, to update $\hat{\boldsymbol{V}}_{m-1}$ to $\hat{\boldsymbol{V}}_m$ 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, $\hat{\boldsymbol{V}}_m = \frac{1}{m^2 s_m}\left(\boldsymbol{A}_m - \bar{\boldsymbol{y}}_m \boldsymbol{b}_m^\top - \boldsymbol{b}_m \bar{\boldsymbol{y}}_m^\top + q_m \bar{\boldsymbol{y}}_m \bar{\boldsymbol{y}}_m^\top\right)$. The formal formulation is presented in Algorithm 2.

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

Corollary 2.4.3. *Under the same conditions of Corollary 2.4.2, we have that*

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

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

$$W(1) \Big/ \left(\int_0^1 (W(r) - g(r)W(1))^2 dr \right)^{1/2} \tag{2.8}$$

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

If we only care about uncertainty of each coordinate $\boldsymbol{x}_j^\star$, for random scaling, we only need to store the diagonal entries of $\hat{\boldsymbol{V}}_T$ from Corollary 2.4.3. Both the storage and computation

Algorithm 2 Online Inference with Local SGD via Random Scaling

```
Input: functions $\{f_k\}_{k=1}^n$, initial point $\boldsymbol{x}_0$, step size $\eta_t$, communication set $\mathcal{J} = \{t_0, t_1, \cdots\}$.

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

    Return: $\bar{\boldsymbol{y}}_m$ and $\hat{\boldsymbol{V}}_m$.
end for
```

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 $\hat{\boldsymbol{G}}_T$ and $\hat{\boldsymbol{S}}_T$ and calculate the diagonal entries of $\hat{\boldsymbol{G}}_T^{-1}\hat{\boldsymbol{S}}_T\hat{\boldsymbol{G}}_T^{-\top}$.

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 (2.8) by $t^\star(\beta)$ and the corresponding critical value by $q_{\alpha,\beta} := \min\{t : \mathbb{P}(t^\star(\beta) \leq t) \geq 1-\alpha\}$. The limiting distribution $t^\star(\beta)$ is mixed normal and symmetric around zero. For easy reference, critical values of $t^\star(\beta)$ are computed via simulations and listed in Table 2.2. In particular, the Brownian motion $W(\cdot)$ 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[71]. 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, Vogelsang, Bunzel [63]. 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

$1-\alpha$ β	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

Table 2.2 Asymptotic critic values $q_{\alpha,\beta}$ of $t^\star(\beta)$ defined by $q_{\alpha,\beta} = \min\{t : \mathbb{P}(t^\star(\beta) \le t) \ge 1-\alpha\}$.

close to 6.747 computed in Abadir, Paruolo [72].

2.5 Proof Sketch of Theorem 2.4.2

We provide a short proof sketch for Theorem 2.4.2 to illustrate our proof technique in this section. As argued, Theorem 2.3.1 can be easily derived from Theorem 2.4.2 by the continuous mapping theorem. We follows the perturbed iterate framework that is derived by Mania, Pan, Papailiopoulos, Recht, Ramchandran, Jordan [73] and is widely used in recent works[10, 13-16, 51, 74]. 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 $m \ge 0$ and consider $t_m \le t < t_{m+1}$. 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

$$\bar{\boldsymbol{x}}_{t+1} = \bar{\boldsymbol{x}}_t - \eta_m \bar{\boldsymbol{g}}_t, \quad \text{where} \quad \bar{\boldsymbol{g}}_t = \sum_{k=1}^{K} p_k \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k). \tag{2.9}$$

Define $\boldsymbol{s}_m = \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star$ and recall that $E_m = t_{m+1} - t_m$ and $\gamma_m = \eta_m E_m$. Iterating (2.9) from $t = t_m$ to $t_{m+1} - 1$ gives

$$\boldsymbol{s}_{m+1} = \boldsymbol{s}_m - \eta_m \sum_{t=t_m}^{t_{m+1}-1} \bar{\boldsymbol{g}}_t = \boldsymbol{s}_m - \gamma_m \boldsymbol{v}_m \quad \text{where} \quad \boldsymbol{v}_m = \frac{1}{E_m} \sum_{t=t_m}^{t_{m+1}-1} \bar{\boldsymbol{g}}_t. \tag{2.10}$$

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

$$\begin{aligned}\boldsymbol{v}_m &= \boldsymbol{G}\boldsymbol{s}_m + \left(\nabla f(\bar{\boldsymbol{x}}_{t_m}) - \boldsymbol{G}\boldsymbol{s}_m\right) + (\boldsymbol{h}_m - \nabla f(\bar{\boldsymbol{x}}_{t_m})) + (\boldsymbol{v}_m - \boldsymbol{h}_m)\\ &:= \boldsymbol{G}\boldsymbol{s}_m + \boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m \end{aligned} \tag{2.11}$$

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

$$\boldsymbol{h}_m = \frac{1}{E_m}\sum_{t=t_m}^{t_{m+1}-1}\sum_{k=1}^{K} p_k \nabla f_k(\bar{\boldsymbol{x}}_{t_m}; \xi_t^k). \tag{2.12}$$

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 (2.10) and (2.11), we have

$$\boldsymbol{s}_{m+1} = (\boldsymbol{I} - \gamma_m \boldsymbol{G})\boldsymbol{s}_m - \gamma_m(\boldsymbol{r}_m + \boldsymbol{\varepsilon}_m + \boldsymbol{\delta}_m) := \boldsymbol{B}_m \boldsymbol{s}_m - \gamma_m \boldsymbol{U}_m, \tag{2.13}$$

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 (2.13) gives

$$\boldsymbol{s}_{m+1} = \left(\prod_{j=0}^{m}\boldsymbol{B}_j\right)\boldsymbol{s}_0 - \sum_{j=0}^{m}\left(\prod_{i=j+1}^{m}\boldsymbol{B}_i\right)\gamma_j \mathbf{U}_j. \tag{2.14}$$

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

$$h(r,T) = \max\left\{ n \in \mathbb{Z}_+ \middle| r\sum_{m=1}^{T}\frac{1}{E_m} \geq \sum_{m=1}^{n}\frac{1}{E_m}\right\}. \tag{2.15}$$

From Assumption 2.3.4, 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 (2.14) from $m = 0$ to $h(r,T)$ gives

$$\begin{aligned}\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\boldsymbol{s}_{m+1} &= \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\left[\left(\prod_{j=0}^{m}\boldsymbol{B}_j\right)\boldsymbol{s}_0 - \sum_{j=0}^{m}\left(\prod_{i=j+1}^{m}\boldsymbol{B}_i\right)\gamma_j\mathbf{U}_j\right]\\ &= \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)}\left(\prod_{j=0}^{m}\boldsymbol{B}_j\right)\boldsymbol{s}_0 - \frac{\sqrt{t_T}}{T}\sum_{j=0}^{h(r,T)}\sum_{m=j}^{h(r,T)}\left(\prod_{i=j+1}^{m}\boldsymbol{B}_i\right)\gamma_j\mathbf{U}_j. \end{aligned}\tag{2.16}$$

Lemma 2.5.1 (Lemma 1 in[30]). *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*

$$\boldsymbol{A}_j^n = \sum_{l=j}^{n}\left(\prod_{i=j+1}^{l}\boldsymbol{B}_i\right)\gamma_j. \tag{2.17}$$

Under Assumption 2.3.3, 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 (2.16) 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

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

where for simplicity we denote

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

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

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: $\sup_{r\in[0,1]}\|\mathscr{T}_0\|$, $\sup_{r\in[0,1]}\|\mathscr{T}_1\|$, $\sup_{r\in[0,1]}\|\mathscr{T}_2\|$, and $\sup_{r\in[0,1]}\|\mathscr{T}_4\|$ 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 2.5.2 will ease our proof. The following lemma proves the first statement of Theorem 2.3.1. The second statement of

Theorem 2.3.1 follows directly from Theorem 3.3.1 which we are going to prove via an argument of the continuous mapping theorem.

Lemma 2.5.2 (Almost surely and L_2 convergence)**.** *Under Assumptions 2.3.1, 3.2.2, and 2.3.3, $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^\star$ almost surely when m goes to infinity. In addition, there exists some $\tilde{C}_0 > 0$ such that*

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

Proof of Lemma 2.5.2. The proof can be found in Appendix A.1. □

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

Lemma 2.5.3. *Under Assumptions 2.3.1, 3.2.2, 2.3.3 and 2.3.4, the functional martingale CLT holds, namely, for any $r \in [0,1]$,*

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

where $h(r,T)$ is defined in (2.15) *and $\boldsymbol{W}(r)$ is the d-dimensional standard Brownian motion.*

Proof of Lemma 2.5.3. The proof can be found in Appendix A.2. □

Step two: Uniform negligibility of $\mathcal{T}_0$ Lemma 2.5.1 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 $T \to \infty$.

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

$$\begin{aligned}
\sup_{r\in[0,1]} \|\mathcal{T}_1\| &= \sup_{r\in[0,1]} \frac{\sqrt{t_T}}{T}\left\|\sum_{m=0}^{h(r,T)} \boldsymbol{A}_m^{h(r,T)}(\boldsymbol{r}_m + \boldsymbol{\delta}_m)\right\| \\
&\le \sup_{r\in[0,1]} \frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} C_0(\|\boldsymbol{r}_m\| + \|\boldsymbol{\delta}_m\|) \\
&= \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T} C_0(\|\boldsymbol{r}_m\| + \|\boldsymbol{\delta}_m\|),
\end{aligned}$$

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 2.5.4. *Under Assumptions 2.3.1, 3.2.2 and 2.3.3, we have that*

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

Proof of Lemma 2.5.4. The proof can be found in Appendix A.4. □

Lemma 2.5.5. *Under Assumptions 2.3.1, 3.2.2 and 2.3.3, we have that*

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

Proof of Lemma 2.5.5. The proof can be found in Appendix A.5. □

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

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

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

$$\mathbb{E}\left\|\boldsymbol{\varepsilon}_m\right\|^2 = \frac{1}{E_m}\mathbb{E}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^2 \le \frac{1}{E_m}\left(C_1 + C_2\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m}-\boldsymbol{x}^\star\|^2\right) \lesssim \frac{1}{E_m}, \tag{2.18}$$

where the first inequality is from Lemma A.1.1 with C_1, C_2 two universal constants defined therein and the second inequality uses Lemma 2.5.2. Using the last result, we have that

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

By Lemma 2.5.1, it follows that as $T \to \infty$,

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

Lemma 2.5.6 implies that $\mathbb{E}\sup_{r\in[0,1]}\|\mathscr{T}_2\|^2 = o(1)$.

Lemma 2.5.6. *Let $\{E_m\}$ be a positive-integer-valued sequence satisfying Assumption 2.3.4 and $\{a_{m,T}\}_{m\in[T],T\geq 1}$ 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*

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

Proof of Lemma 2.5.6. The proof can be found in Appendix A.6. □

Step five: Uniform negligibility of $\mathscr{T}_3$ It is subtle to handle $\mathscr{T}_3$ because its coefficient depends on r.

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

where the last inequality uses

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

Lemma 2.5.7 shows that $\sup_{r\in[0,1]}\|\mathscr{T}_3\| = o_{\mathbb{P}}(1)$.

Lemma 2.5.7. *Under Assumptions 3.2.2 and 2.3.4, it follows that*

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

Proof of Lemma 2.5.7. The proof can be found in Appendix A.7. □

Remark 2.5.1. *There is a more user-friendly version of Lemma 2.5.7 for a plug-and-play use. Define an auxiliary sequence* $\{\mathbf{Y}_m\}_{m\geq 0}$ *as following:* $\mathbf{Y}_0 = \mathbf{0}$ *and for* $m \geq 0$,

$$\mathbf{Y}_{m+1} = \boldsymbol{B}_m \mathbf{Y}_m + \gamma_m \boldsymbol{\varepsilon}_m = (\boldsymbol{I} - \gamma_m \boldsymbol{G})\mathbf{Y}_m + \gamma_m \boldsymbol{\varepsilon}_m. \tag{2.19}$$

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 2.5.7 is equivalent to

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

More formally, we have the following lemma which one can prove from Lemma 2.5.7.

Lemma 2.5.8. *If the martingale difference sequence* $\{\boldsymbol{\varepsilon}_m\}_{m\geq 0}$ *satisfies* $\sup_{m\geq 0} \mathbb{E}\|\boldsymbol{\varepsilon}_m\|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and Assumption 2.3.4 holds with* $E_m \equiv 1$*, for the sequence* $\{\mathbf{Y}_m\}_{m\geq 0}$ *defined in* (2.19) *with* $\boldsymbol{G}$ *positive definite, we have*

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

2.6 Related Work

Local SGD in Federated Learning Federated learning enables a large amount of edge computing devices to jointly learn a global model without data sharing[7]. The seminal paper[5] 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[51]. There has been a rapidly growing line of work concerning various aspects of FedAvg and its variants recently, including the effect of non-i.i.d. data[75], client sampling[76], decentralized optimization[12, 14], acceleration[77], composite optimization[78], and privacy[79]. 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[33, 36, 80-81].

The biggest difference is that Local SGD lowers the communication frequency, while others do not. This simple change improves communication efficiency greatly[11].

Analysis on Local SGD 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, Patel, Stich, Dai, Bullins, Mcmahan, Shamir, Srebro [15], Woodworth, Patel, Srebro [16] 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[82] that shows E_m can be set as large as $O(t_m^{1/2})$ for convergence. Later, Haddadpour, Kamani, Mahdavi, Cadambe [83] 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 by Li, Mou, Wainwright, Jordan [68] who instead proposed a new SGD variant to achieve both together. In addition, Karimireddy, Kale, Mohri, Reddi, Stich, Suresh [84], Liang, Shen, Liu, Pan, Chen, Cheng [85], Pathak, Wainwright [86], Zhang, Hong, Dhople, Yin, Liu [87] 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[88-89].

Recently, there are some works studying the efficiency of Local SGD via a continuous perspective. Viewing FL as a linearly constrained optimization problem, Liang, Han, Li, Zhang [90] modeled intermittent communication as a probabilistic projection and proposed a loop-less algorithm[①] to solve it. Using a novel jump diffusion approximation, they showed that the trajectories connecting those properly scaled last iterates weakly converge to the solution of specific stochastic differential equations (SDEs) that are driven by either a Brownian

[①] In the context of FL, this algorithm can be viewed as Local SGD where the periodic communication is replaced by a probabilistic communication.

motion or a Poisson process. Gu, Lyu, Huang, Arora [91] derived a SDE that captures the long-term behavior of Local SGD and provide a theoretical explanation why Local SGD generalizes better than SGD. Deng, Ma, Song, Zhang, Lin [92] proposed a federated averaging Langevin algorithm (FA-LD) for uncertainty quantification and mean predictions with distributed clients. They then study several factors including communication, accuracy, and privacy for this algorithm.

Statistical inference via SGD and its variants Statistical estimation and inference via SGD attracts great attention. Polyak, Juditsky [30], Ruppert [58] showed averaging iterates along the SGD trajectory has favorable statistical properties in the asymptotic setting, while Anastasiou, Balasubramanian, Erdogdu [31], Mou, Li, Wainwright, Bartlett, Jordan [32] supplemented it with a non-asymptotic analysis. Many papers recently developed iterative algorithms for constructing asymptotically valid confidence intervals[93]. Chen, Lee, Tong, Zhang, et al. [61] proposed a consistent plug-in estimator. However, the computation of the Hessian matrix of loss function is not always tractable. Then, Chen, Lee, Tong, Zhang, et al. [61] adapted the non-overlapping batch-means method[94] and obtained an offline consistent covariance estimator by using time-increasing batch sizes. Later on, Zhu, Chen, Wu [39] extended it to a fully online setting via a recursive counterpart using overlapping batches. In one latest work, Lee, Liao, Seo, Shin [62] 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 2.3.2), 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, Xu, Yang [95], Fang [96] proposed online bootstrap procedures for the estimation of confidence intervals via randomly perturbed SGD. Meanwhile, Su, Zhu [67], Li, Liu, Kyrillidis, Caramanis [97], Liang, Su [98] proposed variants of the SGD algorithm to facilitate inference in a non-asymptotic fashion.

2.7 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

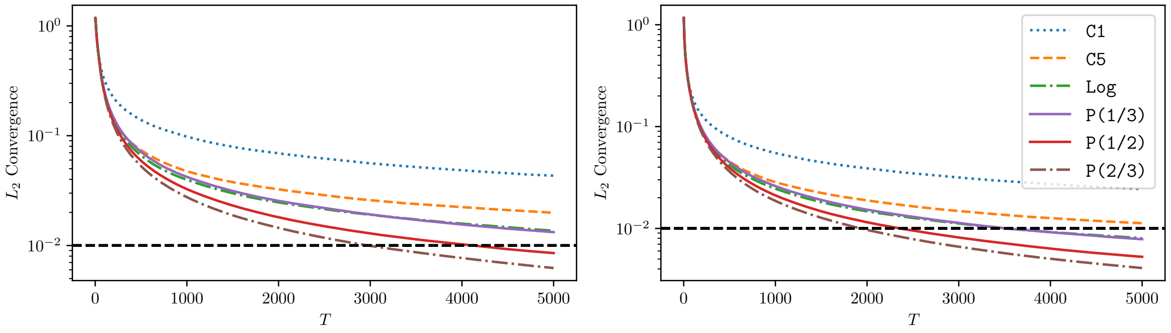


Figure 2.1 L_2 convergence $\|\bar{\boldsymbol{y}}_T - \boldsymbol{x}^\star\|$ in terms of communication T. Left: Results of linear regression. Right: Results of logistic regression. Black dashed line denotes the nominal coverage rate of 95%.

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, $b_t^k = (\boldsymbol{a}_t^k)^\top \boldsymbol{x}_k^\star + \varepsilon_t^k$ where the ε_t^k are i.i.d. according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\boldsymbol{x}_k^\star$ is the true local parameter which we also generate from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. In this case, the global parameter $\boldsymbol{x}^\star$ is the average of $\boldsymbol{x}_k^\star$'s.
- In logistic regression, $b_t^k \in \{0, 1\}$ is generated to be 1 with probability $\sigma((\boldsymbol{a}_t^k)^\top \boldsymbol{x}^\star)$ and 0 with probability $1 - \sigma((\boldsymbol{a}_t^k)^\top \boldsymbol{x}^\star)$. 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 $\boldsymbol{x}^\star$. Here $\boldsymbol{x}^\star$ is equi-spaced on the interval $[0, 1]$ following previous works[61-62].

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 $\boldsymbol{x}_1^\star$ 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 2.1, we find the faster E_m grows, the faster the L_2 convergence in terms of communication, which

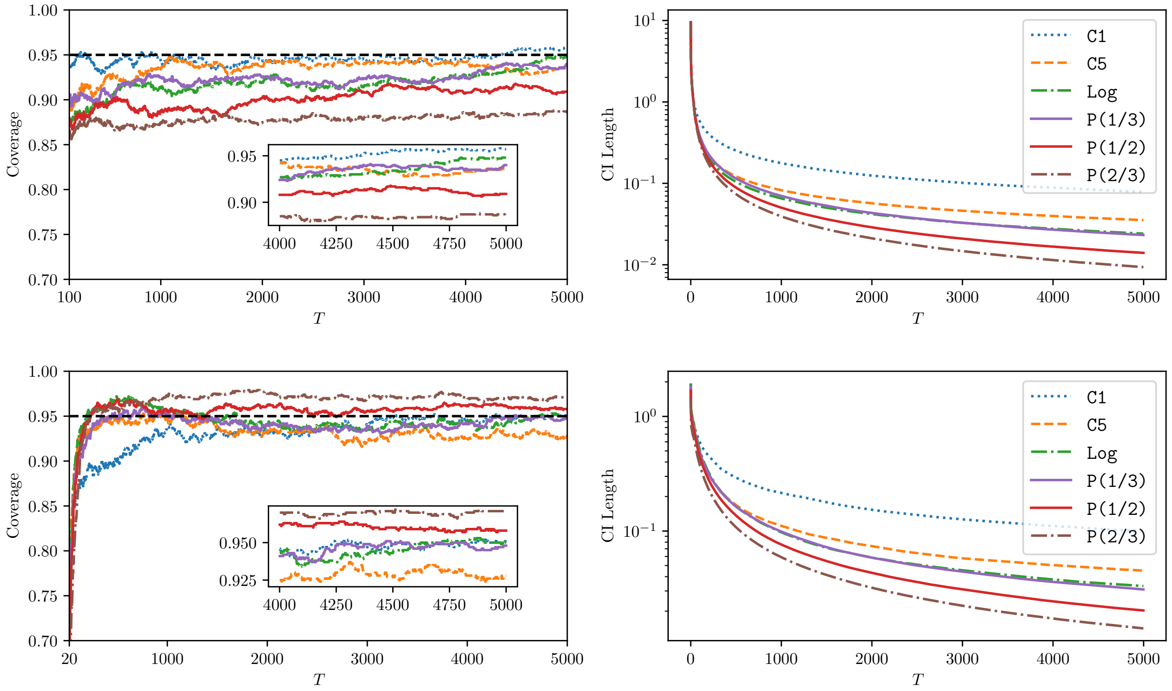


Figure 2.2 Comparison of the plug-in (the top row) and random scaling (the bottom row) in linear regression. Left: Empirical coverage rate against the number of communication. Black dashed line denotes the nominal coverage rate of 95%. Right: Length of confidence intervals.

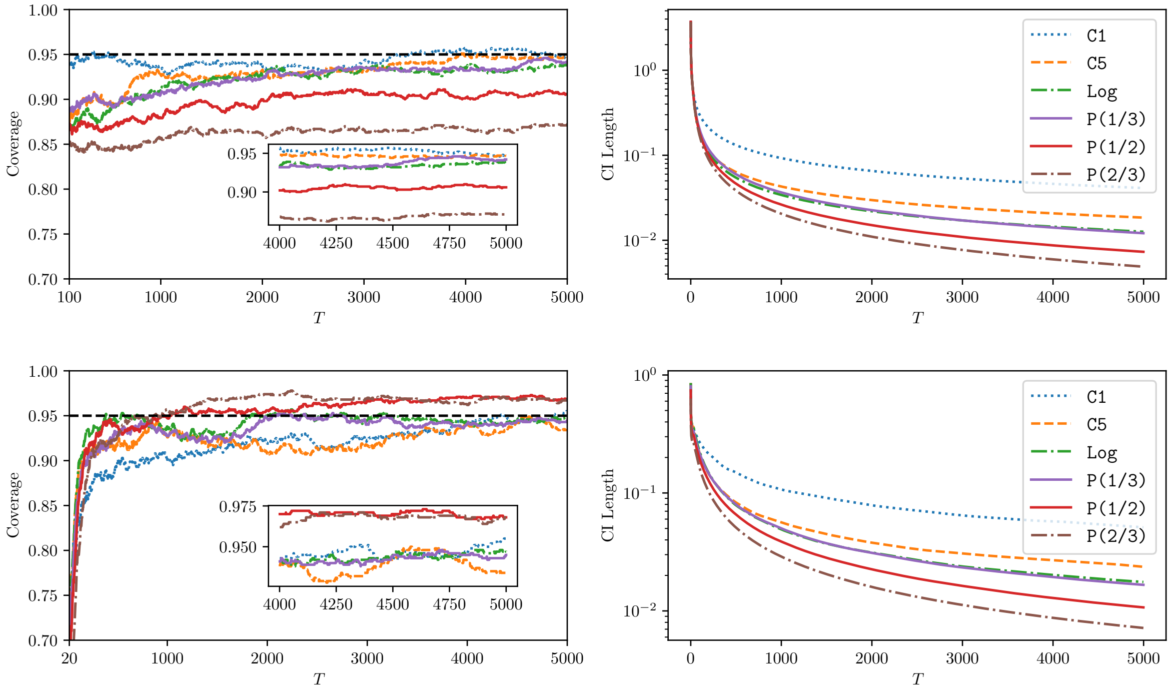


Figure 2.3 Comparison of the plug-in (the top row) and random scaling (the bottom row) estimators in logistic regression. Left: Empirical coverage rate against the number of communication. Black dashed line denotes the nominal coverage rate of 95%. Right: Length of confidence intervals.

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

is consistent with previous studies from optimization perspective[5, 11]. Figure 2.2 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 2.3. 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 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 2.3 shows the empirical performance of the six methods under linear

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)

Table 2.4 Simulation results of logistic regression with $d = 5$. The standard errors of coverage rates $\hat{p}$ are computed via $\sqrt{\hat{p}(1-\hat{p})/1000} \times 100\%$ and reported inside the parentheses.

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 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 3.3.1.

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 2.4. 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.

2.8 Conclusion

This chapter 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.

In literature, stochastic gradient descent (SGD) is considered an instance of the stochastic approximation (SA) method. SA is a more general framework that can be applied to a wider range of optimization problems. The aim of SA is to iteratively update an estimate of the root based on noisy or incomplete data, in order to find the root of a specific stationary equation. Q-Learning, introduced by Watkins [17], is another important example of SA and has recently gained popularity in reinforcement learning[9]. The stationary equation in SGD is simply $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, while in Q-Learning it is $\mathscr{T}\boldsymbol{Q} = \boldsymbol{Q}$, where $\mathscr{T}$ is the Bellman operator and $\boldsymbol{Q}$ is the vectorized Q-value function (which is the counterpart of $\boldsymbol{x}$ in this chapter). There are two main differences that distinguish Q-Learning from SGD. Firstly, $\mathscr{T}$ is typically not smooth, unlike the gradient ∇f, which has a continuous derivative. Secondly, the data used to evaluate $\mathscr{T}$ is typically generated along a Markov chain, whereas in SGD, the data is assumed to be independent. These differences imply that inference methods for SGD cannot be applied directly to Q-Learning. As a result, we are motivated to explore how to perform statistical inference for stochastic approximation using a single trajectory of Markov data.

Chapter 3 Online Statistical Inference for Nonlinear Stochastic Approximation with Markovian Data

3.1 Introduction

Stochastic approximation (SA) is a class of iterative methods for solving root-finding problems in which only noisy observations of objectives are available[99]. The aim is to find the root $\boldsymbol{g}(\boldsymbol{x}^\star) = \boldsymbol{0}$, where $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ is expressed as an integral over the data points ξ drawn from a distribution π on a Polish space Ξ:

$$\boldsymbol{g}(\boldsymbol{x}) := \int_{\Xi} \boldsymbol{H}(\boldsymbol{x}, \xi)\pi(d\xi) = \boldsymbol{0}. \tag{3.1}$$

When $\boldsymbol{g}$ is a linear function of $\boldsymbol{x}$, the method is referred to as linear SA, otherwise, it is referred to as nonlinear SA. A typical SA algorithm is given by the d-dimensional recursion:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}(\boldsymbol{x}_t, \xi_t), \tag{3.2}$$

in which $\{\eta_t\}_{t\geq 0}$ is the non-negative step-size sequence and $\{\xi_t\}_{t\geq 0}$ denotes the sequential data point. Over the past two decades, SA has gained significant attention, driven by applications in reinforcement learning and stochastic optimization[52, 100-102]. Despite the numerous SA methods developed and even the establishment of minimax optimal instance-dependent estimation bounds[21, 32, 100, 103-104], there is still a need for methods and theories that quantify estimation uncertainty and provide precise procedures for constructing confidence intervals.

Uncertainty quantification provides many benefits for practical sequential decision problems. By providing valid confidence intervals around predicted point estimates, it enables decision makers to make more informed and confident decisions with improved stability of recommendation quality[105]. In addition, confidence intervals provide a solid basis for risk management, allowing decision-makers to consider the potential consequences of various courses of action in the presence of uncertainty. This is particularly important in domains such as autonomous driving and personalized medicine where decisions have significant impacts.

In these applications, the sample-generating mechanism behind $\{\xi_t\}_{t\geq 0}$ is commonly modeled using a Markov chain. However, the introduction of Markovian data brings several challenges. Firstly, modeling arbitrary relationships between variables in Markovian data is difficult. Secondly, the distribution of each ξ_t changes over time and is unlikely to equal the desired distribution π, causing $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$ to become a biased estimate of $\boldsymbol{g}(\boldsymbol{x}_t)$ given the history. Finally, given a point estimate, there is currently no known method to estimate the asymptotic variance in the presence of Markovian data, though either a plug-in estimator or a batch-mean estimator could help under i.i.d. data[38-39].

To address these challenges, Ramprasad, Li, Yang, Wang, Sun, Cheng [43] proposed an online bootstrap method in linear SA with Markovian data. This method maintains multiple perturbed iterates $\{\boldsymbol{x}_t^b\}_{b\in[B]}$ from which confidence intervals can be constructed by estimating the asymptotic variance or quantiles from its empirical distribution over $b \in \boldsymbol{B}$. However, the per iteration update of each $\{\boldsymbol{x}_t^b\}_{b\in[B]}$ relies on the *multiple oracles* that evaluate the values of all $\{\boldsymbol{H}(\boldsymbol{x}_t^b, \xi_t)\}_{b\in[B]}$ at different parameters $\boldsymbol{x}_t^b$'s but with the same data point ξ_t. Due to limited control over real environments, multiple oracles typically are not feasible in scenarios where one-trajectory sampling is prevalent. Another limitation of the method is that it is heavily dependent on the linear nature of linear SA problems. As for the more general nonlinear SA (see Section 3.2.2 for examples), its effectiveness remains uncertain.

3.1.1 Contribution

In this study, we are motivated to inquire whether we can propose an efficient online inference method that does not require multiple oracles and can handle Markovian data in nonlinear SA. We provide an affirmative answer to this question.

Theoretical contribution In the absence of multiple oracles, we focus on utilizing the longitudinal dependence between consecutive iterates, rather than the crosswise dependence among perturbed iterates used in the online bootstrap method. To that end, we establish a functional central limit theorem (FCLT) in Theorem 3.3.1 that describes the asymptotic behavior of the partial-sum process $\boldsymbol{\phi}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}(\boldsymbol{x}_t - \boldsymbol{x}^*)$. Recall that $\mathsf{D}_{[0,1],\mathbb{R}^d} = \{\boldsymbol{\phi}$: càdlàg function $\boldsymbol{\phi}(r) \in \mathbb{R}^d, r \in [0,1]\}$ collects all d-dimensional functions that are right-continuous with left limits. As a random element in $\mathsf{D}_{[0,1],\mathbb{R}^d}$, this partial-sum process $\boldsymbol{\phi}_T$ weakly converges to a scaled Brownian motion $\boldsymbol{\psi} := \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}$ in the Skorohod topology, where $\boldsymbol{W}$ is the standard d-dimensional Brownian motion and $\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}$ is the unknown scale matrix. By the continuous mapping theorem, $f(\boldsymbol{\phi}_T)$ weakly converges to $f(\boldsymbol{\psi}) = f(\boldsymbol{W})$ for any continuous scale-invariant functional $f : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$ that satisfies $f(\boldsymbol{A}\boldsymbol{\phi}) = f(\boldsymbol{\phi})$ for any non-singular $\boldsymbol{A}$ and càdlàg process $\boldsymbol{\phi}$. $f(\boldsymbol{\phi}_T)$ is a measurable function of the observed data points $\{\xi_t\}_{t\in[T]}$ and the target parameter $\boldsymbol{x}^\star$, while $f(\boldsymbol{W})$ has a known distribution whose quantiles can be computed via stochastic simulation. It implies $f(\boldsymbol{\phi}_T)$ is an asymptotic pivotal statistic, from which an asymptotically valid confidence interval can be constructed.

To offer a comprehensive understanding of the FCLT, we further establish two additional results. The first result, outlined in Theorem 3.3.3, presents a semiparametric efficient lower bound that demonstrates the asymptotic variance of any regular asymptotic linear (RAL, see Definition 3.3.1) estimator $\boldsymbol{T}_n$, computed using the first n observed data points, is asymptotically lower bounded by $\frac{1}{n}\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$ in the sense that $\lim_{n\to\infty} n \cdot \mathbb{E}(\boldsymbol{T}_n - \boldsymbol{x}^\star)(\boldsymbol{T}_n - \boldsymbol{x}^\star)^\top \geq \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$. In Theorem 3.3.4, we find that for each fraction $r \in (0, 1]$, $\boldsymbol{\phi}_T(r)$ is the most efficient RAL estimator with its asymptotic variance matching the efficiency lower bound. This result answers an open question of efficiency in linear stochastic approximation raised by Ramprasad, Li, Yang, Wang, Sun, Cheng [43] and provides evidence of the statistical optimality of the partial-sum process $\boldsymbol{\phi}_T$ in terms of asymptotic variance.

The second result establishes a non-asymptotic upper bound on the functional weak convergence rate measured in the Lévy-Prokhorov distance, denoted by $d_{\mathrm{P}}(\cdot,\cdot)$. More specifically, Theorem 3.3.5 relates $d_{\mathrm{P}}(\boldsymbol{\theta}^\top\boldsymbol{\phi}_T, \boldsymbol{\theta}^\top\boldsymbol{\psi})$, the dissimilarity of the probability measures generated by the two càdlàg processes $\boldsymbol{\theta}^\top\boldsymbol{\phi}_T$ and $\boldsymbol{\theta}^\top\boldsymbol{\psi}$, to the iteration number T and the mixing time t_{mix} of the underlying Markov chain. Here, $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector with a unit dual norm satisfying $\|\boldsymbol{\theta}\|_* = 1$. To the best of our knowledge, it is the first non-asymptotic bound of functional weak convergence for the nonlinear iterative algorithm (3.2). It highlights the impact of several factors, including the underlying Markovian data, the degree of non-linearity, and the trade-off in step size parameter selection.

Methodological contribution The idea of applying a continuous scale-invariant functional to a partial-sum process, and constructing asymptotic pivotal statistic from it, has been adopted in the econometrics literature. This inference method is considered robust, as it not only eliminates the need to estimate the unknown scale matrix (e.g., $\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}$), but also works well for a wide range of linear series models with heteroskedasticity[63, 72]. Recently, Lee, Liao, Seo, Shin [62] extended this technique by proposing an online statistical inference method named "random scaling" for nonlinear SGD iterates. Following this line of research, subsequent works have further developed this approach for specific iterates $\{\boldsymbol{x}_t\}_{t\geq 0}$ under i.i.d. data[21, 50, 106]. In our work, we extend this concept to the more general setting of nonlinear SA with Markovian data. Additionally, we consider a family of adequate functionals f_m indexed by $m \in \mathbb{N}$. We study various aspects of confidence intervals generated by f_m, including their online computation efficiency, rejection probability, and confidence length. Finally, we evaluate the efficacy of different f_m's through numerical experiments.

Technical contribution The main difficulty in analysis is the establishment of the corresponding FCLT. If the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ is defined in a simpler manner, substantial research has been conducted to establish weak convergence for its partial-sum process in probability literature. The celebrated Donsker' s invariance principle concerns an i.i.d. sequence of $\{\boldsymbol{x}_t\}_{t\geq 0}$, while subsequent works have extended it to weakly dependent random variable[107], including stationary sequences[108] and martingale-like nonstationary structures[109]. However, the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ we consider here is defined recursively through (3.2). There are several reasons why the weakly dependent scenario from previous works is not applicable to our situation. Firstly, even if we assume $\{\xi_t\}_{t\geq 0}$ is sampled from a uniformly ergodic Markov chain with an arbitrary initialization distribution (in Assumption 3.2.4), the decaying step size $\{\eta_t\}_{t\geq 0}$ implies $\{\boldsymbol{x}_t\}_{t\geq 0}$ is not stationary.[①] Secondly, $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$ usually does not behave like a martingale difference and neither does each $\boldsymbol{x}_t$. Lastly, the conditions to control the degree of sequence dependence (e.g., various mixing conditions) in previous probability-oriented works are often difficult to verify in real-world applications. Therefore, we establish weak convergence from scratch by constructing a martingale-remainder decomposition. The idea behind this approach is to decompose the partial-sum of $\{\boldsymbol{x}_t\}_{t\geq 0}$ into the sum of partial-sums of martingale difference arrays and remainders, the latter vanishing asymptotically and uniformly under appropriate regularity conditions on $\boldsymbol{H}(\cdot,\cdot)$ (see Section 3.2.1). By doing so, we can further establish weak convergence rates by leveraging existing rates for martingale difference arrays[110] once those rates for the decomposed remainders are available.

We make several technical contributions along the martingale-remainder approach. The decomposition idea originates from the seminal work[30] for pointwise weak convergence and recently is extended to functional weak convergence by Li, Liang, Chang, Zhang [50], Lee, Liao, Seo, Shin [62] in the context of i.i.d. online convex stochastic optimization. However, for nonlinear SA with Markovian data, several difficulties arise. The Markovian noise precludes the direct use of martingale central limit theory and necessitates a Martingale approximation to decompose $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(\boldsymbol{x}_t)$.[②] Furthermore, the recursive update scheme (3.2), as well as the generality of nonlinear SA, bring difficulty to validate the uniform asymptotic

[①] Decaying the step size is necessary to obtain an asymptotically unbiased estimator for $\boldsymbol{x}^\star$.

[②] In the i.i.d. case, we have $\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)|\mathscr{F}_{t-1}] = \boldsymbol{g}(\boldsymbol{x}_t)$ as a result of the assumption $\xi_t \overset{i.i.d.}{\sim} \pi$. Therefore, $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(x_t)$ is a martingale difference adapted to $\mathscr{F}_t$ and thus the martingale central limit theory could apply. However, it is often not true for Markov cases.

vanishing of one particular remainder sequence.[③] To address the first issue, we utilize an existing martingale-residual-coboundary decomposition introduced by Liang [111]. In (3.15), it decomposes $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(\boldsymbol{x}_t)$ into the sum of a martingale term, a residual term, and a so-called coboundary term, with the last two terms having ignorable impacts on our target partial-sum process (see Lemma 3.4.1). For the second difficulty, we devise a novel technical Lemma 3.4.3 that drills down the particular recursion structure, from which functional weak convergence rates for the remainder sequence can be further derived. See Section 3.4.1 for more details.

3.1.2 Related Work

This study investigates the use of stochastic approximation algorithms for conducting statistical inference on Markovian data. Our findings have important implications for both reinforcement learning and stochastic optimization. So as to put our results into context, we provide more background on previous research in these areas.

Stochastic approximation on Markovian data The use of recursive stochastic procedures for root-finding problems dates back to the pioneering works of Robbins, Monro [99], as well as Kiefer, Wolfowitz [112], who established asymptotic convergence for derivative-free one-dimensional problems. Since then, stochastic approximation (SA) has been studied extensively, with a focus on its convergence and rate, parametric dependence, and qualitative properties. Except for the iterative analysis used to derive pointwise convergence, an ordinary differential equation (ODE) approach has been proposed and developed to track the trajectory behavior of SA procedures[52, 113-114]. The reader is referred to the monographs[52-54].

In many applications, the sample-generating mechanism behind $\{\xi_t\}_{t\geq 0}$ is modeled using an underlying Markov chain. Asymptotic convergence of SA algorithms with Markovian data can be established using either the ODE method[52] or the Poisson equation method[53]. Our paper falls into the second category with a specific interest in functional weak convergence. While other works assume that $\{\xi_t\}_{t\geq 0}$ comes from a state-dependent Markov chain[111, 115-116], it is beyond the scope of our paper. However, we believe that our analysis and methodology could be applied in this area with a stronger assumption on the existence of a solution to a Poisson equation. Our focus is on asymptotic analysis, but non-asymptotic estimation rates for SA algorithms with Markov data can be established if the Markov chain has a bounded mixing time. These rates have been studied in a general manner[103-104, 117], or in special cases, including two-timescale algorithms[118-120], gradient-based optimization[121-123], and estimation

[③] This troublesome process refers to the $\boldsymbol{\psi}_3$ in (3.20).

in autoregressive models[124]. Our contribution is orthogonal to these results, providing rates of functional weak convergence for the entire partial-sum process, in terms of the number of samples and mixing time, instead of moment convergence rates for point estimation.

Statistical inference via averaging stochastic approximation By averaging the iterates of SA procedures, it is known that one can obtain both an improved convergence rate and a Gaussian limiting behavior[30, 58, 125]. The form of this limiting distribution is optimal in the sense of local asymptotic minimax optimality[2, 59, 126]. Therefore, iterate averaging provides automatic optimal uncertainty quantification, laying the foundations of online statistical inference.

In the field of online stochastic optimization, several methods for statistical inference have been proposed. Zhu, Chen, Wu [39], Chen, Lee, Tong, Zhang, et al. [61] developed batch-means estimators for the limiting covariance matrix of asymptotic normality. Several variants of SGD-type algorithms have been proposed to either simplify inference procedures, such as implicit SGD[96, 127], resampling-based SGD[95, 97], and moment-adjusted variants[98], or address structured problems, such as online decision making[38] and sparse generalized linear models[128]. Other works establish Donsker-style generalization to the asymptotic normality to use trajectory information. Su, Zhu [67] took advantage of the asymptotic independence between the averaged iterates of different threads in a tree-structured scheme, while Lee, Liao, Seo, Shin [62] embraced the dependence between consecutive iterates and showed it was asymptotically negligible for a partial-sum process via a functional central limit theorem (FCLT). This partial-sum FCLT leads to a computationally efficient and memory-friendly online inference procedure that has proven effective in practice[62]. Subsequent work has extended this approach to areas such as federated learning[50], synchronous reinforcement learning[21], gradient-free optimization[129], and non-smooth regression[106].

A limitation in the above statistical inference methods and theories is that they assume i.i.d. data points $\{\xi_t\}_{t\geq 0}$. However, in asynchronous reinforcement learning (RL)[130-131], data is generated along a single Markov chain, precluding the use of stochastic optimization methods. Inspired by resampling-based inference methods in stochastic optimization, Bootstrap-based methods have been developed for linear policy evaluation tasks[40-43]. However, they are not suitable for nonlinear tasks, such as quantifying randomness in the optimal value function. The only available approach for this nonlinear task is considered by Shi, Zhang, Lu, Song [132] which uses sieve methods to approximate the Q-function and constructs two-scale confidence intervals, but it relies on batch updates from an offline dataset, making it computationally inefficient for sequential data scenarios. By contrast, we take the advantage of the partial-sum

FCLT and provide a fully online inference method for nonlinear stochastic approximation with Markovian data.

Trajectory behaviors in stochastic approximation To understand the asymptotic behaviors of SA trajectories, functional central limit theorems (FCLTs) are established to show weak convergence of a properly constructed process to a limit process. For discrete iterative algorithms, such as (3.2), the so-called ODE method introduced by Ljung [113] implies that, asymptotically, the noise effects average out or normally distributed once properly scaled, allowing the asymptotic behavior to be effectively determined by a mean ODE or an SDE (e.g., the Ornstein-Uhlenbeck equation). Following the spirit, works like[54, 102, 133] construct piecewise linear or piecewise constant interpolated processes by connecting properly centered and shifted iterates $\boldsymbol{x}_t - \boldsymbol{x}^\star$. These processes have a left-shifted initial point and a time-scale adjustment to approximate the mean ODE or SDE with increasing accuracy. Other SGD-type algorithms have used similar last-iterate interpolated processes. Chao, Cheng [134] studied weak convergence of the trajectories from generalized regularized dual averaging algorithms (gRDA) for online ℓ_1 penalized problems, while Negrea, Yang, Feng, Roy, Huggins [135] established a joint step-size–sample-size scaling asymptotic limit for stochastic gradient Langevin dynamics (SGLD). Our focus is the partial-sum process associated with $\{\boldsymbol{x}_t - \boldsymbol{x}^\star\}_{t\geq 0}$ in nonlinear SA with Markovian data, whereas most results focus on i.i.d. data[21, 50, 62, 106, 129, 136]. We chose not to utilize the ODE/SDE approach to demonstrate the FCLT, as it appears unsuitable for partial-sum processes on account of the imposed shifting initial point or the time scale.

Chapter organization The remainder of this chapter is organized as follows. We introduce the main assumptions and provide three examples of nonlinear SA in Section 3.2. We present the main asymptotic theoretical results in Section 3.3 and the online inference method in Section 3.5. We revisit the three examples and conduct numerical experiments in Section 3.6. We summarize our results and discuss future research directions in Section 3.7.

Notation Given a vector $\boldsymbol{v} = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we associate with it a norm $\|\cdot\|$ and denote its dual norm as $\|\cdot\|_*$, i.e., $\|\boldsymbol{v}\|_* = \sup_{\|\boldsymbol{u}\|\leq 1} |\langle \boldsymbol{v}, \boldsymbol{u}\rangle|$. We will denote $\|\boldsymbol{v}\|_1 := \sum_{i\in[d]} |v_i|$, $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i\in[d]} v_i^2}$, and $\|\boldsymbol{v}\|_\infty = \max_{i\in[d]} |v_i|$. By $\overset{d}{\to}$ we denote the point-wise weak convergence and by $\overset{p}{\to}$ we denote the convergence in probability. We use the standard Loewner order notation $\boldsymbol{A} \succeq \boldsymbol{0}$ if a matrix $\boldsymbol{A}$ is positive semi-definite. We denote $[n] := \{1, 2, \cdots, n\}$, the floor function $\lfloor\cdot\rfloor$ that is the greatest integer less than or equal to the input number, and ceiling function $\lceil \cdot \rceil$ that is the smallest integer greater than or equal to the input number. For two non-negative numbers a, b, we denote t $a \lesssim b$ if there exists a positive number C such that $a \leq Cb$ with C depending on parameters of no interest. Let $\mathscr{F}_t = \sigma(\{\xi_\tau\}_{0 \leq \tau \leq t})$ be the σ-fields generated by all randomness before iteration t and then $\boldsymbol{x}_t$ is $\mathscr{F}_{t-1}$-measurable.

3.2 Problem Setup and Motivating Examples

Recall from our earlier set-up that we are interested in providing confidence intervals for the root $\boldsymbol{x}^\star$ of (3.1) by using *only* the iterates $\{\boldsymbol{x}_t\}_{t \in [T]}$ produced through the iterative algorithm (3.2) with the data $\{\xi_t\}_{t \geq 0}$ sampled from a *single* Markov chain. We do not assume multiple evaluation oracles or access to derivatives of $\boldsymbol{H}(\boldsymbol{x}, \xi)$ with respect to $\boldsymbol{x}$. In summary, our target is a nonparametric inference method that is suitable for single-trajectory data.

3.2.1 Assumptions

We first introduce and discuss the assumptions that underlie our analysis.

Definition 3.2.1 (Hurwitz matrix or stable matrix). *We say $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a Hurwitz (or stable) matrix if* $\mathrm{Re}\lambda_i(\boldsymbol{A}) < 0$ *for $i \in [d]$. Here $\lambda_i(\cdot)$ denotes the i-th eigenvalue.*

Assumption 3.2.1 (Local linearity). *There exist constants $L_G, \lambda, \delta_G > 0$ and a Hurwitz $-\boldsymbol{G} \in \mathbb{R}^{d \times d}$ such that*

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{G}(\boldsymbol{x} - \boldsymbol{x}^\star)\| \leq L_G \|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 \textit{ for any } \|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \delta_G.$$

We consider a generally non-linear $\boldsymbol{g}$ which is locally linear at the neighborhood of the root $\boldsymbol{x}^\star$. We assume the linear coefficient $-\boldsymbol{G}$ is a Hurwitz matrix, a matrix whose every eigenvalue has a strictly positive real part. In engineering and stability theory, only using a Hurwitz matrix could make the linear system $\dot{\boldsymbol{x}} = -\boldsymbol{G}\boldsymbol{x}$ have a converging and stable solution. Such a kind of matrices have also been viewed as a generalization of positive definite matrices in the stochastic approximation literature[30, 32].

Assumption 3.2.2 (Regularized noises at the root). *There exist $p > 2$ and $\sigma > 0$ such that*

$$\sup_{t \geq 0} \sqrt[p]{\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^p} < \infty \textit{ and } \sup_{\xi \in \Xi} \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \leq \sigma,$$

where we denote $\mathscr{P}\boldsymbol{H}(\boldsymbol{x}, \xi) = \int_\Xi \boldsymbol{H}(\boldsymbol{x}, \xi')P(\xi, d\xi')$.

Assumption 3.2.2 adds moment conditions on the noise at the root $\boldsymbol{H}(\boldsymbol{x}^\star, \xi)$ (noting that $\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) = \boldsymbol{g}(\boldsymbol{x}^\star) = \boldsymbol{0}$). In particular, we assume that $\{\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\}_{t\geq 0}$ has uniformly bounded $p > 2$ moments so that we can use the martingale central limit theorem to establish asymptotic normality.

Assumption 3.2.3 (Lipschitz continuity). *Assume $\boldsymbol{H}(\cdot, \xi)$ is a uniformly averaged-L_H-Lipschitz continuous function in the sense that*

$$(\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)\|^p)^{\frac{1}{p}} \leq L_H\|\boldsymbol{x} - \boldsymbol{y}\| \textit{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \textit{ and } \xi \in \Xi, \tag{3.3}$$

where $\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)\|^p := \int_\Xi \|\boldsymbol{H}(\boldsymbol{x}, \xi') - \boldsymbol{H}(\boldsymbol{y}, \xi')\|^p P(\xi, d\xi')$ with p given in Assumption 3.2.2.

Assumption 3.2.3 provides a Lipschitz continuous condition that for any two parameters $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the L_p-norm of $\|\boldsymbol{H}(\boldsymbol{x}, \xi') - \boldsymbol{H}(\boldsymbol{y}, \xi')\|$ is uniformly and linearly bounded in terms of the difference $\|\boldsymbol{x} - \boldsymbol{y}\|$. Here ξ' denotes the data transited one step from the initial one $\xi \in \Xi$. This condition serves as a bridge to connect the running increment $\{\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\}_{t\geq 0}$ and the root-point-around noise $\{\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\}_{t\geq 0}$. In this way, once $\boldsymbol{x}_t$ converges and stays close to $\boldsymbol{x}^\star$, we would expect $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) \approx \boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)$, which together with Assumption 3.2.1 imply that the dynamic of the iterative procedure (3.2) is captured by a linear system up to a high-order approximation error.

Under the idealized i.i.d. setting (i.e., ξ_t is i.i.d. according to π), the condition in (3.3) simplifies to the L_H-averaged Lipschitz continuity, with $(\mathbb{E}_{\xi\sim\pi}\|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)\|^p)^{\frac{1}{p}} \leq L_H\|\boldsymbol{x} - \boldsymbol{y}\|$ and the σ defined in Assumption 3.2.2 is equal to zero. A sufficient condition for (3.3) is almost surely Lipschitz continuity, meaning that $\|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)\| \leq L_H|\boldsymbol{x} - \boldsymbol{y}|$ holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$. This type of condition is commonly used in machine learning, as demonstrated by the A2 condition in[103].

Assumption 3.2.4 (Uniformly ergodic Markov chain sampling). *We assume $\xi_t \in \Xi$ is generated from a time-homogeneous and uniformly ergodic Markov chain $\mathscr{M}$ with π the unique stationary distribution. Furthermore, there exist $\kappa \geq \mathrm{e}, \rho \in [0, 1)$ such that for any initial $\xi \in \Xi$,*

$$d_{\mathrm{TV}}(P^t(\xi, \cdot), \pi) \leq \frac{\kappa\rho^t}{2}, \tag{3.4}$$

where $d_{\mathrm{TV}}(\cdot, \cdot)$ denotes the total variation (TV) distance of probability measures and $P^t(\xi, \cdot)$ denotes the distribution of ξ_t with the initial state as $\xi_0 = \xi$.

A Markov process that satisfies Assumption 3.2.4 with the parameter (κ, ρ) is called a ρ-geometrical ergodic or uniformly ergodic process. Irreducible finite-state Markov chains are always uniformly ergodic. In general, if $\mathscr{M}$ satisfies a drift condition and a minorization condition, as stated in Proposition 5.1 in Andrieu, Moulines, Priouret [137] or Theorem 1.2 in Hairer, Mattingly [138], then (3.4) holds. In practical applications, when ξ_t is a concatenation of random variables taking values in a finite space, such as the current state in an MDP, and exogenous independent observation noises, such as independent stochastic rewards, (3.4) typically holds.

An important consequence from Assumption 3.2.4 is that for any bounded function $\boldsymbol{h} : \Xi \to \mathbb{R}^d$, $\mathscr{P}^t \boldsymbol{h}(\xi)$ would converge to $\mathbb{E}_{\xi\sim\pi}\boldsymbol{h}(\xi)$ exponentially fast uniformly over $\xi \in \Xi$.

Lemma 3.2.1. *Under Assumption 3.2.4, for any measurable uniformly bounded function $\boldsymbol{h} : \Xi \to \mathbb{R}^d$, we we have for any $t \geq 0$,*

$$\sup_{\xi\in\Xi} \|\mathscr{P}^t \boldsymbol{h}(\xi) - \mathbb{E}_{\xi\sim\mathscr{D}}\boldsymbol{h}(\xi)\| \leq \kappa\rho^t \cdot \sup_{\xi\in\Xi} \|\boldsymbol{h}(\xi) - \mathbb{E}_{\xi\sim\mathscr{D}}\boldsymbol{h}(\xi)\|. \tag{3.5}$$

Proof of Lemma 3.2.1. Define an auxiliary function $\boldsymbol{h}_0(\xi) := \boldsymbol{h}(\xi) - \mathbb{E}_{\xi\sim\pi}\boldsymbol{h}(\xi)$. Since $\boldsymbol{h}(\cdot)$ is uniformly bounded, so is $\boldsymbol{h}_0(\cdot)$. Furthermore, $\mathbb{E}_{\xi\sim\pi}\boldsymbol{h}_0(\xi) = \mathbf{0}$. By Strassen' s duality theorem, let $\xi_\infty \in \Xi$ denote the random variable with distribution π that satisfies $d_{\mathrm{TV}}(P^t(\xi,\cdot),\pi) = \mathbb{P}(\xi_t \neq \xi_\infty | \xi_0 = \xi)$. Then

$$\begin{aligned}
\|\mathscr{P}^t \boldsymbol{h}_0(\xi)\| &= \|\mathbb{E}[\boldsymbol{h}_0(\xi_t)|\xi_0 = \xi]\| = \|\mathbb{E}[\boldsymbol{h}_0(\xi_t) - \boldsymbol{h}_0(\xi_\infty)|\xi_0 = \xi]\| \\
&= \|\mathbb{E}[(\boldsymbol{h}_0(\xi_t) - \boldsymbol{h}_0(\xi_\infty)) \cdot 1_{\xi_t\neq\xi_\infty}|\xi_0 = \xi]\| \\
&\leq \sup_{\xi_t,\xi_\infty} \|\boldsymbol{h}_0(\xi_t) - \boldsymbol{h}_0(\xi_\infty)\| \cdot \mathbb{P}(\xi_t \neq \xi_\infty|\xi_0 = \xi) \\
&\leq 2\sup_{\xi\in\Xi} \|\boldsymbol{h}_0(\xi)\| \cdot d_{\mathrm{TV}}(\mathscr{P}^t(\xi,\cdot),\pi) \\
&\leq \kappa\rho^t \cdot \sup_{\xi\in\Xi} \|\boldsymbol{h}_0(\xi)\|.
\end{aligned}$$

□

We comment that we allow $\mathscr{M}$ to be initialized arbitrarily rather than from its stationary distribution π. One important quantity is the mixing time, that is, the time to approach stationarity (in terms of the TV distance) from the worst initial state. For the uniformly ergodic Markov chain above, the mixing time to accuracy ε is $t_{\mathrm{mix}}(\varepsilon) = \lceil \log_\rho \frac{2\varepsilon}{\kappa} \rceil$ so that $\frac{\kappa}{2}\rho^{t_{\mathrm{mix}}(\varepsilon)} \leq \varepsilon$.

With a special interest in the halving accuracy time,[1] we also define

$$t_{\text{mix}} = \begin{cases} 0 & \text{if } \rho = 0, \\ \frac{\ln \kappa}{1-\rho} & \text{if } \rho \in (0,1). \end{cases} \tag{3.6}$$

One can show that t_{mix} is an upper bound for $t_{\text{mix}}(0.5)$ by using the inequality $1 - \frac{1}{u} \le \ln u$ for all $u > 0$. If ξ_t's are i.i.d., then $\rho = 0$ and the t_{mix} is also zero. If ξ_t's follow from the Markov sampling, ρ becomes positive and t_{mix} goes to infinity when it approaches one.

Lemma 3.2.2. *Under Assumptions 3.2.2, 3.2.3, and 3.2.4, there exists a unique bivariate function $\boldsymbol{U}(\boldsymbol{x}, \xi)$ satisfies*

1. *It is the solution to the Poisson equation, where $\mathscr{P}\boldsymbol{U}(\boldsymbol{x}, \xi) := \int_{\Xi} \boldsymbol{U}(\boldsymbol{x}, \xi')P(\xi, d\xi')$,*

$$\boldsymbol{U}(\boldsymbol{x}, \xi) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}, \xi) = \boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{g}(\boldsymbol{x}). \tag{3.7}$$

2. *It is bounded in the sense that for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\xi \in \Xi$,*

$$\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}, \xi)\| \le \kappa t_{\text{mix}} \cdot \left(2L_H\|\boldsymbol{x} - \boldsymbol{x}^\star\| + \sigma\right).$$

3. *It is mean-zero in the sense that $\mathbb{E}_{\xi\sim\pi}\boldsymbol{U}(\boldsymbol{x}, \xi) = \boldsymbol{0}$ for any $\boldsymbol{x} \in \mathbb{R}^d$.*
4. *It is uniformly averaged Lipschitz continuous in the sense that*

$$(\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}, \xi) - \boldsymbol{U}(\boldsymbol{y}, \xi)\|^p)^{\frac{1}{p}} \le L_U\|\boldsymbol{x} - \boldsymbol{y}\| \textit{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \textit{ and } \xi \in \Xi,$$

where $L_U = \mathcal{O}(L_H(1 + \kappa t_{\text{mix}}))$ with $\mathcal{O}(\cdot)$ hiding universal constants. Here we denote $\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}, \xi) - \boldsymbol{U}(\boldsymbol{y}, \xi)\|^p := \int_{\Xi} \|\boldsymbol{U}(\boldsymbol{x}, \xi') - \boldsymbol{U}(\boldsymbol{y}, \xi')\|^p P(\xi, d\xi')$ with p given in Assumption 3.2.2.

Proof of Lemma 3.2.2. Define

$$\boldsymbol{U}(\boldsymbol{x}, \xi) := \sum_{t=0}^{\infty} \left(\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{g}(\boldsymbol{x})\right).$$

We first claim $\boldsymbol{U}(\boldsymbol{x}, \xi)$ is finite almost surely and thus well-defined. When setting $\boldsymbol{h}_{\boldsymbol{x}}(\xi) = \boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{g}(\boldsymbol{x})$, we know that $\mathscr{P}\boldsymbol{h}_{\boldsymbol{x}}(\xi) = \mathscr{P}\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{g}(\boldsymbol{x})$ is bounded by $\sigma_{\boldsymbol{x}} := 2L_H\|\boldsymbol{x} - \boldsymbol{x}^\star\| + \sigma$ uniformly over $\xi \in \Xi$ due to

$$\begin{aligned} \|\mathscr{P}\boldsymbol{h}_{\boldsymbol{x}}(\xi)\| &\le \mathscr{P}\|\boldsymbol{h}_{\boldsymbol{x}}(\xi) - \boldsymbol{h}_{\boldsymbol{x}^\star}(\xi)\| + \|\mathscr{P}\boldsymbol{h}_{\boldsymbol{x}^\star}(\xi)\| \\ &\le \mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| + \|\boldsymbol{g}(\boldsymbol{x})\| + \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \\ &\le 2L_H\|\boldsymbol{x} - \boldsymbol{x}^\star\| + \sigma = \sigma_{\boldsymbol{x}}, \end{aligned}$$

[1] Note that different accuracy ε's affect $t_{\text{mix}}(\varepsilon)$ only mildly. We take a concrete value of ε for notation simplicity.

where the last inequality uses Assumption 3.2.3 and Assumption 3.2.4. Therefore, under Assumption 3.2.4, we have $\mathbb{E}_{\xi\sim\mathscr{D}}\boldsymbol{h}_{\boldsymbol{x}}(\xi) = 0$ and $\|\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| \leq \kappa\rho^{t-1}\sigma_{\boldsymbol{x}}$ from Lemma 3.2.1. As a result, we have

$$\begin{aligned}
\|\boldsymbol{U}(\boldsymbol{x},\xi)\| &\leq \|\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| + \sum_{t=1}^{\infty}\|\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| \\
&\leq \|\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| + \kappa\sum_{t=0}^{\infty}\rho^t\sigma_{\boldsymbol{x}} \\
&\leq \|\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| + \frac{\kappa\sigma_{\boldsymbol{x}}}{1-\rho} < \infty.
\end{aligned}$$

Similarly, we can show

$$\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x},\xi)\| \leq \sum_{t=1}^{\infty}\|\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\| \leq \frac{\kappa\sigma_{\boldsymbol{x}}}{1-\rho} \leq \kappa\sigma_{\boldsymbol{x}}t_{\text{mix}},$$

which completes the proof for the second item. We then show $\boldsymbol{U}(\boldsymbol{x},\xi)$ is indeed a solution to (3.7) because

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{x},\xi)-\mathscr{P}\boldsymbol{U}(\boldsymbol{x},\xi) &= \sum_{t=0}^{\infty}\left(\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\right) - \mathscr{P}\sum_{t=0}^{\infty}\left(\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\right) \\
&= \sum_{t=0}^{\infty}\left(\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\right) - \sum_{t=1}^{\infty}\left(\mathscr{P}^t\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x})\right) \\
&= \boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{g}(\boldsymbol{x}).
\end{aligned}$$

It is also clear that $\mathbb{E}_{\xi\sim\pi}\boldsymbol{U}(\boldsymbol{x},\xi) = \boldsymbol{0}$ since π is the stationary distribution of $\mathscr{P}$ and the equation (3.1). If there exists another solution $\boldsymbol{U}'(\boldsymbol{x},\xi)$ to the same equation (3.7) and satisfying $\mathbb{E}_{\xi\sim\pi}\boldsymbol{U}'(\boldsymbol{x},\xi) = \boldsymbol{0}$ for any $\boldsymbol{x}\in\mathbb{R}^d$, then there exists a function $c(\boldsymbol{x})$ such that $\boldsymbol{U}'(\boldsymbol{x},\xi) = \boldsymbol{U}(\boldsymbol{x},\xi)+c(\boldsymbol{x})$ from Proposition 1.1 of Glynn, Meyn [139]. As a result, we have $c(\boldsymbol{x}) = \boldsymbol{0}$ for any $\boldsymbol{x}\in\mathbb{R}^d$, which implies the uniqueness of $\boldsymbol{U}(\boldsymbol{x},\xi)$.

Finally, for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$ and $\xi\in\Xi$, by Lemma 3.2.1,

$$\begin{aligned}
\|\boldsymbol{U}(\boldsymbol{x},\xi)-\boldsymbol{U}(\boldsymbol{y},\xi)\| &= \left\|\sum_{t=0}^{\infty}\left[\mathscr{P}^t\left(\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{H}(\boldsymbol{y},\xi)\right)-\left(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y})\right)\right]\right\| \\
&\leq \|(\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{H}(\boldsymbol{y},\xi))-(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y}))\| \\
&\qquad+\sum_{t=1}^{\infty}\left\|\mathscr{P}^t\left(\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{H}(\boldsymbol{y},\xi)\right)-\left(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y})\right)\right\| \\
&\leq \|\boldsymbol{H}(\boldsymbol{x},\xi)-\boldsymbol{H}(\boldsymbol{y},\xi)\| + \|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y})\|
\end{aligned}$$

$$+ \frac{\kappa}{1-\rho} \sup_{\xi\in\Xi} \|\mathscr{P}\left(\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\right) - (\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y}))\| .$$

By Jensen's inequality, it follows that for the p defined in Assumption 3.2.2,

$$\begin{aligned}\|\boldsymbol{U}(\boldsymbol{x},\xi) - \boldsymbol{U}(\boldsymbol{y},\xi)\|^p &\le 3^{p-1} \|\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\|^p + 3^{p-1} \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|^p \\ &\quad + \frac{3^{p-1}\kappa^p}{(1-\rho)^p} \sup_{\xi\in\Xi} \|\mathscr{P}\left(\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\right) - (\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y}))\|^p \\ &\le 3^{p-1} \|\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\|^p + \left(3^{p-1} + \frac{6^{p-1}\kappa^p}{(1-\rho)^p}\right) \cdot \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|^p \\ &\quad + \frac{6^{p-1}\kappa^p}{(1-\rho)^p} \sup_{\xi\in\Xi} \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x},\xi) - \mathscr{P}\boldsymbol{H}(\boldsymbol{y},\xi)\|^p .\end{aligned}$$

By Assumption 3.2.3, it follows that $\mathscr{P} \|\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\|^p \le L_H^p \|\boldsymbol{x}-\boldsymbol{y}\|^p$ uniformly for $\xi \in \Xi$. Notice that have that $\mathbb{E}_{\xi\sim\pi}\mathscr{P}(\cdot) = \mathbb{E}_{\xi\sim\pi}(\cdot)$ because π is the (unique) stationary distribution of $\mathscr{P}$. Therefore, by conditional Jensen's inequality,

$$\begin{aligned}\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|^p &= \left\|\mathbb{E}_{\xi\sim\pi}\mathscr{P}\left(\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\right)\right\|^p \\ &\le \mathbb{E}_{\xi\sim\pi}\mathscr{P} \|\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\|^p \le L_H^p \|\boldsymbol{x}-\boldsymbol{y}\|^p.\end{aligned}$$

Similarly, we also have that

$$\sup_{\xi\in\Xi} \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x},\xi) - \mathscr{P}\boldsymbol{H}(\boldsymbol{y},\xi)\|^p \le \sup_{\xi\in\Xi} \mathscr{P} \|\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{H}(\boldsymbol{y},\xi)\|^p \le L_H^p \|\boldsymbol{x}-\boldsymbol{y}\|^p.$$

Putting the pieces together, we conclude that there exists a constant $L_U = \mathcal{O}(L_H(1+\kappa t_{\mathrm{mix}}))$ such that

$$\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x},\xi) - \boldsymbol{U}(\boldsymbol{y},\xi)\|^p \le L_U^p \|\boldsymbol{x}-\boldsymbol{y}\|^p.$$

□

The existence of a unique solution to the Poisson equation (3.7) (denoted $\boldsymbol{U}(\boldsymbol{x},\xi)$) is a crucial result from Assumptions 3.2.2, 3.2.3, and 3.2.4. It can also be expressed as $(\mathscr{I} - \mathscr{P})^{-1}(\boldsymbol{H}(\boldsymbol{x},\xi) - \boldsymbol{g}(\boldsymbol{x}))$, where $\mathscr{I}$ is the identity mapping. Lemma 3.2.2 demonstrates that the operator $\mathscr{I} - \mathscr{P}$ is invertible on the mean-zero function class $\{\boldsymbol{h} \in (\mathbb{R}^d)^\Xi : \mathbb{E}_{\xi\sim\pi}\boldsymbol{h}(\xi) = \boldsymbol{0}\}$. Additionally, the function $\boldsymbol{U}(\boldsymbol{x},\xi)$ inherits all the properties of the bivariate function $\boldsymbol{H}(\boldsymbol{x},\xi)$ outlined in Assumptions 3.2.2 and 3.2.3. This function is important in determining the asymptotic variance and the semi-efficiency lower bound, which will be stated later.

Assumption 3.2.5 (Slowly decaying step size)**.** *Assume* (i) $0 < \eta_t \le 1, \eta_t \downarrow 0$*,* $\eta_t \log^2 t \to 0$ *and* $t\eta_t \uparrow \infty$ *as* $t \to \infty$*,* (ii) $\frac{\eta_{t-1}-\eta_t}{\eta_{t-1}} = o(\eta_{t-1})$ *for* $t \ge 1$*,* (iii) $\sum_{t=1}^{\infty} \frac{\log t}{\sqrt{t}}\eta_t < \infty$*, and* (iv)

$\frac{\sum_{t=0}^{T}\eta_t}{T\eta_T} \le C$ *for* $T \ge 1$.

We consider the step size that decays at a sufficiently slow rate satisfying Assumption 3.2.5. A classic example is the polynomial step size $\eta_t = \eta t^{-\alpha}$ with the scale $\eta > 0$ and $\alpha \in (0.5, 1)$.

Definition 3.2.2 ((L^p, b_t)-consistency[140])**.** *For a sequence* $\{\boldsymbol{x}_t\}_{t\ge 0} \subset \mathbb{R}^d$ *and a non-negative sequence* $\{b_t\}_{t\ge 0} \subset \mathbb{R}$, *we say* $\{\boldsymbol{x}_t\}_{t\ge 0}$ *to be* (L^p, b_t)*-consistency if there exists a positive constant* $C_p \ge 1$ *such that for any* $t \ge 0$,

$$\sqrt[p]{\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p} \le C_p b_t.$$

Assumption 3.2.6. *Assume* $\{\boldsymbol{x}_t\}_{t\ge 0}$ *satisfies the* $(L^2, (1+\log t)\sqrt{\eta_t})$*-consistency and* $\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p < \infty$ *with p given in Assumption 3.2.2.*

The final assumption, Assumption 3.2.6, concerns the (L^p, b_t)-consistency introduced by Gadat, Panloup [140]. This refers to the behavior of the SA update procedure in (3.2). It is important to note that (L^p, b_t)-consistency implies (L^q, b_t)-consistency for $0 < q \le p$ with $1 \le C_q \le C_p$, as per the Jensen inequality. In our case, we only require $(L^2, (1+\log t)\sqrt{\eta_t})$-consistency, a weaker condition than the original work that assumes $(L^4, \sqrt{\eta_t})$-consistency[140].

3.2.2 Examples of Nonlinear Stochastic Approximation

We now present some examples of nonlinear SA which we would revisit in the numerical experiments.

3.2.2.1 Stochastic Gradient Descent

The most celebrated example is stochastic gradient descent (SGD) that is originally introduced by Robbins, Monro [99]. Due to its simplicity and efficiency, SGD probably becomes the most powerful method for solving optimization problems in machine learning. The standard task is to minimize an (unknown) objective function $F : \mathbb{R}^d \to \mathbb{R}$ in the form $F(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\pi}F(\boldsymbol{x}, \xi)$. We have access to the noisy samples of the gradient $\nabla F(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\pi}\nabla F(\boldsymbol{x}, \xi)$ where ξ is the observed data. When having complete control over data collection (e.g. the case of offline training), we can assume each data $\xi_t \overset{i.i.d.}{\sim} \pi$ for granted. In the streaming data setting, it is more practical to assume the data $\{\xi_t\}_{t\ge 0}$ sampled from a Markov chain with π the unique stationary distribution (see Assumption 3.2.2). In this case, during the t-th gradient oracle, we input a parameter $\boldsymbol{x}_t$ and observe a stochastic gradient vector $\nabla F(\boldsymbol{x}_t, \xi_t)$ as the sample of $\nabla F(\boldsymbol{x}_t)$. We then perform SGD to update $\boldsymbol{x}_t$ via

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}(\boldsymbol{x}_t, \xi_t) \text{ with } \boldsymbol{H}(\boldsymbol{x}_t, \xi_t) = \nabla F(\boldsymbol{x}_t, \xi_t).$$

For convex F's, one can find that its minimizer $\boldsymbol{x}^\star$ is exactly the root of its gradient function, i.e., $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = \{\boldsymbol{x} \in \mathbb{R}^d : \nabla F(\boldsymbol{x}) = \boldsymbol{0}\}$. When $\nabla^2 F(\boldsymbol{x})$ further satisfies a local continuity condition around the root $\boldsymbol{x}^\star$ where $\|\nabla^2 F(\boldsymbol{x}) - \nabla^2 F(\boldsymbol{x}^\star)\| \leq 2L_G \|\boldsymbol{x} - \boldsymbol{x}^\star\|$ for any $\|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \delta_G$, Assumption 3.2.1 is satisfied. This local continuity condition is used by Li, Liang, Chang, Zhang [50], Su, Zhu [67], Chen, Lai, Li, Zhang [129] to ensure local linearity in their applications. To ensure (3.3), a sufficient condition is almost surely Lipschitz continuity that $\|\nabla F(\boldsymbol{x}, \xi) - \nabla F(\boldsymbol{y}, \xi)\| \leq L_H \|\boldsymbol{x} - \boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$. Assumption 3.2.2 relies on how the data ξ_t interacts with the gradients $\nabla F(\boldsymbol{x}^\star, \xi_t)$, while Assumption 3.2.4 purely depends on the data generation mechanism, both of which require a case-by-case discussion.

In the following, we provide two more concrete examples of $\nabla F(\boldsymbol{x}, \xi_t)$ and see how they satisfy the assumptions we imposed.

- The first example is linear regression with autoregressive noises. We receive data $\xi_t = (\boldsymbol{a}_t, y_t)$ where $y_t = \langle \boldsymbol{a}_t, \boldsymbol{x}^\star \rangle + \zeta_t$. Here $m : \mathbb{R} \to \mathbb{R}$ is a transformation function, the covariate $\boldsymbol{a}_t \overset{i.i.d.}{\sim} \pi_{\boldsymbol{a}}$, and each infused noise ζ_t is sampled from an autoregressive model with π_ζ the stationary distribution. The stationary distribution π corresponds to the joint distribution of $(\boldsymbol{a}, y)$ where $\boldsymbol{a} \sim \pi_{\boldsymbol{a}}$ and $y = \langle \boldsymbol{a}, \boldsymbol{x}^\star \rangle + \zeta$ with $\zeta \sim \pi_\zeta$ independent of $\boldsymbol{a}$. We use the squared loss $F(\boldsymbol{x}, \xi_t) = \frac{1}{2}(y_t - \langle \boldsymbol{a}_t, \boldsymbol{x} \rangle)^2$ and thus $\nabla F(\boldsymbol{x}, \xi_t) = (\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle - y_t)\boldsymbol{a}_t$. One can show that Assumption 3.2.1 holds with $\boldsymbol{G} = \mathbb{E}_{a \sim \pi_{\boldsymbol{a}}} \boldsymbol{a}\boldsymbol{a}^\top$ and $(\delta_G, L_G) = (\infty, 0)$. Once $\mathbb{E}_{\boldsymbol{a} \sim \pi_{\boldsymbol{a}}} \|\boldsymbol{a}\|^p < \infty$, Assumptions 3.2.2 and 3.2.3 follow.
- The second example is generalized linear model with Markovian data. In the observed data $\xi_t = (\boldsymbol{a}_t, y_t)$, the covariate $\{\boldsymbol{a}_t\}_{t\geq 0}$ is generated according to an autoregressive model with $\pi_{\boldsymbol{a}}$ its stationary distribution and $\{y_t\}_{t\geq 0}$ is generated from the canonical generalized linear model $p_y(y|\boldsymbol{a}_t) \propto \exp\left(\theta_t y - b(\theta_t)\right)$ with $\theta_t = \langle \boldsymbol{a}_t, \boldsymbol{x}^\star \rangle$. The stationary distribution π is $\pi_{\boldsymbol{a}}(d\boldsymbol{a}) \times p_y(dy|\boldsymbol{a})$. We use the negative log-likelihood loss $F(\boldsymbol{x}_t, \xi_t) = b(\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle) - \langle \boldsymbol{a}_t, \boldsymbol{x} \rangle y_t$ and thus $\nabla F(\boldsymbol{x}, \xi_t) = (b'(\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle) - y_t)\boldsymbol{a}_t$ where b' is the derivative of b. Standard choices of b include the identity map for linear regression and the logistic function for logistic regression. Assumption 3.2.1 is satisfied with $\boldsymbol{G} = \nabla^2 F(\boldsymbol{x}^\star)$ and some finite (δ_G, L_G) if we assume b'' is non-negative and uniformly bounded with $\sup_{t\geq 0} \mathbb{E}\|\boldsymbol{a}_t\|^2 < \infty$. Assumptions 3.2.2 and 3.2.3 are satisfied if we further assume b' is Lipschitz continuous and uniformly bounded together with $\sup_{t\geq 0} \mathbb{E}\|\boldsymbol{a}_t\|^p < \infty$ for $p > 2$.

In these cases, the uniform ergodicity of ξ_t in Assumption 3.2.4 is reduced to that of either ζ_t or $\boldsymbol{a}_t$, both autoregressive processes. Uniform ergodicity has already been established for a wide range of first-order linear autoregressive (a.k.a. AR(1)) models[141].

3.2.2.2 Asynchronous Q-Learning

Reinforcement learning algorithms are often studied in terms of the Markov decision process (MDP) with a finite state space $\mathscr{S}$ and action space $\mathscr{A}$[9]. An MDP contains a collection of probability transition kernels $\{\boldsymbol{P}(\cdot|s,a)\}_{(s,a)\in\mathscr{S}\times\mathscr{A}} \subseteq \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|\times|\mathscr{S}|}$ where the transition kernel $\boldsymbol{P}(s'|s,a)$ denotes the probability of transiting to s' when action $a \in \mathscr{A}$ is taken at the state $s \in \mathscr{S}$. The MDP is also equipped with a random reward function $\boldsymbol{R} \in \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|}$ and $\boldsymbol{R}(s,a)$ corresponds to the immediate reward collected in state $s \in \mathscr{S}$ upon performing the action $a \in \mathscr{A}$. We denote $\boldsymbol{r} = \mathbb{E}\boldsymbol{R}$ by the expected reward function. A policy $\pi : \mathscr{S} \to \Delta(\mathscr{A})$ is a mapping from the state space $\mathscr{S}$ to the simplex of action space $\mathscr{A}$ (denoted $\Delta(\mathscr{A})$). In discounted MDPs, a common objective is to maximize the expected long-term reward. For a given policy π, the expected long-term reward is measured by its Q-function $\boldsymbol{Q}^\pi$ defined as

$$\boldsymbol{Q}^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \boldsymbol{R}_t(s_t,a_t)\middle| s_0 = s, a_0 = a, \pi\right],$$

where the trajectory is generated according to $a_t \sim \pi(s_t), s_{t+1} \sim \boldsymbol{P}(\cdot|s_t,a_t)$, and $\boldsymbol{R}_t(s_t,a_t) \sim \boldsymbol{R}(s_t,a_t)$. Classic results show that the optimal Q-function $\boldsymbol{Q}^\star(s,a) := \max_\pi \boldsymbol{Q}^\pi(s,a)$ is uniquely determined by the fixed point of the Bellman equation $\boldsymbol{Q}^\star = \boldsymbol{r} + \gamma\boldsymbol{P}\mathscr{T}\boldsymbol{Q}^\star$ where $\mathscr{T} : \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|} \to \mathbb{R}^{|\mathscr{S}|}$ is a blockwise max operator defined by $(\mathscr{T}\boldsymbol{Q})(s) := \sup_{a\in\mathscr{A}} \boldsymbol{Q}(s,a)$ for any $s \in \mathscr{S}$.

Q-Learning is perhaps the most popular model-free approach to seek the optimal value function[17]. In the so-called asynchronous RL, a generative data simulator is not available and data access is limited to the Markov chain introduced by a given behavior policy π_{b}[130]. At iteration t, the agent performs action $a_t \sim \pi_{\mathrm{b}}(s_t)$ from the current state s_t, then receives a random reward $\boldsymbol{R}_t(s_t,a_t)$, and transits to the next state $s_{t+1} \sim \boldsymbol{P}(\cdot|s_t,a_t)$. With the data $\xi_t = (s_t, a_t, \boldsymbol{R}_t(s_t,a_t), s_{t+1})$, Q-Learning updates an estimate $\boldsymbol{Q}_t$ for $\boldsymbol{Q}^\star$ via

$$\boldsymbol{Q}_{t+1}(s,a) = \begin{cases} \boldsymbol{Q}_t(s,a) & \text{if } (s,a) \neq (s_t,a_t), \\ (1-\eta_t)\cdot\boldsymbol{Q}_t(s,a) + \eta_t\left(\boldsymbol{R}_t(s_t,a_t) + \gamma\sup_{a\in\mathscr{A}}\boldsymbol{Q}_t(s_{t+1},a)\right) & \text{if } (s,a) = (s_t,a_t). \end{cases}$$

Denote by $\boldsymbol{R}_t \in \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|}$ the one-hot vector with only the (s_t,a_t)-th entry non-zero with value $\boldsymbol{R}_t(s_t,a_t)$ and by $\boldsymbol{P}_t \in \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|\times|\mathscr{S}|}$ the sparse matrix with only the (s_t,a_t,s_{t+1})-th entry nonzero with value one. Then one can rewrite (3.8) in a matrix form

$$\boldsymbol{Q}_{t+1} = \boldsymbol{Q}_t - \eta_t \boldsymbol{H}(\boldsymbol{Q}_t, \xi_t) \quad \text{with} \quad \boldsymbol{H}(\boldsymbol{Q}_t, \xi_t) = \boldsymbol{I}_t(\boldsymbol{Q}_t - \gamma \boldsymbol{P}_t \mathscr{T} \boldsymbol{Q}_t - \boldsymbol{R}_t), \tag{3.8}$$

where $\boldsymbol{I}_t \in \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|}$ is a sparse diagonal matrix with only the (s_t, a_t)-th entry equal to one and $\boldsymbol{Q}_t \in \mathbb{R}^{|\mathscr{S}\times\mathscr{A}|}$ is the vectorized Q-value function. We slightly abuse the notation and always use $\boldsymbol{R}_t, \boldsymbol{P}_t$ to denote the dense observation whose most coordinates are not accessible but filtered out by the sparse matrix $\boldsymbol{I}_t$. By contrast, a data generator is available in synchronous RL that produces independent rewards and next states for all state-action pairs so that $\boldsymbol{I}_t$ is always an identity matrix[56].

Proposition 3.2.1. *Assume that* (i) *the MDP introduced by* π_{b} *is irreducible,* (ii) *the optimal policy is unique and denoted by* $\pi^\star$*, and* (iii) $\{\boldsymbol{R}_t(s,a)\}_{(s,a)\in\mathscr{S}\times\mathscr{A}}$ *is independent on* $\{s_t\}_{t\ge 0}$*, and* $\sup_{(s,a)\in\mathscr{S}\times\mathscr{A}} \mathbb{E}|\boldsymbol{R}(s,a)|^p < \infty$*. Then the iterates* $\{\boldsymbol{Q}_t\}_{t\ge0}$ *in* (3.8) *satisfies Assumptions 3.2.1-3.2.4.*

The consequences of the assumptions in Proposition 3.2.1 are as follows. Firstly, under the assumptions, the stationary distribution of ξ_t is given by $d_{\pi_{\mathrm{b}}}(ds)\pi_{\mathrm{b}}(da|s)p_r(dr|s,a)\boldsymbol{P}(ds'|s,a)$, where $d_{\pi_{\mathrm{b}}}(\cdot)$ is the state stationary distribution of the MDP determined by π_{b} and $p_r(\cdot|s,a)$ is the probability density function of $\mathbf{R}(s,a)$. As a result, $\boldsymbol{g}(\boldsymbol{Q}) = \boldsymbol{D}(\boldsymbol{Q} - \gamma \boldsymbol{P}\mathscr{T}\boldsymbol{Q} - \boldsymbol{r})$ with $\boldsymbol{D} = \mathrm{diag}(\{d_{\pi_{\mathrm{b}}}(s)\pi_{\mathrm{b}}(a|s)\}_{(s,a)})$ is a square diagonal matrix with order $|\mathscr{S}\times\mathscr{A}|$. Using the ℓ_∞ norm, Li, Yang, Jiadong, Zhang, Jordan [21] showed Assumption 3.2.1 holds for Q-Learning with $(\delta_G, L_G) = (\infty, \frac{L}{\Delta})$ if the optimal policy $\pi^\star$ is unique, where Δ is the optimality gap defined by $\Delta := \min_s \min_{a\neq\pi^\star(s)} |V^\star(s) - Q^\star(s,a)|$. In this case, when $\|\boldsymbol{Q} - \boldsymbol{Q}^\star\| \lesssim \Delta$, $\mathscr{T}\boldsymbol{Q} = \boldsymbol{\Pi}^{\pi^\star}\boldsymbol{Q}$ behaviors like a linear operator where $\boldsymbol{\Pi}^\pi \in \mathbb{R}^{|\mathscr{S}|\times|\mathscr{S}\times\mathscr{A}|}$ is a projection matrix associated with a given policy π defined by $\boldsymbol{\Pi}^\pi := \mathrm{diag}(\{\pi(\cdot|s)^\top\}_{s\in\mathscr{S}})$. Therefore, the local linearity matrix is $\boldsymbol{G} = \boldsymbol{D}(\boldsymbol{I} - \gamma\boldsymbol{P}\boldsymbol{\Pi}^{\pi^\star})$ whose negative is Hurwitz.[①] Secondly, if each random reward has bounded p-th order moments, Assumption 3.2.2 holds with $\sigma = 0$ due to the boundedness of $\boldsymbol{Q}^\star$. Thirdly, one can show that Assumption 3.2.3 follows with $L_H = 1 + \gamma$. Finally, the Markov chain determined by π_{b} on the finite space $\mathscr{S}\times\mathscr{A}$ is irreducible and thus uniformly ergodic, which along with the i.i.d. nature of $\boldsymbol{R}_t(s,a)$ implies that Assumption 3.2.4 holds.

[①] Note that $\boldsymbol{P}\boldsymbol{\Pi}^{\pi^\star}$ is a Markov transition kernel on $\mathscr{S}\times\mathscr{A}$ and thus has eigenvalues with norm at most 1. As a result of $\gamma\in[0,1)$, $\boldsymbol{I} - \gamma\boldsymbol{P}\boldsymbol{\Pi}^{\pi^\star}$ has eigenvalues with strictly positive real parts and so its negative is Hurwitz. By Liapunov's theorem, $\boldsymbol{A}$ is Hurwitz if and only if there exists symmetric matrices $\boldsymbol{B}_1, \boldsymbol{B}_2$ such that $\boldsymbol{A}^\top\boldsymbol{B}_1 + \boldsymbol{B}_1\boldsymbol{A} = \boldsymbol{B}_2$. Using this equivalence, one can show $-\boldsymbol{G}$ is Hurwitz as well.

3.3 Main Results

We now turn to the statement of our main results, beginning with a FCLT in Section 3.3.1, followed by consistency guarantees in Section 3.3.2, a semi-parametric efficient lower bound in Section 3.3.3, and ended by functional weak convergence rates in Section 3.3.4.

3.3.1 Functional Central Limit Theorem

Theorem 3.3.1 (FCLT)**.** *Under Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5, and 3.2.6, for the iterate* $\{\boldsymbol{x}_t\}_{t\ge 0}$ *defined by* (3.2) *and any* $r \in [0,1]$, *we define the partial-sum process as what follows*

$$\boldsymbol{\phi}_T(r) := \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} (\boldsymbol{x}_t - \boldsymbol{x}^\star).$$

Then, as a random function on $[0,1]$, $\boldsymbol{\phi}_T(\cdot)$ *weakly converges to a scaled Brownian motion* $\boldsymbol{\psi}(\cdot)$, *i.e.,*

$$\boldsymbol{\phi}_T \overset{w}{\to} \boldsymbol{\psi} := \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W} \tag{3.9}$$

in the Skorohod topology where

$$\boldsymbol{S} := \mathbb{E}_{\xi\sim\pi}\left[\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top\right] \tag{3.10}$$

is the covariance matrix and $\boldsymbol{W} = \{\boldsymbol{W}(r) : r \in [0,1]\}$ *is the standard* d*-dimensional Brownian motion.*

Theorem 3.3.1 shows both the cadlag constant function $\boldsymbol{\phi}_t$ weakly converges to the rescaled Brownian motion $\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{W}$. The scale $\boldsymbol{G}^{-1}\boldsymbol{S}$ involves both the local linearity coefficient $\boldsymbol{G}$ and the covariance matrix $\boldsymbol{S}$. One can show $\boldsymbol{S} = \mathbb{E}_{\xi\sim\pi}\mathrm{Var}_{\xi'\sim P(\xi,\cdot)}(\boldsymbol{U}(\boldsymbol{x}^\star,\xi'))$ is the expected conditional covariance matrix of $\boldsymbol{U}(\boldsymbol{x}^\star,\xi')$ with $\xi\sim\pi$ and $\xi'\sim P(\xi,\cdot)$. This functional weak convergence provides stronger characterization for asymptotic behaviors of the SA scheme (3.2) than pointwise weak convergence. By applying the continuous mapping theorem with a continuous functional f, we can arrive at Corollary 3.3.1.

Corollary 3.3.1. *Under the same assumptions in Theorem 3.3.1, for any* $k \ge 1$ *and any* $|||\cdot|||$*-continuous functional* $f : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}^k$, *it follows that as* $T \to \infty$,

$$f(\boldsymbol{\phi}_T) \overset{d}{\to} f(\boldsymbol{\psi}) = f(\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}).$$

By the corollary we could easily establish weaker pointwise weak convergences by picking up a $|||\cdot|||$-continuous functional f. For example, one can recover the standard i.i.d. CLT in[30] by setting $f : \boldsymbol{\phi} \mapsto \boldsymbol{\phi}(1)$. We say $f : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}^d$ is *scale-invariant* if $f(\boldsymbol{A}\boldsymbol{\phi}) = f(\boldsymbol{\phi})$ for any non-singular matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\phi} \in \mathsf{D}_{[0,1],\mathbb{R}^d}$. Moreover, when we choose f as a scale-invariant functional, we immediately have that $f(\boldsymbol{\phi}_T)$ weakly converges to a functional of the standard Brownian motion because $f(\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}) = f(\boldsymbol{W})$ which eliminates out the dependence of the unknown scale $\boldsymbol{G}^{-1}\boldsymbol{S}$. A close inspection reveals that $f(\boldsymbol{\phi}_T)$ is a pivotal quantity involving only collected data and the unobservable root $\boldsymbol{x}^\star$, while $f(\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}) = f(\boldsymbol{W})$ has a known distribution whose quantiles can be computed via simulation. In this way an asymptotic confidence regime can be constructed. This is the reason why the FCLT underpins the theoretical support of our statistical inference method. By making use of randomness along the whole trajectory $\boldsymbol{\phi}_T$, a confidence region can be formulated by reverting an asymptotic pivotal quantity. We provide a proof sketch in Section 3.4.1 and highlight the technical novelty in the proof of Theorem 3.3.1. Before introducing our inference method, we supplement Theorem 3.3.1 with several side results that would deepen one's understanding on our methods and theories.

3.3.2 Consistency Guarantee

A remaining issue is to ensure the $(L^2, (1+\log t)\sqrt{\eta_t})$-consistency and uniformly bounded p-th moment in Assumption 3.2.6. Typically, this can not be done without further assumptions. Previous work[142] assumes the existence of a smooth Lyapunov function to derive non-asymptotic convergence rates, which suffices to address our issue here. However, for non-smooth applications like Q-Learning, such a well-behaved Lyapunov function is not off-the-shelf. Recently, Chen, Maguluri, Shakkottai, Shanmugam[20, 143] develop a regularized Lyapunov approach for SA problems satisfying a general norm contraction by treating the generalized Moreau envelope as the Lyapunov function. In this way, even for non-smooth SA, a smooth counterpart of Lyapunov functions can be constructed and convergence rates can be established. Inspired by their work, we adopt this approach and narrow down our focus to SA problems satisfying both a similar contraction in Assumption 3.3.1 and a growth condition in Assumption 3.3.2. We emphasize the possibility of finding other general conditions to guarantee Assumption 3.2.6. This subsection provides a particular example.

Assumption 3.3.1 (Contraction condition)**.** *There exist* $\gamma \in [0, 1)$ *and* $c > 0$ *such that*

$$\|\mathscr{P}(\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)) - c \cdot (\boldsymbol{x} - \boldsymbol{y})\| \le \gamma c \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \text{ and } \xi \in \Xi.$$

The contraction condition implies the map $\boldsymbol{x} \to \mathscr{P}\boldsymbol{H}(\boldsymbol{x}, \xi) - c\boldsymbol{x}$ is a γ-contraction in the norm $\|\cdot\|$. The condition $\gamma \in [0, 1)$ ensures that $(1-\gamma)c\|\boldsymbol{x}-\boldsymbol{y}\| \le \|\mathscr{P}(\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi))\| \le$

$(1+\gamma)c\|\boldsymbol{x}-\boldsymbol{y}\|$ uniformly over $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$. This inequality can be viewed as a strengthened version of (3.3) when p defined therein equals to one. For μ-strongly convex and L-smooth convex function[①] $F(\cdot, \xi)$, $\boldsymbol{H}(\boldsymbol{x}, \xi) = \nabla F(\boldsymbol{x}, \xi)$ satisfies Assumption 3.3.1 in the ℓ_2 norm with $c = \frac{2}{L+\mu}$ and $\gamma = \frac{L-\mu}{L+\mu}$. For Q-Learning in (3.8), Assumption 3.3.1 follows in the ℓ_∞ norm with $c = 1$ and γ the discount factor.

Assumption 3.3.2 (Growth condition)**.** *There exist $M > 0$ and a non-negative function $g : \Xi \to \mathbb{R}$ such that*

$$\|\boldsymbol{H}(\boldsymbol{x}, \xi)\| \le M(\|\boldsymbol{x}\| + g(\xi)) \textit{ for any } \boldsymbol{x} \in \mathbb{R}^d \textit{ and } \xi \in \Xi.$$

Furthermore, we assume $\sup_{t\ge 0} \mathbb{E}\|g(\xi_t)\|^p < \infty$ with $p > 2$ given in Assumption 3.2.2.

The growth condition requires the incremental update $\|\boldsymbol{H}(\boldsymbol{x}, \xi)\|$ grows at most linearly in both $\|\boldsymbol{x}\|$ and a non-negative function $g : \Xi \to \mathbb{R}$ that captures the contribution of data ξ to the norm growth of $\|\boldsymbol{H}(\boldsymbol{x}, \xi)\|$. It would be emphasized that we assume $\{g(\xi_t)\}_{t\ge 0}$ has uniformly bounded p-th moments, much milder than previous almost surely uniformly boundedness[20, 103, 123].

Remark 3.3.1. *We impose a slightly stronger contraction condition than previous work[20, 103]. Their counterpart condition is $\|(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})) - c \cdot (\boldsymbol{x} - \boldsymbol{y})\| \le \gamma c \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$ uniformly over $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$ under our notation. This is because we assume a much weaker growth condition than theirs. Under our notation, they all assume* (i) $\sup_{\xi\in\Xi} \|\boldsymbol{H}(\boldsymbol{x}, \xi) - \boldsymbol{H}(\boldsymbol{y}, \xi)\| \le A\|\boldsymbol{x} - \boldsymbol{y}\|$ *uniformly and* (ii) $\sup_{\xi\in\Xi} \|\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \le B$ *for two constants $A, B > 0$. The conditions imply $\|\boldsymbol{H}(\boldsymbol{x}, \xi)\| \le A(\|\boldsymbol{x}\| + \|\boldsymbol{x}^\star\|) + B$ which essentially requires $g(\cdot)$ in Assumption 3.3.2 to be a constant function, excluding the possibility of unbounded observation noises. Take Q-Learning as an example. The theories in[20, 103] work only for (almost surely) uniformly bounded random reward $\boldsymbol{R}(s, a)$'s, while ours allow them to have p-th order moments.*

Our second result is the consistency guarantee under a weaker growth condition.

Theorem 3.3.2. *Under Assumptions 3.2.2, 3.2.3, 3.2.4, 3.2.5, 3.3.1, and 3.3.2, $\{\boldsymbol{x}_t\}_{t\ge 0}$ updated according to* (3.2) *satisfies the $(L^p, \max\{a_t, 1\} \cdot \sqrt{\eta_t})$-consistency with $p > 2$ given in Assumption 3.2.2 and*[①]

$$a_t = \left\lceil t_{\mathrm{mix}}\left(\frac{\eta_t}{2\sigma}\right)\right\rceil = \begin{cases} 0 & \textit{if } \rho = 0, \\ \lceil \log_\rho \frac{\eta_t}{\sigma\kappa} \rceil & \textit{if } \rho \in (0, 1). \end{cases} \tag{3.11}$$

[①] It means $\mu \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \le \|\nabla F(\boldsymbol{x}, \xi) - \nabla F(\boldsymbol{y}, \xi)\| \le L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\xi \in \Xi$.

[①] If $\sigma = 0$, we make a convention that $a_t = 0$.

Theorem 3.3.2 implies that the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ satisfies Assumption 3.2.6. This is because when $p > 2$, the $(L^p, \max\{a_t, 1\}\sqrt{\eta_t})$-consistency naturally implies the $(L^2, \max\{a_t, 1\}\sqrt{\eta_t})$-consistency by Jensen's inequality, which further implies the $(L^2, (1+\log t)\cdot\sqrt{\eta_t})$-consistency due to $a_t = \mathcal{O}(\log t)$.[②] Furthermore, the $(L^p, \max\{a_t, 1\}\sqrt{\eta_t})$-consistency also leads to the uniformly bounded p-th order moment as a result of $\eta_t \log^2 t \to 0$ in Assumption 3.2.5.

Corollary 3.3.2. *Under the same conditions of Theorem 3.3.2, Assumption 3.2.6 holds.*

Previous work requiring i.i.d. data often establish the $(L^p, \sqrt{\eta_t})$-consistency[129, 140, 144]. We comment on the additional a_t factor in Theorem 3.3.2. This is because most analyses for Markovian randomness (including ours) use a conditioning argument of the geometric mixing[20, 123]. Roughly speaking, this argument attempts to address the issue of $\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)|\mathcal{F}_{t-1}] \neq \boldsymbol{0}$ by replacing it with $\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)|\mathcal{F}_{t-a_t-1}] = \mathcal{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-a_t-1})$. The geometric mixing in Assumption 3.2.4 implies that the latter could be exponentially small given a_t is sufficiently large. More specifically, we have $\|\mathcal{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \leq \kappa\sigma\rho^{a_t}$ uniformly over $\xi \in \Xi$ from Lemma 3.2.1. To derive the consistency result, it suffices to set $\|\mathcal{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \leq \eta_t$, which explains the choice of a_t's in (3.11). However, several approximation errors occur before this replacement is taken, whose addressing requires a further elaborate analysis which we defer in the Section 3.4.3 due to the technical complexity. As a result, the square estimation error $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$ typically depends linearly on the squared mixing time a_t^2 (e.g., Theorem 2.1 in Chen, Maguluri, Shakkottai, Shanmugam [20]). Our result provides a more complete characterization on the mixing time a_t's, that is, $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p = \mathcal{O}(a_t^p \eta_t^{\frac{p}{2}})$ depends linearly on a_t^p.

3.3.3 Semiparametric Efficient Lower Bound

Theorem 3.3.1 shows the asymptotic variance of $\boldsymbol{\phi}_T(r)$ at any fraction number $r \in [0, 1]$ is $r\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$. It is of theoretical interest to investigate whether this asymptotic variance matrix is efficient or not. This question has already been addressed in the context of i.i.d. observations; the asymptotic variance of the averaged iterate under this scheme (3.2) is known to achieve the Cramér-Rao lower bound[30, 59]. However, the counterpart result for our Markovian root-finding problem is unclear, which is our target in this subsection.

Before presenting the semi-parametric efficiency lower bound, we first formally describe the estimation task. The parameter of interest $\boldsymbol{x}^\star$ is the root of the equation $\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}, \xi) = \boldsymbol{0}$ where π is the stationary distribution of the transition kernel $P(\xi, d\xi')$. We do not parameterize

[②] This is because by $t\eta_t \uparrow \infty$, we have $\eta_t \gtrsim \frac{1}{t}$.

the kernel P in a finite-dimensional space and thus enter the semiparametric world. We assume a dataset $\mathscr{D} = \{\xi_i\}_{i\in[n]}$ with ξ_i's collected by following the Markov kernel $\boldsymbol{P}$. Here, we denote by n (instead of T) the size of $\mathscr{D}$ following the notation in Greenwood, Wefelmeyer [145].

We define the following perturbed transition kernel $P_{nh}(\xi, d\xi')$,

$$\frac{P_{nh}(\xi, d\xi')}{P(\xi, d\xi')} = 1 + \frac{1}{\sqrt{n}}\boldsymbol{h}(\xi, \xi'),$$

where $\boldsymbol{h}$ is a function on $\Xi \times \Xi$ belonging to the following function class

$$\mathscr{B} := \left\{\boldsymbol{h} \in \mathbb{R}^{\Xi\times\Xi} : \boldsymbol{h} \text{ is bounded, measurable and } \mathbb{E}_{\xi'\sim P(\xi,\cdot)}\boldsymbol{h}(\xi, \xi') = 0 \text{ for all } \xi \in \Xi\right\}.$$

The boundedness of $\boldsymbol{h}$ implies P_{nh} is well-defined as long as n is large enough. By π_{nh} we denote the stationary distribution of P_{nh} and by $\boldsymbol{x}^\star_{nh}$ the root of the equation $\mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}, \xi) = 0$.

Definition 3.3.1 (Regular asymptotic linearity). *We say an estimator $\boldsymbol{T}_n$ (which is a measurable function of $\mathscr{D}$) to be regular for $\boldsymbol{x}^\star$ with limit $\boldsymbol{L}$, if for all $\boldsymbol{h} \in \mathscr{B}$,*

$$n^{1/2}(\boldsymbol{T}_n - \boldsymbol{x}^\star_{nh}) \overset{d}{\to} \boldsymbol{L} \text{ under } P_{nh}.$$

Furthermore, we say $\boldsymbol{T}_n$ to be regular asymptotically linear (RAL) if $\boldsymbol{T}_n$ is both regular for $\boldsymbol{x}^\star$ and asymptotically linear with a measurable function $\boldsymbol{\varphi}$ such that

$$n^{1/2}(\boldsymbol{T}_n - \boldsymbol{x}^\star) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varphi}(\xi_{i-1}, \xi_i) + o_{P^n}(1),$$

where $\boldsymbol{\varphi}$ is referred to as an influence function. It is satisfied that $\mathbb{E}_{\xi\sim\pi,\xi'\sim P(\xi,\cdot)}\boldsymbol{\varphi}(\xi, \xi')\boldsymbol{\varphi}(\xi, \xi')^\top$ is non-singular and $\mathbb{E}_{\xi\sim\pi,\xi'\sim P(\xi,\cdot)}\boldsymbol{\varphi}(\xi, \xi') = 0$.

To establish efficiency lower bound, we focus on an important class of estimators, the *regular asymptotically linear (RAL)* estimators. Tsiatis [146] argued that RAL estimators provide a good tradeoff between expressivity and tractability. Informally speaking, an estimator is regular if its limiting distribution is unaffected by local changes in the data-generating process. In Definition 3.3.1, it means even we perturb the data-generating transition kernel from $\boldsymbol{P}$ to another P_{nh}, the asymptotic distribution of $n^{1/2}(\boldsymbol{T}_n - \boldsymbol{x}^\star_{nh})$ remains unchanged as $\boldsymbol{L}$. This regularity excludes super-efficient estimators, whose asymptotic variance can be smaller than the Cramér-Rao lower bound for some parameter values, but which perform poorly in the neighborhood of points of super-efficiency.

Our third result is the efficiency lower bound for the asymptotic variance of any RAL estimators for $\boldsymbol{x}^\star$. By Definition 3.3.1, any influence function $\boldsymbol{\varphi}$ determines an asymptotic linear estimator for $\boldsymbol{x}^\star$. Theorem 3.3.3 serves as a concrete target in constructing the influence function, and any influence function that achieves this bound is the most efficient among all RAL estimators. Theorem 3.3.3 is also helpful in understanding recent non-asymptotic instance-dependent estimation bounds for Markovian linear SA[104]. These bounds show that $\inf_{\boldsymbol{T}_n} \sup_{P_{nh}} \mathbb{E}\|\boldsymbol{T}_n - \boldsymbol{x}^\star\|_2^2$, the minimax estimation bounds in ℓ_2-norm, is lower bounded by an instance-dependent quantity $\frac{1}{n}\|\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\|_F^2$, which can be computed directly from (3.12).[1] Similar correspondence is found in i.i.d. nonlinear SA[21, 100].

Theorem 3.3.3 (Semiparametric efficient lower bound)**.** *Under Assumptions 3.2.1, 3.2.3, 3.2.4, and 3.3.1, for any RAL estimator $\boldsymbol{T}_n$ for $\boldsymbol{x}^\star$ that is computed from $\mathscr{D} = \{\xi_i\}_{i\in[n]}$, we have*

$$\lim_{n\to\infty} n \cdot \mathbb{E}(\boldsymbol{T}_n - \boldsymbol{x}^\star)(\boldsymbol{T}_n - \boldsymbol{x}^\star)^\top \succeq \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top} \tag{3.12}$$

with both $\boldsymbol{G}$ and $\boldsymbol{S}$ defined in Theorem 3.3.1.

Theorem 3.3.4 implies that for each $r \in [0,1]$, $\frac{1}{\lfloor Tr\rfloor}\sum_{t=0}^{\lfloor Tr\rfloor} \boldsymbol{x}_t$ is most efficient estimator among all RAL estimators and the influence function is given by $\boldsymbol{\varphi}(\xi,\xi') = \boldsymbol{U}(\boldsymbol{x}^\star,\xi') - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)$. One can show this by the fact that $\boldsymbol{\varphi}(\xi,\xi')$ is mean-zero and its covariance matrix is exactly $\boldsymbol{S}$ when $(\xi,\xi') \sim \pi(d\xi)\times P(\xi,d\xi')$. This theorem implies the partial-sum process $\boldsymbol{\phi}_T$ has the optimal asymptotic variance at each fraction $r \in [0,1]$. By contrast, the scaled last-iterate process typically fails to achieve it[129, 144].

Theorem 3.3.4. *Under the same conditions of Theorem 3.3.1, for any $r \in [0,1]$, the partial-sum value $\frac{1}{\lfloor Tr\rfloor}\sum_{t=0}^{\lfloor Tr\rfloor} \boldsymbol{x}_t$ is a RAL estimator for $\boldsymbol{x}^\star$ with the following decomposition*

$$\boldsymbol{\phi}_T(r) := \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}(\boldsymbol{x}_t - \boldsymbol{x}^\star) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}[\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})] + o_{\mathbb{P}}(1),$$

where $\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1}) = \mathbb{E}[\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)|\mathscr{F}_{t-1}]$ and $o_{\mathbb{P}}(1)$ denotes a random function whose uniform norm $|||\cdot|||$ converge to zero in probability.

Proof of Theorem 3.3.4. In Section 3.4.1, we have analyzed $\boldsymbol{\phi}_T$. We will make use of many results obtained therein. We decompose $\boldsymbol{\phi}_T$ into several terms in (3.20)

$$\tilde{\boldsymbol{\phi}}_T(r) - \frac{1}{\sqrt{T}}\sum_{j=0}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{u}_t = \boldsymbol{\psi}_0(r) + \boldsymbol{\psi}_1(r) + \boldsymbol{\psi}_2(r) + \boldsymbol{\psi}_3(r).$$

[1] For any standard basis $\boldsymbol{e}_j \in \mathbb{R}^d$, we have $\lim_{n\to\infty} n \cdot \mathbb{E}\boldsymbol{e}_j^\top(\boldsymbol{T}_n - \boldsymbol{x}^\star)(\boldsymbol{T}_n - \boldsymbol{x}^\star)^\top\boldsymbol{e}_j \geq \boldsymbol{e}_j^\top\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\boldsymbol{e}_j$. Summing over $j \in [d]$, we arrive at $\lim_{n\to\infty} n\mathbb{E}\|\boldsymbol{T}_n - \boldsymbol{x}^\star\|_2^2 \geq \|\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\|_F^2$.

We also show that $\left|\!\left|\!\left|\tilde{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_T\right|\!\right|\!\right| = o_{\mathbb{P}}(1)$ and $\left|\!\left|\!\left|\boldsymbol{\psi}_k\right|\!\right|\!\right| = o_{\mathbb{P}}(1)$ for $0 \le k \le 3$ as in Lemma 3.4.2. In the proof of Lemma 3.4.1, we also decompose $\boldsymbol{u}_t := \boldsymbol{u}_{t,1} + \boldsymbol{u}_{t,2}$ into two terms in (B.1) such that we have

$$\left|\!\left|\!\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\boldsymbol{u}_{t,1}\right|\!\right|\!\right| = o_{\mathbb{P}}(1) \text{ and } \boldsymbol{u}_{t,2} = \left[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\right].$$

Putting these results together and using Slutsky's theorem, we complete the proof. □

3.3.4 Functional Weak Convergence Rate

In this subsection, we provide a more quantitative result that specifies the rate at which the weak convergence in Theorem 3.3.1 takes place. For two random processes $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ in $\mathsf{D}_{[0,1],\mathbb{R}^d}$, we denote by d_{P} the Lévy-Prokhorov distance between the probability measures introduced in[147-148], that is,

$$d_{\mathrm{P}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) := \inf\left\{\varepsilon : \mathbb{P}(\boldsymbol{\phi}_1 \in B) \le \mathbb{P}(\boldsymbol{\phi}_2 \in B^\varepsilon) + \varepsilon, \ \forall B \in \mathscr{D}_{[0,1],\mathbb{R}^d}\right\}, \tag{3.13}$$

where $\boldsymbol{B}^\varepsilon := \left\{\boldsymbol{\phi}_1 \in \mathsf{D}_{[0,1],\mathbb{R}^d} : \inf_{\boldsymbol{\phi}_2 \in B} d_{\mathrm{S}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) < \varepsilon\right\}$ is the ε-neighborhood of $\boldsymbol{B}$. Since $\mathsf{D}_{[0,1],\mathbb{R}^d}$ with the Skorokhod metric is separable, convergence in the Lévy–Prokhorov metric is equivalent to weak convergence of the corresponding measures, as a result of which, we have $d_{\mathrm{P}}(\boldsymbol{\phi}_T, \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}) \to 0$ from Theorem 3.3.1.

Assumption 3.3.3 (Further regularity conditions)**.** (i) *Assume the initial data $\xi_0 \sim \pi$ and*

$$\sup_{\xi\in\Xi}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\|^2 := \sup_{\xi\in\Xi}\int_\Xi \|\boldsymbol{H}(\boldsymbol{x}^\star, \xi')\|^2 P(\xi, d\xi') < \infty.$$

(ii) *$\{\boldsymbol{x}_t\}_{t\ge0}$ satisfies the $(L^p, (1+\log t)\cdot\sqrt{\eta_t})$-consistency with $p > 2$ given in Assumption 3.2.2.*

Theorem 3.3.5 (Functional weak convergence rate)**.** *Let Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, and 3.3.3 hold and choose the step size to be $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$. It follows that for any vector $\boldsymbol{\theta} \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}\|_* = 1$,*

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top\boldsymbol{\phi}_T, \boldsymbol{\theta}^\top\boldsymbol{\psi}) = \tilde{\mathscr{O}}\left(T^{-J_1(\alpha)} + (c_r + t_{\mathrm{mix}})^{\frac{p}{2+p}} \cdot T^{-J_2(\alpha)} + T^{-(1-\alpha)\left[\frac{p-2}{2(p+1)}\wedge\frac{1}{3}\right]} + t_{\mathrm{mix}}^{\frac{1}{6}}T^{-\frac{1}{6}}\right), \tag{3.14}$$

where $p > 2$ is given in Assumption 3.2.2, $c_r := \max\left\{L_G, \frac{L_H+\|\boldsymbol{G}\|}{\delta_G}\right\}$, t_{mix} is the mixing time *defined in* (3.6)*, and both* $\boldsymbol{J}_1$ *and* $\boldsymbol{J}_2$ *are increasing functions of* α *given by*[1]

$$\boldsymbol{J}_1(\alpha) = \begin{cases} \frac{\alpha p}{2(2+p)} & \textit{if } \alpha \in \left(0.5, \frac{2}{p}\right], \\ \frac{\alpha}{\alpha+2} & \textit{if } \alpha \in \left[\frac{2}{p}, 1\right), \end{cases} \quad \textit{and} \quad \boldsymbol{J}_2(\alpha) = \begin{cases} (\alpha-0.5)\frac{p}{2+p} & \textit{if } \alpha \in \left(0.5, \frac{2}{p}\right], \\ \frac{\alpha-0.5}{\alpha+1} & \textit{if } \alpha \in \left[\frac{2}{p}, 1\right). \end{cases}$$

Here we hide dependence on uninterested parameters and the log factors in $\tilde{\mathcal{O}}(\cdot)$.

Our last result is the functional weak convergence rate (3.14) for the one-dimensional projected partial-sum process $\boldsymbol{\theta}^\top \boldsymbol{\phi}_T$. To establish this theorem, we impose an additional Assumption 3.3.3. It requires ξ_0 is initialized as the stationary distribution π and assumes a uniform bound for $\sup_{\xi\in\Xi} \mathcal{P}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\|^2$. The former condition is standard in nonasymptotic analysis for Markovian data[104], while the later condition is mildly weaker than the uniform boundedness used in the literature (see Remark 3.3.1). The discussion in Section 3.3.2 reveals the $(L^p, (1+\log t)\cdot\sqrt{\eta_t})$-consistency follows when Assumptions 3.3.1 and 3.3.2 hold. In short, Assumption 3.3.3 is mild and standard.

The bound (3.14) is an analog of the Berry-Esseen bounds on the distance between the distributions of cádág functions in $\mathsf{D}_{[0,1],\mathbb{R}}$ measured in the Lévy-Prokhorov metric. To the best of our knowledge, it is the first non-asymptotic bound of functional weak convergence for the nonlinear iterative algorithm (3.2) in the existence of Markovian data. If $\{\boldsymbol{\theta}^\top(\boldsymbol{x}_t - \boldsymbol{x}^\star)\}_{t\geq 0}$ is i.i.d. with zero mean and bounded p-th order moments ($p \in [2, 3]$), Borovkov [149] showed the bound for $d_{\mathrm{P}}(\boldsymbol{\theta}^\top\boldsymbol{\phi}_T, \boldsymbol{\theta}^\top\boldsymbol{\psi})$ is $\mathcal{O}\left(T^{-\frac{p-2}{2(p+1)}}\right)$. Haeusler [110], Kubilius [150] showed the same bound holds for martingale difference sequences under specific moment conditions. However, our result has a slower rate, as the third term in (3.14) alone is already slower than $\mathcal{O}\left(T^{-\frac{p-2}{2(p+1)}}\right)$. The main cause is that the sequence $\{\boldsymbol{\theta}^\top(\boldsymbol{x}_t - \boldsymbol{x}^\star)\}_{t\geq 0}$ is neither stationary nor martingale differences. The non-stationarity of $\boldsymbol{x}_t$, not remaining at $\boldsymbol{x}^\star$, introduces additional errors, slowing down the rate. More exactly, Theorem 3.3.4 states that for any $r \in [0, 1]$, $\boldsymbol{\phi}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor} \boldsymbol{G}^{-1}[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathcal{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})] + o_{\mathbb{P}}(1)$, with $\{\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathcal{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\}_{t\geq 1}$ being a fast mixing martingale difference under Assumptions 3.2.4 and 3.3.3. According to the existing result[150], the Lévy–Prokhorov distance between the partial-sum process $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor} \boldsymbol{\theta}^\top\boldsymbol{G}^{-1}[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathcal{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})]$ and the scaled Brownian motion $\boldsymbol{\theta}^\top\boldsymbol{\psi} = \boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}$ is roughly $\tilde{\mathcal{O}}(T^{-\frac{p-2}{2(p+1)}\wedge\frac{1}{3}} + t_{\mathrm{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}})$. Therefore, the remaining $o_{\mathbb{P}}(1)$ term causes the slow convergence rate in (3.14). The detailed proof of Theorem 3.3.5 is collected in the Appendix D.

[1] Here we interpret $\left(0.5, \frac{2}{p}\right] = \varnothing$ if $p > 4$.

There are many implications from the bound (3.14). We list them below.

- Markovian data slows down the functional convergence rate polynomially due to the second and fourth terms of (3.14). It together with Theorem 3.3.2 implies Markovian data with a bounded mixing time has limited consequences on both the estimation error $\sqrt[p]{\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p}$ and the weak convergence rate $d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi})$.
- It is vital to ensure $p > 2$, otherwise the third term in (3.14) might blow up or keep non-diminishing. Furthermore, the bound (3.14) mildly changes when the moment order p increases.
- The non-linearity attacks the weak convergence rate via the quantity $c_r = \max\left\{L_G, \frac{L_H + \|\boldsymbol{G}\|}{\delta_G}\right\}$. If we consider linear SA, then $\delta_G = \infty$ and $L_G = 0$ in Assumption 3.2.1, which implies $c_r = 0$. For nonlinear SA, c_r serves as a measure that quantifies the degree of nonlinearity.
- There exists a trade-off for the step size parameter $\alpha \in (0.5, 1)$. Indeed, since both J_1 and J_2 are increasing, the first and second terms in (3.14) decrease in α, while the third term increases in α. It is of theoretical interest to investigate the optimal $\alpha^\star$ and the resulting weak convergence rates. As Corollary 3.3.3 shows, the optimal rate in linear and i.i.d. case nearly matches the pointwise rate $\tilde{\mathcal{O}}(T^{-\frac{1}{6}})$[151] when the moment order p is sufficiently large. However, it deteriorates almost by half once either non-linearity or Markovian data gets involved.

Corollary 3.3.3. *Under the same conditions of Theorem 3.3.5, if $t_{\mathrm{mix}} = c_r = 0$, then for any small $\varepsilon > 0$,*

$$\min_{\alpha \in [0.5+\varepsilon, 1)} d_{\mathrm{P}} P\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}\right) = \tilde{\mathcal{O}}\left(T^{-\left[\frac{p-2}{4(p+1)} \wedge \frac{1}{6}\right](1-2\varepsilon)}\right),$$

with the optimum achieved by $\alpha^\star = 0.5 + \varepsilon$. If either $t_{\mathrm{mix}} > 0$ or $c_r > 0$, it follows that

$$\min_{\alpha \in (0.5, 1)} d_{\mathrm{P}} P\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}\right) = \tilde{\mathcal{O}}\left(\left[(c_r + t_{\mathrm{mix}})^{\frac{p}{2+p}} + 1\right] \cdot T^{-J(p)}\right),$$

where $J(\cdot)$ is defined as follows

$$J(p) = \begin{cases} \frac{(p-2)p}{2(3p^2+2p-1)} & \text{if } p \in (2, p_0], \\ \frac{(2p-1)-\sqrt{3(p^2-p+1)}}{2(p+1)} & \text{if } p \in [p_0, 8], \\ \frac{5-\sqrt{19}}{6} \approx 0.107 & \text{if } p \in [8, \infty), \end{cases}$$

with $p_0 \in (3,4)$ *a number making* $\boldsymbol{J}(\cdot)$ *continuous and and with optimum achieved by*

$$\alpha^\star(p) = \begin{cases} \frac{2p^2+p-4}{3p^2+2p-4} & \textit{if } p \in (2, p_0], \\ \frac{\sqrt{3(p^2-p+1)}-(p+1)}{p-2} & \textit{if } p \in [p_0, 8], \\ \frac{\sqrt{19}-3}{2} \approx 0.679 & \textit{if } p \in [8, \infty). \end{cases}$$

Proof of Corollary 3.3.3. The proof can be found in Appendix D.1. □

3.4 Proof Sketches

Before introducing our inference method, we provide the proof sketches for the four theorems present in the last chapter and highlight our technical contributions therein. The proofs for several technical lemmas are deferred in the appendix.

3.4.1 Proof of Theorem 3.3.1

The proof for Theorem 3.3.1 contains three steps. Proofs for technical lemmas are deferred in Appendix B.

Step one: Martingale-residual-coboundary decomposition Recall that the update rule is $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$. We decompose $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$ into two terms:

$$\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) = \boldsymbol{g}(\boldsymbol{x}_t) + \left[\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(\boldsymbol{x}_t)\right].$$

By Lemma 3.2.2, there exists a unique bivariate function $\boldsymbol{U}(\boldsymbol{x}, \xi)$ such that $\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(\boldsymbol{x}_t) = \boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t)$. We further decompose $\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t)$ into three terms:

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) = & \underbrace{\left[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\right]}_{martingale} + \underbrace{\left[\frac{\eta_{t+1}}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1}, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t)\right]}_{residual} \\ & + \underbrace{\left[\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1}) - \frac{\eta_{t+1}}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1}, \xi_t)\right]}_{coboundary}. \end{aligned} \tag{3.15}$$

We refer to the last equation as martingale-residual-coboundary decomposition which is reminiscent of the martingale-coboundary decomposition that is originally proposed to establish FCLTs for stationary sequences[152-153]. This martingale-residual-coboundary decomposition is recently used in the asymptotic analysis for stochastic approximation MCMC algorithms[111]. The telescoping structure in the coboundary term motivates us to introduce an auxiliary process $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 0}$ to remove its effect where

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta_t \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1}).$$

As a result, we have

$$\tilde{\boldsymbol{x}}_{t+1} = \tilde{\boldsymbol{x}}_t - \eta_t \left[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1}) + \frac{\eta_{t+1}}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1}, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) \right].$$

We then focus on $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 0}$ and simplify the last equation by introducing the following shortcuts: $\boldsymbol{\Delta}_t = \tilde{\boldsymbol{x}}_t - \boldsymbol{x}^\star$ and

$$\boldsymbol{r}_t = \boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{G}\boldsymbol{\Delta}_t, \tag{3.16}$$

$$\boldsymbol{u}_t = \left[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\right], \tag{3.17}$$

$$\boldsymbol{v}_t = \frac{\eta_{t+1}}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1}, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_t). \tag{3.18}$$

With the notation, the update rule becomes

$$\boldsymbol{\Delta}_{t+1} = \boldsymbol{\Delta}_t - \eta_t\left[\boldsymbol{G}\boldsymbol{\Delta}_t + \boldsymbol{r}_t + \boldsymbol{u}_t + \boldsymbol{v}_t\right] = (\boldsymbol{I} - \eta_t\boldsymbol{G})\boldsymbol{\Delta}_t + \eta_t\left[\boldsymbol{r}_t + \boldsymbol{u}_t + \boldsymbol{v}_t\right]. \tag{3.19}$$

The following lemma explains the reason why we perform the decomposition (3.15). It shows, while $\{\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{g}(\boldsymbol{x}_t)\}_{t\geq 0}$ is not a martingale difference sequence, the decomposed $\{\boldsymbol{u}_t\}_{t\geq 0}$ is. Furthermore, $\{\boldsymbol{u}_t\}_{t\geq 0}$ admits an FCLT via a standard argument of multidimensional martingale FCLT (e.g., Theorem 2.1 in Whitt [154]). The remaining terms $\{\boldsymbol{r}_t\}_{t\geq 0}$ and $\{\boldsymbol{v}_t\}_{t\geq 0}$ have negligible effects because they vanish asymptotically.

Lemma 3.4.1 (Properties of decomposed terms)**.** *Under the same conditions of Theorem 3.3.1,*

1. *It follows that as* $T \to \infty$, $\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{r}_t\| \to 0$;
2. $\{\boldsymbol{u}_t\}_{t\geq 0}$ *is a martingale difference sequence satisfying* $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{u}_t\|^p < \infty$ *where* $p > 2$ *is given in Assumption 3.2.2. Furthermore, the following FCLT holds* $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_t \xrightarrow{w} \boldsymbol{S}^{1/2}\boldsymbol{W}(r)$;
3. *It follows that as* $T \to \infty$, $\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{v}_t\| \to 0$.

Proof of Lemma 3.4.1. The proof can be found in Appendix B.1. □

Step two: Martingale-remainder (or partial-sum) decomposition Setting $\boldsymbol{B}_t = \boldsymbol{I} - \eta_t\boldsymbol{G}$ and recurring (3.19) give

$$\boldsymbol{\Delta}_{t+1} = \left(\prod_{j=0}^{t}\boldsymbol{B}_j\right)\boldsymbol{\Delta}_0 + \sum_{j=0}^{t}\left(\prod_{i=j+1}^{t}\boldsymbol{B}_i\right)\eta_j\left[\boldsymbol{r}_j + \boldsymbol{u}_j + \boldsymbol{v}_j\right].$$

Here we use the convention that $\prod_{j=t+1}^{t} \boldsymbol{B}_j = \boldsymbol{I}$ for any $t \geq 0$. As a result, for any $r \in [0,1]$,

$$
\begin{aligned}
\tilde{\boldsymbol{\phi}}_T(r) &:= \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} (\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^\star) \\
&= \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left\{ \left(\prod_{j=0}^{t} \boldsymbol{B}_j \right) \boldsymbol{\Delta}_0 + \sum_{j=0}^{t} \left(\prod_{i=j+1}^{t} \boldsymbol{B}_i \right) \eta_j \left[\boldsymbol{r}_j + \boldsymbol{u}_j + \boldsymbol{v}_j \right] \right\} \\
&= \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left(\prod_{j=0}^{t} \boldsymbol{B}_j \right) \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=0}^{\lfloor Tr \rfloor} \sum_{t=j}^{\lfloor Tr \rfloor} \left(\prod_{i=j+1}^{t} \boldsymbol{B}_i \right) \eta_j \left[\boldsymbol{r}_j + \boldsymbol{u}_j + \boldsymbol{v}_j \right].
\end{aligned}
$$

In the following, for simplicity we define

$$
\boldsymbol{A}_j^n := \sum_{t=j}^{n} \left(\prod_{i=j+1}^{t} \boldsymbol{B}_i \right) \eta_j.
$$

Using the notation, we further simplify the last equation as

$$
\tilde{\boldsymbol{\phi}}_T(r) = \frac{1}{\sqrt{T}\eta_0} \boldsymbol{A}_0^{\lfloor Tr \rfloor} \boldsymbol{B}_0 \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{j=0}^{\lfloor Tr \rfloor} \boldsymbol{A}_j^{\lfloor Tr \rfloor} \left[\boldsymbol{r}_j + \boldsymbol{u}_j + \boldsymbol{v}_j \right].
$$

Arrangement yields

$$
\begin{aligned}
\tilde{\boldsymbol{\phi}}_T(r) - \frac{1}{\sqrt{T}} \sum_{j=0}^{\lfloor Tr \rfloor} \boldsymbol{G}^{-1} \boldsymbol{u}_t &= \frac{1}{\sqrt{T}\eta_0} \boldsymbol{A}_0^{\lfloor Tr \rfloor} \boldsymbol{B}_0 \boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{A}_t^{\lfloor Tr \rfloor} (\boldsymbol{r}_t + \boldsymbol{v}_t) \\
&\quad + \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left(\boldsymbol{A}_t^T - \boldsymbol{G}^{-1} \right) \boldsymbol{u}_t + \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left(\boldsymbol{A}_t^{\lfloor Tr \rfloor} - \boldsymbol{A}_t^T \right) \boldsymbol{u}_t \\
&:= \boldsymbol{\psi}_0(r) + \boldsymbol{\psi}_1(r) + \boldsymbol{\psi}_2(r) + \boldsymbol{\psi}_3(r). \qquad (3.20)
\end{aligned}
$$

Step three: Establishment of FCLT By (3.20), we are ready to prove the Theorem 3.3.1. First, from Lemma 3.4.1, the functional weak convergence follows that $\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{G}^{-1} \boldsymbol{u}_t \overset{w}{\to} \boldsymbol{\psi}(r) = \boldsymbol{S}^{1/2} \boldsymbol{G}^{-1} \boldsymbol{W}(r)$ uniformly over $r \in [0,1]$. Second, $\mathbb{E} \left|\!\left|\!\left| \tilde{\boldsymbol{\phi}}_T - \boldsymbol{\phi}_T \right|\!\right|\!\right| \lesssim \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_T \to 0$ because of Lemma 3.2.2, Assumption 3.2.5 and 3.2.6. It implies the random function $\boldsymbol{\phi}_T$ has the same asymptotic behavior as $\tilde{\boldsymbol{\phi}}_T$, i.e., $\boldsymbol{\phi}_T = \tilde{\boldsymbol{\phi}}_T + o_{\mathbb{P}}(1)$. To complete the proof, it suffices to show that

$$
\left|\!\left|\!\left| \tilde{\boldsymbol{\phi}}_T - \boldsymbol{\psi} \right|\!\right|\!\right| = \sup_{r \in [0,1]} \left\| \tilde{\boldsymbol{\phi}}_T(r) - \boldsymbol{\psi}(r) \right\| = o_{\mathbb{P}}(1). \qquad (3.21)
$$

In this way, one has $\tilde{\boldsymbol{\phi}}_T = \boldsymbol{\psi} + o_{\mathbb{P}}(1)$ and thus $\boldsymbol{\phi}_T = \boldsymbol{\psi} + o_{\mathbb{P}}(1)$ due to Slutsky's theorem. Lemma 3.4.2 provides a sufficient condition to (3.21) where the four separate terms

$\sup_{r\in[0,1]} \|\boldsymbol{\psi}_k(r)\|(0 \le k \le 3)$ respectively converge to zero in probability.

Lemma 3.4.2. *Under the same conditions of Theorem 3.3.1, for all* $0 \le k \le 3$*, when* $T \to \infty$*,*

$$\left\|\left\|\boldsymbol{\psi}_k\right\|\right\| = \sup_{r\in[0,1]} \|\boldsymbol{\psi}_k(r)\| = o_{\mathbb{P}}(1).$$

Proof of Lemma 3.4.2. The proof can be found in Appendix B.2. □

Difficulty of analyzing $\boldsymbol{\psi}_3$ In the proof of Lemma 3.4.2, the largest difficulty is to analyze the last process $\boldsymbol{\psi}_3$. Because $\boldsymbol{\psi}_3(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\left(\boldsymbol{A}_t^{\lfloor Tr\rfloor} - \boldsymbol{A}_t^T\right)\boldsymbol{u}_t$ is a weighted sum of martingale differences $\boldsymbol{u}_t$'s whose weights depend on the fraction r, we can't apply Doob's inequality to bound $\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|$. We made a novel technical contribution towards an elaborate analysis for $\sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\|$. In particular, a close inspection reveals that

$$\left\|\left\|\boldsymbol{\psi}_3\right\|\right\| = \sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\| \lesssim \sup_{n\in[T]}\left\|\frac{1}{\sqrt{T}}\frac{1}{\eta_{n+1}}\sum_{t=0}^{n}\left(\prod_{i=t+1}^{n}\boldsymbol{B}_i\right)\eta_t\boldsymbol{u}_t\right\|. \tag{3.22}$$

In Lemma 3.4.3, we show that the right-hand side of (3.22) is indeed $o_{\mathbb{P}}(1)$.

Lemma 3.4.3. *Let* $\{\boldsymbol{\varepsilon}_t\}_{t\ge0}$ *be a martingale difference sequence adapting to the filtration* $\mathcal{F}_t$*. Define an auxiliary sequence* $\{\boldsymbol{y}_t\}_{t\ge0}$ *as follows:* $\boldsymbol{y}_0 = \boldsymbol{0}$ *and for* $t \ge 0$*,*

$$\boldsymbol{y}_{t+1} = (\boldsymbol{I} - \eta_t\boldsymbol{G})\boldsymbol{y}_t + \eta_t\boldsymbol{\varepsilon}_t. \tag{3.23}$$

It is easily verified that

$$\boldsymbol{y}_{t+1} = \sum_{j=0}^{t}\left(\prod_{i=j+1}^{t}\left(\boldsymbol{I} - \eta_i\boldsymbol{G}\right)\right)\eta_j\boldsymbol{\varepsilon}_j. \tag{3.24}$$

Let $\{\eta_t\}_{t\ge0}$ *satisfy Assumption 3.2.5. If* $\mathrm{Re}\lambda_i(\boldsymbol{G}) > 0$ *for all* $i \in [d]$ *and* $\sup_{t\ge0}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^p < \infty$ *for* $p > 2$*, then we have that when* $T \to \infty$

$$\left\|\left\|\bar{\boldsymbol{y}}_T\right\|\right\| \xrightarrow{p} 0 \quad where \quad \bar{\boldsymbol{y}}_T(r) = \frac{\boldsymbol{y}_{\lfloor (T+1)r\rfloor}}{\sqrt{T}\eta_{\lfloor (T+1)r\rfloor}} \quad for \quad r \in [0,1]. \tag{3.25}$$

Furthermore, if setting $\eta_t = t^{-\alpha}$ *with* $\alpha \in (0.5, 1)$*, we have that for any* $p' \in [2, p]$*,*

$$\tilde{d}(\bar{\boldsymbol{y}}_T) := \inf_{\varepsilon\ge0}\varepsilon \vee \mathbb{P}(\left\|\left\|\bar{\boldsymbol{y}}_T\right\|\right\| \ge \varepsilon) = \mathcal{O}\left(p' \cdot T^{-(1-\alpha)\frac{p'-2}{2(p'+1)}}\right). \tag{3.26}$$

Proof of Lemma 3.4.3. The proof can be found in Appendix B.3. □

Although some works establish similar counterparts of Lemma 3.4.3 for SA algorithms, our Lemma 3.4.3 is the most general in three aspects. First, it relaxes the restriction on $-\boldsymbol{G}$

from being negative definite to Hurwitz[50, 62, 129]. Second, it requires uniformly bounded $p(>2)$-th order moments on the martingale difference sequence $\{\boldsymbol{\varepsilon}_t\}_{t\geq 0}$ rather than bounded forth moments[21, 62]. Last, it accommodates a general step size in Assumption 3.2.5 instead of simple polynomial step sizes[129]. We made this improvement from a key observation that Lemma 3.4.3 is easy to prove via a similar argument in Lemma 2.5.8 when $\boldsymbol{G}$ is further diagonalizable. For the general non-diagonalizable case, without loss of generality, we assume $\boldsymbol{G}$ is a matrix of Jordan canonical form by utilizing its Jordan decomposition. Then, the fact that $\boldsymbol{G}$ would be upper triangular motivates an induction proof to relate the projection components of $\boldsymbol{y}_{t+1}$ on non-diagonalizable Jordan blocks to those on diagonalizable ones, completing the proof for the asymptotic result (3.25). The proof idea also motivates a method to quantify the rate (3.26) of convergence in probability. One can show that $\tilde{d}(\bar{\boldsymbol{y}}_T) \to 0$ is equivalent to $|||\bar{\boldsymbol{y}}_T||| \overset{p}{\to} 0$. This quantitative bound (3.26) provides a great help in establishing the weak convergence rate in Theorem 3.3.5. We believe it would benefit future quantitative studies on weak convergence of iterative algorithms.

3.4.2 Establishment of $(L^2, a_t\sqrt{\eta_t})$-consistency

We present a proof for Theorem 3.3.2 in the following two subsections. The first subsection establishes $(L^2, a_t\sqrt{\eta_t})$-consistency, while the second subsection deals with $(L^p, a_t\sqrt{\eta_t})$-consistency. We begin with $(L^2, a_t\sqrt{\eta_t})$-consistency because it is easier to establish using existing techniques. Additionally, based on this result, one can more easily understand the way we prove $(L^p, a_t\sqrt{\eta_t})$-consistency. At a high level, we adapt the generalized Lyapunov approach developed in[20, 143] to our case. Throughout these subsections, we use the $\ell_{\bar{p}}$-norm, denoted by $\|\cdot\|_{\bar{p}}$, defined in $\mathbb{R}^d$. Readers should note that the $\bar{p}$ used in this section has no relation to the p defined in Assumption 3.2.2.

Lemma 3.4.4 (Smoothness and approximation of the envelope, Lemma 2.1 in Chen, Maguluri, Shakkottai, Shanmugam [143]). *Let $\|\cdot\|_{\bar{p}}$ denote the $\ell_{\bar{p}}$-norm defined in $\mathbb{R}^d$. Define the Moreau envelope of $\frac{1}{2}\|\cdot\|^2$ w.r.t. $\frac{1}{2}\|\cdot\|_{\bar{p}}^2$ as*

$$M(\boldsymbol{x}) = \min_{\boldsymbol{u}\in\mathbb{R}^d}\left[\frac{1}{2}\|\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_{\bar{p}}^2\right].$$

We then have the following results

1. *$\boldsymbol{M}(\boldsymbol{x})$ is convex in $\boldsymbol{x}$ and is $(\bar{p}-1)\lambda$-smooth w.r.t. the norm $\|\cdot\|_{\bar{p}}$.*

2. *Suppose $l_{\bar{p}}\|\cdot\| \le \|\cdot\|_{\bar{p}} \le u_{\bar{p}}\|\cdot\|$, then for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\left(1+\frac{\lambda}{u_{\bar{p}}^2}\right) M(\boldsymbol{x}) \le \frac{1}{2}\|\boldsymbol{x}\|^2 \le \left(1+\frac{\lambda}{l_{\bar{p}}^2}\right) M(\boldsymbol{x}).$$

3. *There exists one norm $\|\cdot\|_M$ such that $M(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_M^2$ for all $\boldsymbol{x} \in \mathbb{R}^d$.*

Proof of Theorem 3.3.2. Recall that the update rule is $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$. Hence, it follows that

$$\begin{aligned} M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star) &\le M(\boldsymbol{x}_t - \boldsymbol{x}^\star) + \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{(\bar{p}-1)\lambda}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_{\bar{p}}^2 \\ &= M(\boldsymbol{x}_t - \boldsymbol{x}^\star) - \eta_t \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t) \rangle + \frac{(\bar{p}-1)\lambda\eta_t^2}{2}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^2. \end{aligned} \tag{3.27}$$

Let $\mathscr{G}_t = \sigma(\{\xi_\tau\}_{0\le\tau<t})$ be the σ-filed generated by all random variables $\{\xi_\tau\}_{0\le\tau<t}$ strictly before iteration t. Clearly, $\boldsymbol{x}_t$ is $\mathscr{G}_t$-measurable. We denote $\mathbb{E}_t(\cdot)$ by $\mathbb{E}[\cdot|\mathscr{G}_t]$ for simplicity and $\mathbb{E}(\cdot)$ takes all randomness. For one thing,

$$\begin{aligned} \mathbb{E}_t\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^2 &\le \mathbb{E}_t\left(\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|_{\bar{p}} + \|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|_{\bar{p}}\right)^2 \\ &\le 2\left(\mathbb{E}_t\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|_{\bar{p}}^2 + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|_{\bar{p}}^2\right) \\ &\le 2u_{\bar{p}}^2\left(\mathbb{E}_t\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^2 + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^2\right) \\ &\le 2u_{\bar{p}}^2\left(\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\|^2 + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^2\right) \\ &\overset{(a)}{\le} 2u_{\bar{p}}^2\left(L_H^2\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^2\right) \\ &\overset{(b)}{\le} 2u_{\bar{p}}^2\left(2L_H^2\left(1+\frac{\lambda}{l_{\bar{p}}^2}\right)M(\boldsymbol{x}_t - \boldsymbol{x}^\star) + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^2\right), \end{aligned} \tag{3.28}$$

where (a) uses Assumption 3.2.3 and (b) uses the Item 2 in Lemma 3.4.4.

For another thing, we decompose the cross term into three part as following

$$\begin{aligned} &\mathbb{E}_t\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle \\ &\quad= \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1})\rangle \\ &\quad= \underbrace{\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) - c(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle}_{I} \\ &\qquad+ \underbrace{c\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle}_{II} + \underbrace{\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\rangle}_{III}. \end{aligned} \tag{3.29}$$

We are going to analyze the three terms in (3.29) respectively in the following.

For the term I From the Item 3 in Lemma 3.4.4, we have $M(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_M^2$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Hence, $\nabla M(\boldsymbol{x}) = \|\boldsymbol{x}\|_M \boldsymbol{v}_{\boldsymbol{x}}$ where $\boldsymbol{v}_{\boldsymbol{x}} \in \partial \|\boldsymbol{x}\|_M$ is a subgradient of the function $\|\boldsymbol{x}\|_M$ at $\boldsymbol{x}$. Let $\|\cdot\|_M^\star$ denote the dual norm of $\|\cdot\|_M$, defined by $\|\boldsymbol{x}\|_M^\star = \sup_{\|\boldsymbol{y}\|_M \le 1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$. Since $\|\cdot\|_M$ is a 1-Lipschitz w.r.t. the norm itself, we have $\|\boldsymbol{x}\|_M^\star \le 1$ for all $\boldsymbol{x} \in \mathbb{R}^d$. By Assumption 3.3.1, it follows that

$$
\begin{aligned}
|I| &\le \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M \|\boldsymbol{v}_{\boldsymbol{x}_t - \boldsymbol{x}^\star}\|_M^\star \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) - c(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_M \\
&\le \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) - c(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_M
\end{aligned}
$$

For another thing, by the Item 2 in Lemma 3.4.4, $\left(1 + \frac{\lambda}{u_{\bar{p}}^2}\right) M(\boldsymbol{x}) \le \frac{1}{2}\|\boldsymbol{x}\|^2 \le \left(1 + \frac{\lambda}{l_{\bar{p}}^2}\right) M(\boldsymbol{x})$, which is equivalent to $\frac{l_{\bar{p}}}{\sqrt{l_{\bar{p}}^2 + \lambda}}\|\boldsymbol{x}\| \le \|\boldsymbol{x}\|_M \le \frac{u_{\bar{p}}}{\sqrt{u_{\bar{p}}^2 + \lambda}}\|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Hence,

$$
\begin{aligned}
&\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) - c(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_M \\
&\le \frac{u_{\bar{p}}}{\sqrt{u_{\bar{p}}^2 + \lambda}} \|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) - c(\boldsymbol{x}_t - \boldsymbol{x}^\star)\| \\
&\le \frac{c\gamma u_{\bar{p}}}{\sqrt{u_{\bar{p}}^2 + \lambda}} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \le c\gamma \cdot \frac{u_{\bar{p}}\sqrt{l_{\bar{p}}^2 + \lambda}}{l_{\bar{p}}\sqrt{u_{\bar{p}}^2 + \lambda}} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M.
\end{aligned}
$$

As a result,

$$
|I| \le 2c\gamma \frac{u_{\bar{p}}\sqrt{l_{\bar{p}}^2 + \lambda}}{l_{\bar{p}}\sqrt{u_{\bar{p}}^2 + \lambda}} M(\boldsymbol{x}_t - \boldsymbol{x}^\star). \tag{3.30}
$$

For the term II Since $\|\cdot\|_M$ is a convex function of $\boldsymbol{x}$, we have by the definition of convexity that $\|\boldsymbol{0}\|_M - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M \ge \langle \boldsymbol{v}_{\boldsymbol{x}_t - \boldsymbol{x}^\star}, -(\boldsymbol{x}_t - \boldsymbol{x}^\star)\rangle$. Hence,

$$
II = c\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M \langle \boldsymbol{v}_{\boldsymbol{x}_t - \boldsymbol{x}^\star}, \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle \ge c\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_M^2 = 2cM(\boldsymbol{x}_t - \boldsymbol{x}^\star). \tag{3.31}
$$

For the term III The the term III exists due to Markovian data. Note that $\boldsymbol{x}_t, \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star) \in \mathscr{G}_t$ and $\xi_{t-1} \in \mathscr{G}_t$. By Lemma 3.2.1 and Assumption 3.2.3, for any $t \ge 0$,

$$
\|\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t+1})|\xi_0 = \xi]\| = \|\mathscr{P}^{t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \le \kappa\rho^t \sup_{\xi \in \Xi}\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \le \kappa\sigma\rho^t.
$$

Therefore, we are motivated to define

$$
a_t = \lceil \log_\rho \frac{\eta_t}{\sigma\kappa} \rceil \text{ if } \rho > 0;\ a_t = 0 \text{ otherwise,} \tag{3.11}
$$

for each $t \geq 0$ such that for any $\xi \in \Xi$,

$$\|\mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\| = \|\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{a_t+1})|\xi_0 = \xi]\| = \|\mathbb{E}[\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)|\xi_{t-a_t-1} = \xi]\| \leq \eta_t, \tag{3.32}$$

where the last equality holds because we consider a time-homogeneous Markov chain. Then,

$$\begin{aligned}\mathbb{E}III = \mathbb{E}\langle\nabla\boldsymbol{M}(\boldsymbol{x}_t-\boldsymbol{x}^\star) - \nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\rangle \\ + \mathbb{E}\langle\nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star), \mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-a_t-1})\rangle.\end{aligned} \tag{3.33}$$

Let $\bar{q} \geq 1$ be the real number satisfying $\bar{q}^{-1} + \bar{p}^{-1} = 1$ for the given $\bar{p}$. By Hölder's inequality,

$$\begin{aligned}&|\langle\nabla\boldsymbol{M}(\boldsymbol{x}_t-\boldsymbol{x}^\star) - \nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\rangle| \\ &\quad\leq \|\nabla\boldsymbol{M}(\boldsymbol{x}_t-\boldsymbol{x}^\star) - \nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)\|_{\bar{q}}\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|_{\bar{p}} \\ &\quad\leq u_{\bar{p}} \cdot \|\nabla\boldsymbol{M}(\boldsymbol{x}_t-\boldsymbol{x}^\star) - \nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)\|_{\bar{q}}\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|.\end{aligned} \tag{3.34}$$

By the Item 1 in Lemma 3.4.4, we have

$$\|\nabla\boldsymbol{M}(\boldsymbol{x}_t-\boldsymbol{x}^\star)-\nabla\boldsymbol{M}(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)\|_{\bar{q}} \leq (\bar{p}-1)\lambda\cdot\|\boldsymbol{x}_t-\boldsymbol{x}_{t-a_t}\|_{\bar{p}} \leq (\bar{p}-1)u_{\bar{p}}\lambda\cdot\|\boldsymbol{x}_t-\boldsymbol{x}_{t-a_t}\|. \tag{3.35}$$

To proceed the proof, we introduce three useful lemmas in the following.

Lemma 3.4.5 (Properties of a_t's)**.** *Define* $\{a_t\}_{t\geq 0}$ *according to* (3.11)*. Under Assumption 3.2.5,it follows that* (i) $a_t = \mathcal{O}(\log t)$*,* (ii) $a_t\eta_{t-a_t}\log t = o(1)$ *when t goes to infinity, as a result of which, there exists* $K > 0$ *such that any* $t \geq K$*, we have*

$$Ma_t\eta_{t-a_t} \leq \log 2,$$

(iii) $\eta_{t-a_t}/\eta_t = \mathcal{O}(1)$*, and* (vi) $a_t \leq a_{t+1} \leq a_t + 1$ *for any sufficiently large t.*

Proof of Lemma 3.4.5. The proof can be found in Appendix C.1 □

Lemma 3.4.6. *With* $\{a_t\}_{t\geq 0}$ *defined in* (3.11)*, we introduce*

$$g_{t-1} = \begin{cases} \sup\limits_{t-a_t\leq\tau\leq t-1} g(\xi_\tau) & \textit{if } a_t \geq 1; \\ 0 & \textit{if } a_t = 0. \end{cases} \tag{3.36}$$

Then under Assumption 3.2.3, 3.2.4, 3.2.5 and 3.3.2, for any $t \geq K$*,*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-a_t}\| \leq 6Ma_t\eta_{t-a_t}(\|\boldsymbol{x}_t\| + g_{t-1}) \leq 2(\|\boldsymbol{x}_t\| + g_{t-1}).$$

Proof of Lemma 3.4.6. The proof can be found in Appendix C.2 □

Lemma 3.4.7. *Under Assumption 3.3.2, we have* $\mathbb{E}g_{t-1} \le (\mathbb{E}g_{t-1}^{\frac{p}{2}})^{\frac{2}{p}} = \mathcal{O}(a_t)$ *where* $\mathcal{O}(\cdot)$ *hides the linear dependence on* $\sup_{t\ge0}(\mathbb{E}|g(\xi_t)|^{\frac{p}{2}})^{\frac{2}{p}}$.

Proof of Lemma 3.4.7. The proof can be found in Appendix C.3 □

It then follows that for any $t \ge K$,

$$
\begin{aligned}
&|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star) - \nabla \boldsymbol{M}(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\rangle| \\
&\overset{(a)}{\le} (\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \|\boldsymbol{x}_t - \boldsymbol{x}_{t-a_t}\|\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\| \\
&\overset{(b)}{\le} 6\boldsymbol{M}(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot a_t\eta_{t-a_t}(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \|\boldsymbol{x}^\star\| + g_{t-1})\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\| \\
&\overset{(c)}{\le} 6\sigma\boldsymbol{M}(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot a_t\eta_{t-a_t}(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \|\boldsymbol{x}^\star\| + g_{t-1}) \\
&\overset{(d)}{\le} 6\sigma\boldsymbol{M}(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot a_t\eta_{t-a_t}\left(\frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + 1 + \|\boldsymbol{x}^\star\| + g_{t-1}\right) \\
&\overset{(e)}{\le} 6\sigma\boldsymbol{M}(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot a_t\eta_{t-a_t}\left(\left(1 + \frac{\lambda}{l_{\bar{p}}^2}\right)\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star) + \|\boldsymbol{x}^\star\| + g_{t-1} + 1\right),
\end{aligned}
\tag{3.37}
$$

where (a) follows from (3.34) and (3.35), (b) uses Lemma 3.4.6, (c) uses $\|\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\| \le \sigma$ from Assumption 3.2.2, (d) uses $x \le \frac{x^2+1}{2}$ for any $x \in \mathbb{R}$, and (e) uses the Item 2 in Lemma 3.4.4.

Notice that $\mathbf{0}$ is the unique minimizer of the smooth function $\boldsymbol{M}(\cdot)$, which implies $\nabla\boldsymbol{M}(\mathbf{0}) = \mathbf{0}$. Similarly, we have

$$
\begin{aligned}
&|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star), \mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-a_t-1})\rangle| \\
&\le (\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \|\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star\|\|\mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-a_t-1})\| \\
&\overset{(a)}{\le} \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \|\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star\| \\
&\le \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \left(\|\boldsymbol{x}_{t-a_t} - \boldsymbol{x}_t\| + \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|\right) \\
&\overset{(b)}{\le} \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \left(2\|\boldsymbol{x}_t\| + 2g_{t-1} + \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|\right) \\
&\le \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \left(3\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + 2(\|\boldsymbol{x}^\star\| + g_{t-1})\right) \\
&\le \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \left(\frac{3}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \frac{3}{2} + 2(\|\boldsymbol{x}^\star\| + g_{t-1})\right) \\
&\overset{(c)}{\le} 3\eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot \left(\left(1 + \frac{\lambda}{l_{\bar{p}}^2}\right)\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star) + (\|\boldsymbol{x}^\star\| + g_{t-1} + 1)\right),
\end{aligned}
\tag{3.38}
$$

where (a) follows from (3.32), (b) uses Lemma 3.4.6 and (c) uses the Item 2 in Lemma 3.4.4.

Combining (3.33), (3.37) and (3.38), we have for any $t \ge K$,

$$|\mathbb{E}III| \le 3(\bar{p}-1)u_{\bar{p}}^2\lambda \left[\left(\eta_t + 2M\sigma a_t \eta_{t-a_t}\right)\left(1+\frac{\lambda}{l_{\bar{p}}^2}\right)\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star) \right.$$
$$\left. +(\eta_t + 2M\sigma a_t \eta_{t-a_t})(\mathbb{E}g_{t-1} + \|\boldsymbol{x}^\star\| + 1)\right].$$

Putting them together Plugging the bounds for $\mathbb{E}I, \mathbb{E}II, \mathbb{E}III$ into (3.29) and combing the resulting inequality with (3.27) and (3.28), we have for any $t \ge K$,

$$\begin{aligned}\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star) &\le (1 + A_1\eta_t^2 + A_2 a_t\eta_t\eta_{t-a_t} - A_3\eta_t)\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star) \\ &\quad + A_4\eta_t^2 + A_5 a_t\eta_t\eta_{t-a_t} + (A_6\eta_t^2 + A_7 a_t\eta_t\eta_{t-a_t})\mathbb{E}g_{t-1},\end{aligned} \tag{3.39}$$

where for short we denote

$$A_1 = (\bar{p}-1)u_{\bar{p}}^2\lambda\left(3+2L_H^2\right)\left(\frac{\lambda}{l_{\bar{p}}^2}+1\right),\ A_2 = 6M\sigma(\bar{p}-1)u_{\bar{p}}^2\lambda\left(\frac{\lambda}{l_{\bar{p}}^2}+1\right),$$

$$A_3 = 2c\left(1-\gamma\frac{u_{\bar{p}}\sqrt{l_{\bar{p}}^2+\lambda}}{l_{\bar{p}}\sqrt{u_{\bar{p}}^2+\lambda}}\right),\ A_4 = u_{\bar{p}}^2(\bar{p}-1)\lambda\left(\sup_{t\ge0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\|^2 + 3(\|\boldsymbol{x}^\star\|+1)\right),$$

$$A_5 = 6M\sigma(\bar{p}-1)u_{\bar{p}}^2\lambda\left(1+\|\boldsymbol{x}^\star\|\right),\ A_6 = 3(\bar{p}-1)u_{\bar{p}}^2\lambda,\ A_7 = 6M(\bar{p}-1)u_{\bar{p}}^2\sigma\lambda.$$

Pay attention that by setting λ sufficiently small, we can ensure all A_i's are positive.

Dividing (3.39) by $a_t^2\eta_t$ and simplifying the inequality, we arrive at

$$(1+o(\eta_t))\cdot\frac{\mathbb{E}M(\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star)}{a_{t+1}^2\eta_{t+1}} \le (1 + A_1\eta_t^2 + A_2 a_t\eta_t\eta_{t-a_t} - A_3\eta_t)\frac{\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)}{a_t^2\eta_t} + \mathcal{O}(\eta_t),$$

where we use $\eta_{t+1} = \eta_t(1+o(\eta_t))$, $1 \le a_t \le a_{t+1}$ and $\eta_{t-a_t}/\eta_t = OM(1)$ in Lemma 3.4.5, and $\mathbb{E}g_{t-1} = \mathcal{O}(a_t)$ in Lemma 3.4.7. As long as t is sufficiently large, we have $\frac{1+A_1\eta_t^2+A_2a_t\eta_t\eta_{t-a_t}-A_3\eta_t}{1+o(\eta_t)} \le 1 - B_1\eta_t$ and there exist a constant positive $B_2 > 0$ such that

$$\frac{\mathbb{E}M(\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star)}{a_{t+1}^2\eta_{t+1}} \le \left(1 - B_1\eta_t\right)\frac{\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)}{a_t^2\eta_t} + B_2\eta_t.$$

Using the last inequality and Lemma A.10 in[67], we have

$$\sup_{t\ge0}\frac{\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)}{a_t^2\eta_t} < \infty.$$

By the Item 2 in Lemma 3.4.4, $\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)$ approximates $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$ up to constant factors. It implies $\sup_{t\ge0}\frac{\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2}{a_t^2\eta_t} < \infty$ and thus we establish the $(L^2, a_t\sqrt{\eta_t})$-consistency of $\{\boldsymbol{x}_t\}_{t\ge0}$.

3.4.3 Proof of Theorem 3.3.2

In this subsection, we further establish the $(L^p, a_t\sqrt{\eta_t})$-consistency. Though the main idea is similar to the case of $(L^2, a_t\sqrt{\eta_t})$-consistency, the proof procedure is much more circuitous for the following two reasons.

1. First, following the spirit of the generalized Lyapunov approach, we should consider the recursion of the form $\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star)^{\frac{p}{2}}$ where $\boldsymbol{x}_{t+1}$ is updated according to (3.27). However, $\mathbb{E}M(\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star)^{\frac{p}{2}}$ doesn't has a close-form expansion as the square counterpart $\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star)$. We then have to bound the incremental growth of $\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star)^{\frac{p}{2}}$ with respect to $\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\frac{p}{2}}$ via inequalities. To that end, we derive Lemma 3.4.8.

   Lemma 3.4.8. *For any scalar $A > 0$ and any real number $x \geq -A$, we have*

$$(A+x)^{1+\alpha} \leq \begin{cases} A^{1+\alpha} + (1+\alpha)A^\alpha x + |x|^{1+\alpha} & \text{if } \alpha \in (0,1], \\ A^{1+\alpha} + (1+\alpha)A^\alpha x + \frac{c_\alpha(1+\alpha)}{2}A^{\alpha-1}x^2 + c_\alpha|x|^{1+\alpha} & \text{if } \alpha \in [1,\infty). \end{cases} \tag{3.40}$$

   where c_α in a universal constant depending α and satisfying $\alpha \leq c_\alpha \leq 3^\alpha$.

   Proof of Lemma 3.4.8. The proof can be found in Appendix C.4. □

2. Second, according to (3.40), the specific value of α would affect the inequality we use. It implies we should proceed the proof in two cases.

Now, we formally start the proof. By (3.27), we obtain

$$M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star) \leq M(\boldsymbol{x}_t - \boldsymbol{x}^\star) - \eta_t\delta_t,$$

where

$$\delta_t := \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle - \frac{(\bar{p}-1)\lambda\eta_t}{2}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^2. \tag{3.41}$$

It is clear that $M(\boldsymbol{x}_t - \boldsymbol{x}^\star) - \eta_t\delta_t \geq M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star) \geq 0$. In the following, we set $\alpha = \frac{p}{2} - 1$ for short and have $\alpha > 0$ by assumption.

For the case of $\alpha \in (0,1]$ Taking $(1+\alpha)$-th order moment and using the first scalar inequality in (3.40), we have

$$\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star)^{1+\alpha} \leq \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} - (1+\alpha)\eta_t\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha\delta_t + \eta_t^{1+\alpha}\mathbb{E}|\delta_t|^{1+\alpha}. \tag{3.42}$$

To analyze the second and third term in (3.42), we establish corresponding upper bounds in Lemma 3.4.9 and Lemma 3.4.10.

Lemma 3.4.9. *Let* $d_t = \max_{t-a_t \le \tau \le t} \mathbb{E}M(\boldsymbol{x}_\tau - \boldsymbol{x}^\star)^{1+\alpha}$. *There exists a constant* $A_8 > 0$ *such that*

$$\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \delta_t \ge A_3 \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} - A_8(\eta_t + a_t\eta_{t-a_t})\left(d_t + a_t \cdot d_t^{\frac{\alpha}{1+\alpha}}\right).$$

Here A_8 *depends on* $A_1, A_2, A_5, A_6, \{\eta_t\}_{t\ge 0}$ *and* $\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^p$.

Proof of Lemma 3.4.9. The proof can be found in Appendix C.5. □

Lemma 3.4.10. *With* δ_t *defined in* (3.41)*, there exists a constant* $A_9 > 0$ *such that*

$$\mathbb{E}|\delta_t|^{1+\alpha} \le A_9(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} + \eta_t^{1+\alpha}).$$

Here A_9 *depends on* $\bar{p}, \lambda, u_{\bar{p}}, L_H$ *and* $\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^p$.

Proof of Lemma 3.4.10. The proof can be found in Appendix C.6. □

Denote $v_t = \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}$. With d_t defined in Lemma 3.4.9, we have $v_t \le d_t$ by definition. Plugging the bounds in Lemma 3.4.9 and Lemma 3.4.10 into (3.42), we have

$$v_{t+1} \le \left(1 - (1+\alpha)A_3\eta_t\right) v_t + (1+\alpha)A_8\eta_t(\eta_t + a_t\eta_{t-a_t})\left(d_t + a_t \cdot d_t^{\frac{\alpha}{1+\alpha}}\right) + A_9\eta_t^{1+\alpha}(v_t + \eta_t^{1+\alpha}).$$

We define $\tilde{d}_t = d_t a_t^{-2(1+\alpha)}\eta_t^{-(1+\alpha)}$ and similarly $\tilde{v}_t = v_t a_t^{-2(1+\alpha)}\eta_t^{-(1+\alpha)}$. For sufficiently large t, we would have $1-(1+\alpha)A_3\eta_t \in (0,1)$. Dividing $\eta_t^{1+\alpha}a_t^{2(1+\alpha)}$ on the both sides of the last inequality and using $\eta_{t+1} = \eta_t(1+o(\eta_t))$ and $1 \le a_t \le a_{t+1}$, we arrive at

$$\begin{aligned} \tilde{v}_{t+1}(1+o(\eta_t)) &\le \left[1 - (1+\alpha)A_3\eta_t + (1+\alpha)A_8\eta_t(\eta_t + a_t\eta_{t-a_t}) + A_9\eta_t^{1+\alpha}\right]\tilde{d}_t \\ &\quad + (1+\alpha)A_8\eta_t(1 + a_t\eta_{t-a_t}/\eta_t)a_t^{-1}\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + A_9\eta_t^{1+\alpha} \\ &\le \left[1 - (1+\alpha)A_3\eta_t(1+o(\eta_t))\right]\tilde{d}_t + \mathcal{O}(\eta_t)\cdot\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + A_9\eta_t^{1+\alpha} \end{aligned} \tag{3.43}$$

where the last equality uses $a_t\eta_{t-a_t} = o(1)$ and $\eta_{t-a_t}/\eta_t = \mathcal{O}(1)$ in Lemma 3.4.5.

We assert that

$$\sup_{t\ge 0}\tilde{d}_t < \infty.$$

We prove this statement in the following. For sufficiently large t, we have that $0 < \frac{1-(1+\alpha)A_3\eta_t}{1+o(\eta_t)} \le 1 - B_1\eta_t < 1$ for some constant $B_1 > 0$. Then we can find constants $B_2, B_3 > 0$ and simplify (3.43) as

$$\tilde{v}_{t+1} \le (1 - B_1\eta_t)\tilde{d}_t + B_2\eta_t\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + B_3\eta_t =: \tilde{d}_t - \eta_t h(\tilde{d}_t) \tag{3.44}$$

where $h(x) = B_1 x - B_2 x^{\frac{\alpha}{1+\alpha}} - B_3$ is a helper function. One can show that $h(x)$ is a function defined on $[0, \infty)$ that starts from a negative value, then decreases, and finally increases to infinity. As a result, there is a unique root $d^\star > 0$ such that $h(d^\star) = 0$. With a sufficiently large t, one has $a_t \le a_{t+1} \le a_t + 1$ from Lemma 3.4.5. If $\tilde{d}_t \ge d^\star$, we then have $h(\tilde{d}_t) \ge 0$ and thus $\tilde{v}_{t+1} \le \tilde{d}_t$ from (3.44). As a result of the fact $t + 1 - a_{t+1} \ge t - a_t$, we have

$$\begin{aligned}\tilde{d}_{t+1} &= \max_{t+1-a_{t+1} \le \tau \le t+1} \tilde{v}_t \le \max_{t-a_t \le \tau \le t+1} \tilde{v}_t = \max\left\{ \max_{t-a_t \le \tau \le t} \tilde{v}_t, \tilde{v}_{t+1} \right\} = \max\left\{ \tilde{d}_t, \tilde{v}_{t+1} \right\} \\ &\le \tilde{d}_t.\end{aligned} \tag{3.45}$$

In short, once $\tilde{d}_t \ge d^\star$, $\tilde{d}_{t+1}$ decreases until it is smaller than $d^\star$. Furthermore, if $\tilde{d}_t < d^\star$ and $\tilde{d}_{t+1} \ge d^\star$, from (3.45), we have $\tilde{d}_{t+1} \le \tilde{v}_{t+1}$, which, together with (3.44), implies $\tilde{d}_{t+1} - \tilde{d}_t$ is bounded by a universal constant. As a result, we conclude that $\tilde{d}_t$ is impossible to reach infinity, and thus $\sup_{t \ge 0} \tilde{d}_t < \infty$.

Given $\sup_{t\ge0} \tilde{d}_t < \infty$ and $\frac{p}{2} = 1 + \alpha$, we have that $\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\frac{p}{2}} \le C_p \eta_t^{\frac{p}{2}} a_t^p$ uniformly for $t \ge 0$ and a universal constant $C_p > 0$. By Lemma 3.4.4, we have $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \lesssim M(\boldsymbol{x}_t - \boldsymbol{x}^\star)$. As a result, we have $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p \le C_p \eta_t^{\frac{p}{2}} a_t^p$ (by slightly abusing the notation C_p).

For the case of $\alpha \in (1, \infty)$ Taking $(1 + \alpha)$-th order moment and using the second scalar inequality in (3.40), we have

$$\begin{aligned}\mathbb{E}M(\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star)^{1+\alpha} \le \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} &- (1+\alpha)\eta_t \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \delta_t \\ &+ \frac{c_\alpha(1+\alpha)}{2} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1} \eta_t^2 |\delta_t|^2 + c_\alpha \eta_t^{1+\alpha} \mathbb{E}|\delta_t|^{1+\alpha}.\end{aligned} \tag{3.46}$$

Because most of the terms in (3.46) have been analyzed previously, we only focus on the remaining term $\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\delta_t|^2$.

Lemma 3.4.11. *There exists a positive constant $A_{10} > 0$ such that*

$$\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\delta_t|^2 \le A_{10}\left[\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha+1} + \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}\right)^{\frac{\alpha}{\alpha+1}} + \eta_t^2 (\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha-1}{\alpha+1}}\right].$$

Here A_{10} depends on $L_H, M, \|\boldsymbol{x}^\star\|, \lambda, l_{\bar{p}}, \sup_{t\ge0} \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_t)\|^p$ and $\sup_{t\ge0} \mathbb{E}|g(\xi_t)|^p$.

Proof of Lemma 3.4.11. The proof can be found in Appendix C.7. □

Plugging these bounds in Lemma 3.4.9, Lemma 3.4.10, and Lemma 3.4.11 into (3.46), we have

$$v_{t+1} \le \left(1 - (1+\alpha)A_3\eta_t\right) v_t + (1+\alpha)A_8 \eta_t(\eta_t + a_t \eta_{t-a_t}) \left(d_t + a_t \cdot d_t^{\frac{\alpha}{1+\alpha}}\right)$$

$$+ (1+\alpha)c_\alpha A_{10}\eta_t^2 \left[v_t + v_t^{\frac{\alpha}{1+\alpha}} + \eta_t^2 v_t^{\frac{\alpha-1}{1+\alpha}}\right] + c_\alpha A_9 \eta_t^{1+\alpha}(v_t + \eta_t^{1+\alpha}).$$

Recall that $\tilde{d}_t = d_t a_t^{-2(1+\alpha)}\eta_t^{-(1+\alpha)}$ and $\tilde{v}_t = v_t a_t^{-2(1+\alpha)}\eta_t^{-(1+\alpha)}$. For simplicity, we let $\mathcal{O}(\cdot)$ hide positive constant factors. Then, dividing $\eta_t^{1+\alpha}a_t^{2(1+\alpha)}$ on the both sides of the last equation and using $\eta_{t+1} = \eta_t(1+o(\eta_t))$ and $1 \le a_t \le a_{t+1}$, we arrive at

$$\begin{aligned}
\tilde{v}_{t+1}(1+o(\eta_t)) &\le \left[1-(1+\alpha)A_3\eta_t + \mathcal{O}(\eta_t)\cdot(\eta_t + a_t\eta_{t-a_t}) + \mathcal{O}(\eta_t^{1+\alpha})\right]\tilde{d}_t \\
&\quad + \mathcal{O}(\eta_t)\cdot(1 + a_t\eta_{t-a_t}/\eta_t)a_t^{-1}\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + \mathcal{O}(\eta_t^2)\cdot\tilde{d}_t^{\frac{\alpha-1}{\alpha+1}} + \mathcal{O}(\eta_t^{1+\alpha}) \\
&\overset{(a)}{\le} \left[1-(1+\alpha)A_3\eta_t(1+o(\eta_t))\right]\tilde{d}_t + \mathcal{O}(\eta_t)\cdot\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + \mathcal{O}(\eta_t^2)\cdot\tilde{d}_t^{\frac{\alpha-1}{\alpha+1}} + \mathcal{O}(\eta_t^{1+\alpha}) \\
&\overset{(b)}{\le} \left[1-(1+\alpha)A_3\eta_t(1+o(\eta_t))\right]\tilde{d}_t + \mathcal{O}(\eta_t)\cdot\tilde{d}_t^{\frac{\alpha}{1+\alpha}} + \mathcal{O}(\eta_t^{1+\alpha}), \qquad (3.47)
\end{aligned}$$

where (a) uses $a_t\eta_{t-a_t} = o(1)$ and $\eta_{t-a_t}/\eta_t = \mathcal{O}(1)$ in Lemma 3.4.5 and (b) follows because we can assume $\tilde{d}_t \ge 1$ without loss of generality (which can be achieved by redefining $\tilde{d}_t \leftarrow \max\{\tilde{d}_t, 1\}$).

For sufficiently large t, we can find positive constants $B_1, B_2, B_3 > 0$ such that

$$\tilde{v}_{t+1} \le (1 - B_1\eta_t)\tilde{d}_t + B_2\eta_t\tilde{d}_t^{\frac{\alpha}{\alpha+1}} + B_3\eta_t,$$

which is the inequality we have already analyzed in (3.44). By an identical argument therein, we conclude $\sup_{t\ge 0}\tilde{d}_t < \infty$. Therefore, we also have $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^p \le C_p\eta_t^{\frac{p}{2}}a_t^p$ when $\alpha > 1$.

□

3.4.4 Proof of Theorem 3.3.3

In literature, the semiparametric efficiency for empirical estimators has been well understood when the interest of (unknown) parameter is in an expectation form, i.e., $\mathbb{E}_{\xi\sim\pi}\ell(\xi)$ for a function ℓ. However, our interest parameter is $\boldsymbol{x}^\star$, which is the root of the equation $\boldsymbol{g}(\boldsymbol{x}) := \mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x},\xi) = \boldsymbol{0}$. To make use of the existing result, we provide the following lemma to serve as a bridge.

Lemma 3.4.12. *For any RAL estimator $\boldsymbol{T}_n$ for $\boldsymbol{x}^\star$ on π with limit $\boldsymbol{L}$, we have that $\mathbf{g}(\boldsymbol{T}_n)$ is an RAL estimator for $-\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)$ with limit $\boldsymbol{G}\cdot\boldsymbol{L}$.*

For any RAL estimator $\boldsymbol{T}_n$ for $\boldsymbol{x}^\star$ on π with limit $\boldsymbol{L}$, Lemma 3.4.12 shows the transformed estimator $\mathbf{g}(\boldsymbol{T}_n)$ is an RAL estimator for $-\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)$ with limit $\boldsymbol{G}\cdot\boldsymbol{L}$. Because $-\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)$ is a parameter in the expectation form, the Markovian convolution theorem presented in Greenwood, Wefelmeyer[155] shows that $\boldsymbol{G}\cdot\boldsymbol{L}$ can be represented as $\mathbf{M}+\mathbf{N}$, where $\mathbf{M}$ is independent of $\mathbf{N}$ and $\mathbf{N}$ is Gaussian distributed with zero mean and the covariance $\mathbb{E}_{\xi\sim\pi}\left[(\mathscr{I}-\mathscr{P})^{-1}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\right]\left[(\mathscr{I}-\mathscr{P})^{-1}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\right]^\top$. By Lemma 3.2.2, we know the matrix is exactly $\boldsymbol{S}$. Therefore, under $\boldsymbol{P}$, $\mathrm{Var}(\mathbf{N})\succeq\boldsymbol{S}$ and thus

$$\lim_{n\to\infty} n\mathbb{E}(\boldsymbol{T}_n-\boldsymbol{x}^\star)(\boldsymbol{T}_n-\boldsymbol{x}^\star)^\top=\mathrm{Var}(\boldsymbol{L})=\boldsymbol{G}^{-1}\mathrm{Var}(\mathbf{M}+\mathbf{N})\boldsymbol{G}^{-\top}\succeq\boldsymbol{G}^{-1}\mathrm{Var}(\mathbf{N})\boldsymbol{G}^{-\top}=\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}.$$

At the end of this subsection, we provide the proof for Lemma 3.4.12.

Proof of Lemma 3.4.12. By the definition of RAL estimators, we need to check asymptotic linearity and regularity. We denote by $\mathscr{P}$ and $\mathscr{P}_{nh}$ forward operator of the transition kernels $\boldsymbol{P}$ and $\boldsymbol{P}_{nh}$ respectively.

Asymptotic linearity From the regularity of $\boldsymbol{T}_n$, we have $\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{x}^\star)\xrightarrow{d}\boldsymbol{L}$ under $\boldsymbol{P}$. It implies that $\boldsymbol{T}_n\xrightarrow{p}\boldsymbol{x}^\star$ and $\mathbb{E}_{P^n}\|\boldsymbol{T}_n-\boldsymbol{x}^\star\|^2\lesssim\frac{1}{n}$. By Assumption 3.2.1, it holds that

$$\|\sqrt{n}\left(\mathbf{g}(\boldsymbol{T}_n)-\mathbf{g}(\boldsymbol{x}^\star)\right)-\sqrt{n}\boldsymbol{G}(\boldsymbol{T}_n-\boldsymbol{x}^\star)\|\le L_G\sqrt{n}\|\boldsymbol{T}_n-\boldsymbol{x}^\star\|^2+o_{P^n}(1)=o_{P^n}(1).$$

By the asymptotic linearity of $\boldsymbol{T}_n$, we have $\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{x}^\star)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{\varphi}(\xi_{i-1},\xi)+o_{P^n}(1)$ and thus

$$\sqrt{n}(\mathbf{g}(\boldsymbol{T}_n)-\mathbf{g}(\boldsymbol{x}^\star))=\sqrt{n}\mathbf{g}(\boldsymbol{T}_n)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{G}\boldsymbol{\varphi}(\xi_{i-1},\xi_i)+o_{P^n}(1).$$

Regularity We first control the sum $\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)+\mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)$. By using the fact $\mathbb{E}_{\sim\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)=\mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)=\mathbf{0}$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)+\mathbb{E}_{\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\\
=&\mathbb{E}_{\xi\sim\pi}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]-\mathbb{E}_{\xi\sim\pi_{nh}}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]\\
\overset{(a)}{=}&\mathbb{E}_{\xi\sim\pi}\mathscr{P}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]-\mathbb{E}_{\xi\sim\pi_{nh}}\mathscr{P}_{nh}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]\\
\overset{(b)}{=}&\mathbb{E}_{\xi\sim\pi}\mathscr{P}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]-\mathbb{E}_{\xi\sim\pi_{nh}}\mathscr{P}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)]\\
&\qquad+\frac{1}{\sqrt{n}}\mathbb{E}_{\xi'\sim\pi_{nh}}\mathbb{E}_{\xi\sim P(\xi,\cdot)}\boldsymbol{h}(\xi,\xi')[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi')-\boldsymbol{H}(\boldsymbol{x}^\star,\xi')]\\
=&\int_\Xi\mathscr{P}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi)](\pi(d\xi)-\pi_{nh}(d\xi))\\
&\qquad+\frac{1}{\sqrt{n}}\mathbb{E}_{\xi'\sim\pi_{nh}}\mathbb{E}_{\xi\sim P(\xi,\cdot)}\boldsymbol{h}(\xi,\xi')[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star,\xi')-\boldsymbol{H}(\boldsymbol{x}^\star,\xi')]\\
:=&\mathscr{Z}_1+\mathscr{Z}_2.
\end{aligned}$$

where (a) uses the fact that π and π_{nh} are stationary distributions of $\boldsymbol{P}$ and $\boldsymbol{P}_{nh}$ and (b) uses the choice $\boldsymbol{P}_{nh}(\xi, d\xi') = \boldsymbol{P}(\xi, d\xi')\left(1 + \frac{1}{\sqrt{n}}\boldsymbol{h}(\xi, \xi')\right)$.

We then bound the two term $\mathscr{L}_1$ and $\mathscr{L}_2$ respectively. By the boundedness of $\boldsymbol{h}$ and Assumption 3.2.3, we have

$$\|\mathscr{L}_2\| \lesssim \frac{1}{\sqrt{n}}\mathbb{E}_{\xi\sim\pi_{nh}}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_{nh}^\star, \xi) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \lesssim \frac{1}{\sqrt{n}}\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\|.$$

On the other hand, from Assumption 3.2.4, both of the transition kernels $\boldsymbol{P}$ and $\boldsymbol{P}_{nh}$ are *strongly stable* which is defined in Kartashov [156]. By Theorem 3 in Kartashov [156], it follows that $d_{\text{TV}}(\pi, \pi_{nh}) \lesssim \sup_{\xi\in\Xi} d_{\text{TV}}(\boldsymbol{P}(\xi, \cdot), \boldsymbol{P}_{nh}(\xi, \cdot))$. Therefore,

$$\begin{aligned}
\|\mathscr{L}_1\| &\le \int_{\Xi} \|\mathscr{P}[\boldsymbol{H}(\boldsymbol{x}_{nh}^\star, \xi) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi)]\| \cdot |\pi(d\xi) - \pi_{nh}(d\xi)| \\
&\le \sup_{\xi\in\Xi} \mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_{nh}^\star, \xi) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| \cdot d_{\text{TV}}(\pi, \pi_{nh}) \\
&\overset{(a)}{\lesssim} \|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\| \cdot \sup_{\xi\in\Xi} d_{\text{TV}}(\boldsymbol{P}(\xi, \cdot), \boldsymbol{P}_{nh}(\xi, \cdot)) \\
&\overset{(b)}{\le} \|\boldsymbol{x}_{nh}^\star - x^\star\| \cdot \frac{1}{\sqrt{n}} \sup_{\xi,\xi'\in\Xi} \|\boldsymbol{h}(\xi, \xi')\| \lesssim \frac{1}{\sqrt{n}}\|\boldsymbol{x}_{nh}^\star - x^\star\|,
\end{aligned}$$

where (a) follows from Assumption 3.2.3 and (b) follows from the definition of $\boldsymbol{P}_{nh}(\xi, d\xi')$. Combining these two bounds, we get that

$$\mathbf{g}(\boldsymbol{x}_{nh}^\star) + \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) = \mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}_{nh}^\star, \xi) + \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) \lesssim \frac{\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\|}{\sqrt{n}}.$$

Noting that $\mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) = \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) - \mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)$, we can show $\mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) = \mathcal{O}(1/\sqrt{n})$ by using the same technique in bounding $\|\mathscr{L}_1\|$. Using the last inequality, we have

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{x}^\star) - \boldsymbol{g}(\boldsymbol{x}_{nh}^\star) &= -\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}_{nh}^\star, \xi) - \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) + \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) \\
&= \mathbb{E}_{\xi\sim\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\|\right).
\end{aligned}$$

By Assumption 3.3.1, it is easy to show that $(1-\gamma)c\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\| \le \|\mathbf{g}(\boldsymbol{x}_{nh}^\star) - \mathbf{g}(\boldsymbol{x}^\star)\|$. Hence,

$$\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\| \le \|\mathbb{E}_{\pi_{nh}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\| + \mathcal{O}\left(\frac{1}{\sqrt{n}}\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\|\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\|\right).$$

Rearranging the last inequality yields $\|\boldsymbol{x}_{nh}^\star - \boldsymbol{x}^\star\| \lesssim \frac{1}{\sqrt{n}}$.

Finally, by the regularity of $\boldsymbol{T}_n$, we have $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{x}_{nh}^\star) \overset{d}{\to} \boldsymbol{L}$ under the perturbed distribution $\boldsymbol{P}_{\boldsymbol{nh}}$ and thus $\boldsymbol{T}_n \overset{p}{\to} \boldsymbol{x}^\star$. Therefore,

$$
\begin{aligned}
&\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) + \mathbb{E}_{\xi\sim\pi_{\boldsymbol{nh}}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi))\\
=&\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{x}^\star_{\boldsymbol{nh}})) + \sqrt{n}\left(\mathbf{g}(\boldsymbol{x}^\star_{\boldsymbol{nh}}) + \mathbb{E}_{\xi\sim\pi_{\boldsymbol{nh}}}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)\right)\\
=&\sqrt{n}[(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{x}^\star)) - (\boldsymbol{g}(\boldsymbol{x}^\star_{\boldsymbol{nh}}) - \boldsymbol{g}(\boldsymbol{x}^\star))] + \mathcal{O}(\|\boldsymbol{x}^\star_{\boldsymbol{nh}} - \boldsymbol{x}^\star\|)\\
=&\sqrt{n}[\boldsymbol{G}(\boldsymbol{T}_n - \boldsymbol{x}^\star) - \boldsymbol{G}(\boldsymbol{x}^\star_{\boldsymbol{nh}} - \boldsymbol{x}^\star)] + \mathcal{O}\left(\sqrt{n}\|\boldsymbol{T}_n - \boldsymbol{x}^\star_{\boldsymbol{nh}}\|^2 + \sqrt{n}\|\boldsymbol{x}^\star_{\boldsymbol{nh}} - \boldsymbol{x}^\star\|^2\right) + \mathcal{O}(1/\sqrt{n})\\
=&\boldsymbol{G}\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{x}^\star_{\boldsymbol{nh}}) + \mathcal{O}_{\boldsymbol{P}^n}(1/\sqrt{n}) = \boldsymbol{G}\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{x}^\star_{\boldsymbol{nh}}) + o_{\boldsymbol{P}^n}(1)\\
\overset{d}{\to}&\boldsymbol{G}\cdot\boldsymbol{L}.
\end{aligned}
$$

The last equation means $\boldsymbol{g}(\boldsymbol{T}_n)$ is a regular estimator for $\mathbb{E}_{\xi\sim\pi}\boldsymbol{H}(\boldsymbol{x}^\star, \xi)$ with limit $\boldsymbol{GL}$. □

3.4.5 Preliminaries on the Lévy-Prokhorov Metric

Before presenting the proof of Theorem 3.3.5, we introduce additional preliminaries and notation. We relate the Lévy-Prokhorov metric $d_{\mathrm{S}}(\cdot)$ in (3.13) with a Ky-Fan-metric-type functional $\tilde{d}(\cdot)$ that would be frequently used latter on. For any continuous stochastic process $\boldsymbol{\phi} \in \mathsf{D}_{[0,1],\mathbb{R}^d}$, we denote

$$\tilde{d}(\boldsymbol{\phi}) := \inf_{\varepsilon\ge 0} \varepsilon \vee \mathbb{P}(|||\boldsymbol{\phi}||| > \varepsilon). \tag{3.48}$$

Proposition 3.4.1. *For any $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathsf{D}_{[0,1],\mathbb{R}^d}$, it then follows that*

$$d_{\mathrm{P}}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2, \boldsymbol{\phi}_1) \le \tilde{d}(\boldsymbol{\phi}_2).$$

Proof of Proposition 3.4.1. For each $\boldsymbol{\phi}_2$, we assume the maximum in $\tilde{d}(\boldsymbol{\phi}_2)$ is achieved by ε_2 such that $\tilde{d}(\boldsymbol{\phi}_2) = \varepsilon_2 \vee \mathbb{P}(\left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right| > \varepsilon_2)$. It is obvious that $\varepsilon_2 \le \tilde{d}(\boldsymbol{\phi}_2)$. Recall that $B^\varepsilon := \{\boldsymbol{\phi}_1 : \inf_{\boldsymbol{\phi}_2 \in B} d_{\mathrm{S}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \le \varepsilon\}$. Then, for any $B \in \mathscr{D}_{[0,1],\mathbb{R}^d}$, once $\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 \in B$ and $\left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right| \le \varepsilon$, we have $\boldsymbol{\phi}_1 \in B^\varepsilon$. Therefore,

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 \in B) &= \mathbb{P}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 \in B, \left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right| \le \varepsilon_2) + \mathbb{P}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 \in B, \left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right| > \varepsilon_2)\\
&\le \mathbb{P}(\boldsymbol{\phi}_1 \in B^{\varepsilon_2}) + \mathbb{P}(\left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right| > \varepsilon_2)\\
&\le \mathbb{P}(\boldsymbol{\phi}_1 \in B^{\tilde{d}(\boldsymbol{\phi}_2)}) + \tilde{d}(\boldsymbol{\phi}_2).
\end{aligned}
$$

By taking B as all measurable set in $\mathscr{D}_{[0,1],\mathbb{R}^d}$, we conclude that $d_{\mathrm{P}}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2, \boldsymbol{\phi}_1) \le \tilde{d}(\boldsymbol{\phi}_2)$ by the definition of d_{P} in (3.13). □

Proposition 3.4.2. *Let $g : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathsf{D}_{[0,1],\mathbb{R}^k} (k \ge 1)$ be L-Lipschitz continuous in $|||\cdot|||$ in the sense that $\left|\left|\left|f(\boldsymbol{\phi}_1) - f(\boldsymbol{\phi}_2)\right|\right|\right| \le L \cdot \left|\left|\left|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\right|\right|\right|$ for any $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathsf{D}_{[0,1],\mathbb{R}^d}$. For any*

$\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathsf{D}_{[0,1],\mathbb{R}^d}$, *it follows that*

$$d_{\mathrm{P}}(g(\boldsymbol{\phi}_1), g(\boldsymbol{\phi}_2)) \le (L \vee 1) \cdot d_{\mathrm{P}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2).$$

Proof of Proposition 3.4.2. Let $\tilde{B}$ be any measurable Borel set in $\mathbb{R}^k$ and we define $B = \{\boldsymbol{\phi} \in \mathsf{D}_{[0,1],\mathbb{R}} : g(\boldsymbol{\phi}) \in \tilde{B}\}$. Let $\varepsilon = d_{\mathrm{P}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$. By definition, we have $\mathbb{P}(\boldsymbol{\phi}_1 \in B) \le \mathbb{P}(\boldsymbol{\phi}_2 \in B^{\varepsilon}) + \varepsilon$. Notice that $\mathbb{P}(\boldsymbol{\phi}_1 \in B) = \mathbb{P}(g(\boldsymbol{\phi}_1) \in \tilde{B})$ and $\mathbb{P}(\boldsymbol{\phi}_2 \in B^{\varepsilon}) \le \mathbb{P}(g(\boldsymbol{\phi}_2) \in \tilde{B}^{L\varepsilon})$.

The second inequality uses the result that if $\boldsymbol{\phi}_2 \in B^{\varepsilon}$, then there exists $\boldsymbol{\phi}_3 \in B$ such that $g(\boldsymbol{\phi}_3) \in B$ and $\left|\left|\left|\boldsymbol{\phi}_2 - \boldsymbol{\phi}_3\right|\right|\right| \le \varepsilon$. Therefore, $\left|\left|\left|g(\boldsymbol{\phi}_2) - g(\boldsymbol{\phi}_3)\right|\right|\right| \le L \cdot \left|\left|\left|\boldsymbol{\phi}_2 - \boldsymbol{\phi}_3\right|\right|\right| \le L \cdot \varepsilon$ and thus $g(\boldsymbol{\phi}_2) \in \tilde{B}^{L\varepsilon}$. Hence, by arbitrariness of $\tilde{B}$, $d_{\mathrm{P}}(g(\boldsymbol{\phi}_1), g(\boldsymbol{\phi}_2)) \le (L \vee 1) \cdot d_{\mathrm{P}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$. □

As a direct corollary of Proposition 3.4.2, we have

Corollary 3.4.1. *For any vector $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\theta}\|_* = 1$,*

$$d_{\mathrm{P}}(\boldsymbol{\theta}^{\top}\boldsymbol{\phi}_1, \boldsymbol{\theta}^{\top}\boldsymbol{\phi}_2) \le d_{\mathrm{P}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2).$$

Proposition 3.4.3. *If $\boldsymbol{\phi} \in \mathsf{D}_{[0,1],\mathbb{R}^d}$ satisfies $\mathbb{E}|||\boldsymbol{\phi}|||^p < \infty$ with $p > 0$, then*

$$\tilde{d}(\boldsymbol{\phi}) \le \left(\mathbb{E}|||\boldsymbol{\phi}|||^p\right)^{\frac{1}{p+1}}.$$

Proof of Proposition 3.4.3. With $\varepsilon = \left(\mathbb{E}|||\boldsymbol{\phi}|||^p\right)^{\frac{1}{p+1}}$, Markov's inequality yields that $\mathbb{P}(|||\boldsymbol{\phi}||| > \varepsilon) \le \frac{\mathbb{E}|||\boldsymbol{\phi}|||^p}{\varepsilon^p} = \varepsilon$. Hence, $\tilde{d}(\boldsymbol{\phi}) \le \varepsilon \vee \mathbb{P}(|||\boldsymbol{\phi}||| > \varepsilon) = \varepsilon = \left(\mathbb{E}|||\boldsymbol{\phi}|||^p\right)^{\frac{1}{p+1}}$. □

Proposition 3.4.1 shows that the Lévy-Prokhorov metric between $\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_1$ is exactly bounded by $\tilde{d}(\boldsymbol{\phi}_2)$. Proposition 3.4.3 then implies $\tilde{d}(\boldsymbol{\phi}_2)$ is further bounded by $\left(\mathbb{E}\left|\left|\left|\boldsymbol{\phi}_2\right|\right|\right|^p\right)^{\frac{1}{p+1}}$ if the p-th order moment exists. In this way, we reduce the Lévy-Prokhorov metric between two given random processes to the high-order moments of their difference. The latter is more tractable and thus easier to analyze.

Theorem 3.4.1 (Corollary 1 in Kubilius [150])**.** *Let (X^n, F^n) be a sequence of locally square integrable martingales in $\mathbb{R}$, and (X, F) be a continuous Gaussian martingale. Then for any $T > 0$, and $0 < \delta < 3/2$,*

$$\begin{aligned} d_{\mathrm{P}}\left(X^n, X\right) = \mathscr{O}\Bigg(&\left\{\left(\mathbb{E}\sup_{t\leqslant T}\left|\langle X^n\rangle_t - \langle X\rangle_t\right|\right)^{\frac{1}{3}} + \left(\mathbb{E}\int_0^T\int_{\mathbb{R}}|x|^{2+2\delta}\Pi^n(ds,dx)\right)^{\frac{1}{3+2\delta}}\right\} \\ &\times\left|\ln\left(\mathbb{E}\sup_{t\leqslant T}\left|\langle X^n\rangle_t - \langle X\rangle_t\right| + \mathbb{E}\int_0^T\int_{\mathbb{R}}|x|^{2+2\delta}\Pi^n(ds,dx)\right)\right|^{1/2}\Bigg), \end{aligned} \tag{3.49}$$

where $\langle X \rangle$ is the quadratic characteristic of X and Π^n is the dual predictable projection of the process X^n.

3.4.6 Proof of Theorem 3.3.5

With the preliminaries in the previous subsection, we are ready to prove Theorem 3.3.5.

Proof of Theorem 3.3.5. Let $p = 2(1+\delta)$ for simplicity. Then $p > 2$ is equivalent to $\delta > 0$.

Step one: Finer partial-sum process decomposition We have analyzed the partial-sum decomposition in Section 3.4.1. We will further decompose two terms to proceed proof. Recall that $\tilde{\boldsymbol{\phi}}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}(\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^\star)$ and $\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta_t \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})$. We directly quote the result (3.20) here

$$\begin{aligned}\tilde{\boldsymbol{\phi}}_T(r) - \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{u}_t &= \frac{1}{\sqrt{T}\eta_0}\boldsymbol{A}_0^{\lfloor Tr\rfloor}\boldsymbol{B}_0\boldsymbol{\Delta}_0 + \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{A}_t^{\lfloor Tr\rfloor}(\boldsymbol{r}_t + \boldsymbol{v}_t)\\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\left(\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\right)\boldsymbol{u}_t + \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\left(\boldsymbol{A}_t^{\lfloor Tr\rfloor} - \boldsymbol{A}_t^T\right)\boldsymbol{u}_t\\ &:= \boldsymbol{\psi}_0(r) + \boldsymbol{\psi}_1(r) + \boldsymbol{\psi}_2(r) + \boldsymbol{\psi}_3(r).\end{aligned}\tag{3.20}$$

First, we further decompose $\boldsymbol{\psi}_1(r) := \boldsymbol{\psi}_{1,1}(r) + \boldsymbol{\psi}_{1,2}(r)$ into two terms and arrive at

$$\boldsymbol{\psi}_1(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{A}_t^{\lfloor Tr\rfloor}(\boldsymbol{r}_t+\boldsymbol{v}_t) := \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{A}_t^{\lfloor Tr\rfloor}\boldsymbol{r}_t + \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{A}_t^{\lfloor Tr\rfloor}\boldsymbol{v}_t =: \boldsymbol{\psi}_{1,1}(r)+\boldsymbol{\psi}_{1,2}(r).$$

Second, we decompose the noise $\boldsymbol{u}_t = \boldsymbol{u}_{t,1} + \boldsymbol{u}_{t,2}$ where

$$\begin{aligned}\boldsymbol{u}_{t,1} &= \left[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\right] - \left[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\right],\\ \boldsymbol{u}_{t,2} &= \left[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\right].\end{aligned}\tag{B.1}$$

This decomposition has been used to analyze the asymptotic behavior of $\frac{1}{\sqrt{T}}\sum_{t=0}^{T}\boldsymbol{G}^{-1}\boldsymbol{u}_t$ in Lemma 3.4.1. From the proof of 2 in Lemma 3.4.1, we know that both $\{\boldsymbol{u}_{t,1}\}_{t\ge 0}$ and $\{\boldsymbol{u}_{t,2}\}_{t\ge 0}$ are martingale difference sequences with bounded $(2+2\delta)$-th order moment. For simplicity, we denote

$$\boldsymbol{\psi}_{4,1}(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{u}_{t,1},\quad \boldsymbol{\psi}_{4,2}(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2},\quad \text{and}\quad \boldsymbol{\psi}(r) = \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}(r).$$

Therefore, it follows that

$$\boldsymbol{\phi}_T = (\boldsymbol{\phi}_T - \tilde{\boldsymbol{\phi}}_T) + \boldsymbol{\psi}_0 + \boldsymbol{\psi}_{1,1} + \boldsymbol{\psi}_{1,2} + \boldsymbol{\psi}_2 + \boldsymbol{\psi}_3 + \boldsymbol{\psi}_{4,1} + \boldsymbol{\psi}_{4,2}.$$

By repeatedly using Proposition 3.4.1 and Corollary 3.4.1, it follows that for any $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\theta}\|_* = 1$,

$$\begin{aligned} d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}) \le d_{\mathrm{P}}\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \tilde{\boldsymbol{\phi}}_T\right) + \tilde{d}(\boldsymbol{\psi}_0) + \tilde{d}(\boldsymbol{\psi}_{1,1}) + \tilde{d}(\boldsymbol{\psi}_{1,2}) \\ + \tilde{d}(\boldsymbol{\psi}_2) + \tilde{d}(\boldsymbol{\psi}_3) + \tilde{d}(\boldsymbol{\psi}_{4,1}) + d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\psi}_{4,2}, \boldsymbol{\theta}^\top \boldsymbol{\psi}). \end{aligned} \tag{3.50}$$

Step two: Moment analysis By Proposition 3.4.3, each $\tilde{d}(\boldsymbol{\psi})$ is bounded by the moment $\left(\mathbb{E}|||\boldsymbol{\psi}|||^{v}\right)^{\frac{1}{v+1}}$ for any $1 \le v \le p$. Therefore, analyzing most of the terms in the right-hand side of (3.50) is reduced to analyze their higher-order moment with the moment order v unspecified as a variable. Lemma 3.4.13 provides these higher order moment bounds with λ, m, l, k the corresponding variables. Given the interested parameters include only t_{mix} and T, we will hide other parameter dependence in $\lesssim$, $\mathcal{O}$ and $\tilde{\mathcal{O}}$.

Lemma 3.4.13. *Rewrite $p = 2(1+\delta)$. Under the assumptions of Theorem 3.3.5, it follows that*

$$d_{\mathrm{P}}\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \tilde{\boldsymbol{\phi}}_T\right) = \mathcal{O}\left(t_{\mathrm{mix}}^{\frac{1+m}{2+m}} \cdot T^{\frac{1+m}{2(2+m)}} \cdot \left(\frac{1}{T}\sum_{t=0}^{T} \eta_t^{1+m}\right)^{\frac{1}{2+m}}\right) \quad \forall m \in [0, 2\delta+1], \tag{3.51}$$

$$\tilde{d}(\boldsymbol{\psi}_0) = \mathcal{O}\left(T^{-\frac{1}{2}}\right), \tag{3.52}$$

$$\tilde{d}(\boldsymbol{\psi}_{1,1}) = \tilde{\mathcal{O}}\left(\left(c_r + t_{\mathrm{mix}}\right)^{\frac{1+\lambda}{2+\lambda}} \cdot T^{\frac{1+\lambda}{2(2+\lambda)}} \cdot \left(\frac{1}{T}\sum_{t=0}^{T} \eta_t^{1+\lambda}\right)^{\frac{1}{2+\lambda}}\right) \quad \forall \lambda \in [0, \delta], \tag{3.53}$$

$$\tilde{d}(\boldsymbol{\psi}_{1,2}) = \tilde{\mathcal{O}}\left(\left(1 + t_{\mathrm{mix}}\right)^{\frac{1+m}{2+m}} \cdot T^{\frac{1+m}{2(2+m)}} \cdot \left(\frac{1}{T}\sum_{t=0}^{T} \eta_t^{1+m}\right)^{\frac{1}{2+m}}\right) \quad \forall m \in [0, 2\delta+1], \tag{3.54}$$

$$\tilde{d}(\boldsymbol{\psi}_2) = \mathcal{O}\left(T^{-\frac{1-\alpha}{3}}\right), \tag{3.55}$$

$$\tilde{d}(\boldsymbol{\psi}_3) = \mathcal{O}\left((1+l) \cdot T^{-\frac{l(1-\alpha)}{3+2l}}\right) \quad \forall l \in [0, \delta], \tag{3.56}$$

$$\tilde{d}(\boldsymbol{\psi}_{4,1}) = \tilde{\mathcal{O}}\left(\sqrt{k} C_k \cdot \left(\frac{1}{T}\sum_{t=0}^{T} \eta_t^{k/2}\right)^{\frac{1}{1+k}}\right) \quad \forall k \in [1, 2(1+\delta)], \tag{3.57}$$

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\psi}_{4,2}, \boldsymbol{\theta}^\top \boldsymbol{\psi}) = \mathcal{O}\left(T^{-\frac{\delta}{3+2\delta}} + T^{-(\frac{1}{4}-o(1))} + T^{-\frac{1}{3}} + t_{\mathrm{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}}\right) \textit{ for an infinitesimal } o(1), \tag{3.58}$$

where $c_r := \max\left\{L_G, \frac{L_H + \|\boldsymbol{G}\|}{\delta_G}\right\}$ *and* C_p *is the constant in the* $(L^p, (1+\log t)\sqrt{\eta_t})$ *consistency. Here,* $\tilde{\mathcal{O}}(\cdot)$ *hides uninterested parameters and the log factor* $\log T$.

Proof of Lemma 3.4.13. The proof can be found in Appendix D.2. □

Step three: Variable selection Notice that for any $\beta \geq 0$, we have

$$\frac{1}{T}\sum_{t=0}^{T} \eta_t^{\beta} = \frac{1}{T}\sum_{t=0}^{T} t^{-\alpha\beta} = \begin{cases} \mathcal{O}\left(\frac{1}{1-\alpha\beta} T^{-\alpha\beta}\right) & \text{if } \alpha\beta < 1 \\ \mathcal{O}\left(\frac{\log T}{T}\right) & \text{if } \alpha\beta = 1 \\ \mathcal{O}(\frac{1}{T}) & \text{if } \alpha\beta > 1 \end{cases} = \tilde{\mathcal{O}}\left(T^{-(\alpha\beta)\wedge 1}\right) \tag{3.59}$$

where $\tilde{\mathcal{O}}(\cdot)$ hides the log factor $\log T$ and constant dependence on α, β and $a \wedge b = \min\{a, b\}$.

With the help of (3.59), we simplify the bounds in Lemma 3.4.13 by choosing (nearly) optimal variables λ, m, l and k. Recall that we rewrite $p = 2(1+\delta)$ for simplicity.

- It is easy to verify that

$$J_1(\alpha) := \max_{k \in [1, 2(1+\delta)]} \frac{\alpha k \wedge 2}{2(k+1)} = \begin{cases} \frac{\alpha(1+\delta)}{3+2\delta} & \text{if } \alpha \in \left(0, \frac{1}{1+\delta}\right] \text{ achieved by } k = 2(1+\delta), \\ \frac{\alpha}{2+\alpha} & \text{if } \alpha \in \left[\frac{1}{1+\delta}, 1\right) \text{ achieved by } k = \frac{2}{\alpha}. \end{cases} \tag{3.60}$$

By setting $k = \min\left\{2(1+\delta), \frac{2}{\alpha}\right\}$, we get that

$$\tilde{d}\left(\boldsymbol{\psi}_{4,1}\right) = \tilde{\mathcal{O}}\left(\frac{C_{\frac{2}{\alpha}\wedge p}}{\sqrt{\alpha}} \cdot T^{-J_1(\alpha)}\right) = \tilde{\mathcal{O}}\left(T^{-J_1(\alpha)}\right)$$

where the last equality uses $\alpha \in (0.5, 1)$ and C_k is increasing in k.

- Note that

$$T^{\frac{1+\lambda}{2(2+\lambda)}} \cdot \left(\frac{1}{T}\sum_{t=0}^{T} \eta_t^{1+\lambda}\right)^{\frac{1}{2+\lambda}} = \tilde{\mathcal{O}}\left(T^{-\frac{1+\lambda}{2+\lambda}\left[(\alpha-0.5)\wedge\frac{1-\lambda}{2(1+\lambda)}\right]}\right) = \tilde{\mathcal{O}}\left(T^{-h_0(\lambda)}\right)$$

where we denote

$$h_0(\lambda) = \frac{1+\lambda}{2+\lambda}\left[(\alpha - 0.5) \wedge \frac{1-\lambda}{2(1+\lambda)}\right].$$

One can show that

$$J_2(\alpha) := \max_{\lambda \in [0,\delta]} h_0(\lambda) = \begin{cases} (\alpha - 0.5)\frac{1+\delta}{2+\delta} & \text{if } \alpha \in \left(0.5, \frac{1}{1+\delta}\right] \text{ achieved by } \lambda = \delta, \\ \frac{\alpha - 0.5}{\alpha+1} & \text{if } \alpha \in \left[\frac{1}{1+\delta}, 1\right) \text{ achieved by } \lambda = \frac{1}{\alpha} - 1. \end{cases} \tag{3.61}$$

By setting $\lambda = \min\left\{\delta, \frac{1}{\alpha} - 1\right\}$, we get that

$$\max\left\{\tilde{d}\left(\boldsymbol{\psi}_{1,1}\right), \tilde{d}\left(\boldsymbol{\psi}_{1,2}\right)\right\} = \tilde{\mathcal{O}}\left((1 + t_{\text{mix}})^{\frac{1+\delta}{2+\delta}} \cdot T^{-J_2(\alpha)}\right).$$

- Note that

$$\max_{l \in [0,\delta]} \left[\frac{1}{3} \wedge \frac{l}{3+2l}\right] = \begin{cases} \frac{\delta}{3+2\delta} & \text{if } \delta \in [0,3] \text{ achieved by } l = \delta, \\ \frac{1}{3} & \text{if } \delta \in [3,\infty) \text{ achieved by } l = 3. \end{cases}$$

  By setting $l = \min\{\delta, 3\}$, we have that

$$\max\left\{\tilde{d}\left(\boldsymbol{\psi}_2\right), \tilde{d}\left(\boldsymbol{\psi}_3\right)\right\} = \tilde{\mathcal{O}}\left(T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]}\right).$$

- Finally, we note that $T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]} \geq \max\left\{T^{-\frac{\delta}{3+2\delta}}, T^{-(\frac{1}{4} - o(1))}, T^{-\frac{1}{3}}\right\}$ due to $\alpha \in (0,1)$.

Combining these bounds and using $p = 2(1+\delta)$, we arrive at

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}) = \tilde{\mathcal{O}}\left(T^{-J_1(\alpha)} + (1 + t_{\text{mix}})^{\frac{p}{2+p}} \cdot T^{-J_2(\alpha)} + T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]} + t_{\text{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}}\right). \tag{3.62}$$

A special case: I.i.d. data From the above analysis, one can find that $\boldsymbol{\psi}_{1,2}$ contributes a lot to the bound (3.62). When it comes to the i.i.d. case, $\boldsymbol{U}(\boldsymbol{x}, \xi) = \boldsymbol{H}(\boldsymbol{x}, \xi)$ and $\boldsymbol{g}(\boldsymbol{x}) = \mathscr{P}\boldsymbol{H}(\boldsymbol{x}, \xi)$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\xi \in \Xi$. In this case, there is a refined decomposition where $\boldsymbol{\psi}_{1,2}$ doesn't show up. In contrast, $\boldsymbol{\psi}_{1,2}$ always appears in (3.50) no matter what the case is.

The key idea in the refined decomposition is to use $\boldsymbol{\phi}_T$ rather than $\tilde{\boldsymbol{\phi}}_T$. With a slight of notation abuse, we redefine $\boldsymbol{\Delta}_t = \boldsymbol{x}_t - \boldsymbol{x}^\star$, then similar to (3.19), we have $\boldsymbol{\Delta}_{t+1} = (\boldsymbol{I} - \eta_t \boldsymbol{G})\boldsymbol{\Delta}_t + \eta_t[\boldsymbol{r}_t + \boldsymbol{u}_t]$ where $\boldsymbol{r}_t = \boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{G}(\boldsymbol{x}_t - \boldsymbol{x}^\star)$. The key observation is that once iterating $\boldsymbol{\phi}_T$ rather than $\tilde{\boldsymbol{\phi}}_T$, the sum of the residual term and coboudary term in (3.15) equals to zero because $\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi) = \boldsymbol{g}(\boldsymbol{x}_t)$ for all $\xi \in \Xi$. Hence, by a similar recursion analysis in Section 3.4.1, we have

$$\boldsymbol{\phi}_T = \boldsymbol{\psi}_0 + \boldsymbol{\psi}_{1,1} + \boldsymbol{\psi}_2 + \boldsymbol{\psi}_3 + \boldsymbol{\psi}_{4,1} + \boldsymbol{\psi}_{4,2}.$$

By repeatedly using Proposition 3.4.1 and Corollary 3.4.1, it follows that for any $\boldsymbol{\theta} \in \mathbb{R}^d$

satisfying $\|\boldsymbol{\theta}\|_* = 1$,

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}) \le \tilde{d}(\boldsymbol{\psi}_0) + \tilde{d}(\boldsymbol{\psi}_{1,1}) + \tilde{d}(\boldsymbol{\psi}_2) + \tilde{d}(\boldsymbol{\psi}_3) + \tilde{d}(\boldsymbol{\psi}_{4,1}) + d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\psi}_{4,2}, \boldsymbol{\theta}^\top \boldsymbol{\psi}).$$

It turns out that $d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi})$ doesn't depend on $\tilde{d}(\boldsymbol{\psi}_2)$ any more. We comment that (3.53) is still a valid upper bound for $\tilde{d}(\boldsymbol{\psi}_{1,1})$ even we change the definition of $\boldsymbol{\Delta}_t$ from $\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^\star$ to $\boldsymbol{x}_t - \boldsymbol{x}^\star$.

Taking this special case into consideration, we have

$$\begin{aligned}\max\left\{\tilde{d}\left(\boldsymbol{\psi}_{1,1}\right), \tilde{d}\left(\boldsymbol{\psi}_{1,2}\right)\right\} &= \tilde{\mathcal{O}}\left((c_r + t_{\text{mix}} + (1+t_{\text{mix}})1_{t_{\text{mix}}})^{\frac{1+\delta}{2+\delta}} \cdot T^{-J_2(\alpha)}\right) \\ &= \tilde{\mathcal{O}}\left((c_r + t_{\text{mix}})^{\frac{1+\delta}{2+\delta}} \cdot T^{-J_2(\alpha)}\right),\end{aligned}$$

where $1_{t_{\text{mix}}}$ is an indicator function for the event $\{t_{\text{mix}} > 0\}$ satisfying $1_{t_{\text{mix}}} \le t_{\text{mix}}$. As a result, a finer bound is

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}) = \tilde{\mathcal{O}}\left(T^{-J_1(\alpha)} + (c_r + t_{\text{mix}})^{\frac{p}{2+p}} \cdot T^{-J_2(\alpha)} + T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]} + t_{\text{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}}\right). \tag{3.14}$$

□

3.5 Online Statistical Inference Procedure

In this section, we formally introduce the online statistical inference method. As discussed in Section 3.3.1, the key idea is to find a scale-invariant $|||\cdot|||$-continuous functional f so as to cancel out the dependence of the unknown scale $\boldsymbol{G}^{-1}\boldsymbol{S}$. For analysis facilitation, we continuize the càdlàg function $\boldsymbol{\phi}_T$ by linearly connecting points $\left\{\boldsymbol{\phi}_T\left(\frac{n}{T}\right)\right\}_{n \in [T] \cup \{0\}}$ such that it becomes an element in $\mathsf{C}_{[0,1],\mathbb{R}}$. In particular, we denote the continuous function by $\boldsymbol{\phi}_T^{\mathrm{c}}$ with the following definition that given $n \in [T-1] \cup \{0\}$, when $r \in \left[\frac{n}{T}, \frac{n+1}{T}\right]$,

$$\boldsymbol{\phi}_T^{\mathrm{c}}(r) = \boldsymbol{\phi}_T\left(\frac{n}{T}\right) + (Tr - n)\left[\boldsymbol{\phi}_T\left(\frac{n+1}{T}\right) - \boldsymbol{\phi}_T\left(\frac{n}{T}\right)\right]. \tag{3.63}$$

One can show that $\boldsymbol{\phi}_T^{\mathrm{c}} \overset{w}{\to} \boldsymbol{\psi}$ in the uniform topology effortlessly from Theorem 3.3.1.

Theorem 3.5.1. *Under the same assumptions of Theorem 3.3.1, it follows that*

$$\boldsymbol{\phi}_T^{\mathrm{c}} \overset{w}{\to} \boldsymbol{G}^{-1}\boldsymbol{G}^{1/2}\boldsymbol{W}$$

in the uniform topology with the same $\boldsymbol{G}, \boldsymbol{S}$ *given in Theorem 3.3.1.*

Proof of Theorem 3.5.1. One can show that $\left|\left|\left|\boldsymbol{\phi}_T - \boldsymbol{\phi}_T^{\mathrm{c}}\right|\right|\right| = o_{\mathbb{P}}(1)$. This is because of the equality $|||\boldsymbol{\phi}_T - \boldsymbol{\phi}_T^{\mathrm{c}}||| = \frac{1}{\sqrt{T}} \sup_{n\in[T]} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|$ and the fact that

$$\frac{1}{T}\mathbb{E}\sup_{t\in[T]}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \lesssim \frac{\log T}{T}\sum_{t=1}^{T}\eta_t \to 0.$$

Hence, we know that $\boldsymbol{\phi}_T^{\mathrm{c}} \overset{w}{\to} \boldsymbol{\psi}$ in the Skorokhod topology. That is for any bounded and d_{S}-continuous functional $h : \mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$, we have $\mathbb{E}h(\boldsymbol{\phi}_T^{\mathrm{c}}) \to \mathbb{E}h(\boldsymbol{\psi})$. Note that any bounded and $|||\cdot|||$-continuous functional $h : \mathsf{C}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$ can be viewed as a bounded and d_{S}-continuous functional $\mathsf{D}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$. Hence, $\mathbb{E}h(\boldsymbol{\phi}_T^{\mathrm{c}}) \to \mathbb{E}h(\boldsymbol{\psi})$ holds for any bounded and $|||\cdot|||$-continuous functional $h : \mathsf{C}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$. It is equivalent to $\boldsymbol{\phi}_T^{\mathrm{c}} \overset{w}{\to} \boldsymbol{\psi}$ in the uniform topology. □

For simplicity, we focus on one-dimensional inference via the one-dimensional projected process $\phi_T := \boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$ and consider the one-dimensional scale-invariant functional $f : \mathsf{C}_{[0,1],\mathbb{R}^d} \to \mathbb{R}$. Such a f satisfies $f(a\phi) = f(\phi)$ for any process $\phi \in \mathsf{C}_{[0,1],\mathbb{R}}$ and positive number $a > 0$.

Corollary 3.5.1. *Under the same assumptions in Theorem 3.5.1, for any $\boldsymbol{\theta} \in \mathbb{R}^d$ and any $|||\cdot|||$-continuous scale-invariant functional $f : \mathsf{C}_{[0,1],\mathbb{R}} \to \mathbb{R}$, it follows that as $T \to \infty$,*

$$f(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}) \overset{w}{\to} f(W).$$

where $W = \{W(r) : r \in [0,1]\}$ is the standard one-dimensional Brownian motion on $[0,1]$.

Proof of Corollary 3.5.1. By Theorem 3.5.1, we have $f(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}) \overset{w}{\to} f(\boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W})$. We complete the proof by noting that $\boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W} \overset{d}{=} \|\boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\|_2 W$ and f is a scale-invariant functional so that $f(\|\boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\|_2 W) = f(W)$. □

3.5.1 A Family of Scale-invariant Functional f_m

We then explore possible choices of adequate functional f. In statistics, the t-statistic is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error. It is of great use when the population standard deviation is unknown. For the partial-sum process $\boldsymbol{\phi}_T$, $\boldsymbol{\phi}_T(1)$ is exactly the difference between averaged estimator $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t$ and the hypothesized value $\boldsymbol{x}^\star$ (up to a factor $\sqrt{T}$). Following the spirit of t-statistics, we propose a family of scale-invariant functional $f_m (m \in \mathbb{N})$ by using different normalization terms to remove the scale dependence

$$f_m(\phi) = \frac{\phi(1)}{\sqrt[m]{\int_0^1 |\phi(r) - r\phi(1)|^m dr}}. \tag{3.64}$$

In the econometrics literature, the pivotal statistics $f_2(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T)$ is used to conduct robust testing and result in the fixed bandwidth heteroskedasticity and autocorrelation robust (fixed-b HAR) estimator. Such an estimator takes advantage of the underlying autocorrelation structure in linear autoregressive models and overcomes the series correlation and heteroskedasticity therein[63, 72]. Lee, Liao, Seo, Shin [62] utilizes and generalizes this technique to propose an online statistical inference method named as random scaling for SGD iterates. Subsequent works follow the spirit and propose similar procedures for specific iterates $\{\boldsymbol{x}_t\}_{t\geq 0}$ under i.i.d. data[21, 50, 106]. In our work, we consider a general family of m-th root normalization in (3.64) instead of the square root normalization in f_2.

Proposition 3.5.1. *The functional f_m are scale-invariant and symmetric so that $f_m(-\phi) = -f_m(\phi)$ for any process ϕ and $m \geq 1$. Furthermore, it is $|||\cdot|||$-continuous in the uniform topology.*

As a result of Proposition 3.5.1, the limiting distribution $f_m(W)$ is mixedly normal and symmetric around zero. For better illustration, we show the density probability function of different $f_m(W)$'s in Figure 3.1(a) and compute the corresponding asymptotic critic values $q_{\alpha,m}$ in Table 3.1. We note that Abadir, Paruolo [72] calculates the probability density of $f_2(W)$ explicitly, based on which more accurate asymptotic critic values are accessible. We perform stochastic simulations to approximate each $q_{\alpha,m}$ as what Kiefer, Vogelsang, Bunzel [63] did for simplicity and universality. Numerical experiments in Section 3.6 validate its sufficiency. Finally, the following proposition shows how we can establish the confidence set by inverting the asymptotic pivotal statistics.

Proposition 3.5.2. *Under the same assumptions in Theorem 3.3.1, given $\boldsymbol{\theta} \in \mathbb{R}^d$ and $m \geq 1$, it follows that when $T \to \infty$,*

$$\mathbb{P}\left(\boldsymbol{\theta}^\top \boldsymbol{x}^\star \in \mathscr{C}(\alpha, m)\right) \to 1 - \alpha,$$

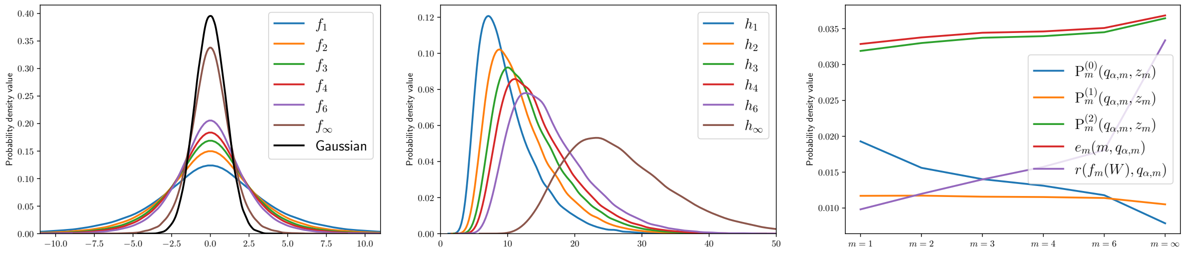
where $\mathscr{C}(\alpha, m)$ is the α-level confidence set defined by

$$\mathscr{C}(\alpha, m) := \left\{\boldsymbol{\theta}^\top \boldsymbol{x}^\star \in \mathbb{R} : |f_m(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}})| \leq q_{\alpha,m}\right\} \tag{3.65}$$

and $q_{\alpha,m}$ is the critical value satisfying $\mathbb{P}(|f_m(W)| \geq q_{\alpha,m}) = \alpha$.

f \ $1-\alpha$	1%	2.5%	5%	10%	50%	90%	95%	97.5%	99%
f_1	-10.705	-8.334	-6.569	-4.749	0.000	4.749	6.569	8.334	10.705
f_2	-8.628	-6.758	-5.316	-3.873	0.000	3.873	5.316	6.758	8.628
f_3	-7.495	-5.899	-4.650	-3.403	0.000	3.403	4.650	5.899	7.495
f_4	-6.798	-5.344	-4.232	-3.108	0.000	3.108	4.232	5.344	6.798
f_6	-5.969	-4.705	-3.728	-2.754	0.000	2.754	3.728	4.705	5.969
f_∞	-3.408	-2.711	-2.175	-1.626	0.000	1.626	2.175	2.711	3.408

Table 3.1 Asymptotic critic values $q_{\alpha,m}$ of $f_m(W)$ defined by $q_{\alpha,m} = \sup\{q : \mathbb{P}(|f_m(W)| \geq q) \leq \alpha\}$. They are computed via simulations. In particular, the Brownian motion W is approximated by normalized sums of i.i.d. $\mathcal{N}(0,1)$ pseudo-random deviates using 1,000 steps and 50,000 replications.



(a) P.d.f. of $f_m(W)$'s (b) P.d.f. of $h_m(W)$'s (c) Components of $e(m, q_{\alpha,m})$

Figure 3.1 (a) shows the probability density functions (p.d.f.) of different $f_m(W)$'s. The black line represents the standard normal distribution. (b) shows the p.d.f. of the denominator of different $f_m(W)$'s, denoted by $h_m(W)$'s. (c) computes the dominant quantities in the bound (3.72).

3.5.2 Online Computation Efficiency

We study per-iteration computation complexity of computing different f_m's in the subsection. We denote $\phi_T = \boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}$ and $\bar{\boldsymbol{x}}_t = \frac{1}{t}\sum_{\tau=0}^{t} \boldsymbol{x}_\tau$ the averaged iterates at iteration t.

Proposition 3.5.3. *$f_m(\phi_T)$ with an even number m can be computed efficiently online.*

We explain this above proposition in the following. First, the numerator is set to be $\phi_T(1) = \boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}(1) = \sqrt{T}\boldsymbol{\theta}^\top(\bar{\boldsymbol{x}}_T - \boldsymbol{x}^\star)$ where $\bar{\boldsymbol{x}}_T$ can be undated in a moving average form, incurring $\mathcal{O}(1)$ additional computation cost per iteration. Second, denoting $\phi_{n,T} = \frac{n}{\sqrt{T}}\boldsymbol{\theta}^\top\left(\bar{\boldsymbol{x}}_n - \bar{\boldsymbol{x}}_T\right)$ for simplicity, we have when $r \in [\frac{n}{T}, \frac{n+1}{T})$ for some $n \in \mathbb{N}$,

$$\phi_T(r) - r\phi_T(1) = \phi_{n,T} + (Tr - n)(\phi_{n+1,T} - \phi_{n,T}),$$

which has nothing to do with the unknown parameter $\boldsymbol{x}^\star$. It is easy to verify that

$$\int_0^1 (\phi_T(r) - r\phi_T(1))^2 dr = \sum_{n=0}^{T-1} \int_{\frac{n}{T}}^{\frac{n+1}{T}} (\phi_{n,T} + (Tr-n)(\phi_{n+1,T} - \phi_{n,T}))^2 dr$$

$$= \sum_{n=0}^{T-1} \frac{(\phi_{n,T})^2 + \phi_{n,T}\phi_{n+1,T} + (\phi_{n+1,T})^2}{3T}.$$

The right-hand side of the last equality can be computed in an online manner. Indeed, by expanding $(\phi_{n,T})^2$ into $\frac{n^2}{T}\left((\theta^\top \bar{x}_n)^2 + 2(\theta^\top \bar{x}_T)^2 + \theta^\top \bar{x}_n \theta^\top \bar{x}_T\right)$ and doing similarly for $\phi_{n,T}\phi_{n+1,T}$ and $(\phi_{n+1,T})^2$, one can find that the sum of each decomposed terms can be updated fully online without passing the observed data twice. A simpler method used for $m = 2$ is to approximate each $\frac{(\phi_{n,T})^2+\phi_{n,T}\phi_{n+1,T}+(\phi_{n+1,T})^2}{3T}$ with $\frac{(\phi_{n+1,T})^2}{T}$[21, 50, 62, 106]. In other words, we use the rectangle rule to compute the integral $\int_{\frac{n}{T}}^{\frac{n+1}{T}} (\phi(r) - r\phi(1))^2 dr$ instead of the Trapezoid rule so as to simplify computation. In this way,

$$\int_0^1 (\phi_T(r) - r\phi_T(1))^2 dr \approx \sum_{n=1}^{T} \frac{(\phi_n^T)^2}{T} = \frac{1}{T^2}\sum_{n=1}^{T} n^2 \left[(\theta^\top \bar{x}_n)^2 + (\theta^\top \bar{x}_T)^2 + 2\theta^\top \bar{x}_n \theta^\top \bar{x}_T\right] \tag{3.66}$$

can be constructed in a simpler online fashion via only two iterative updates of $(\theta^\top \bar{x}_n)^2$ and $\theta^\top \bar{x}_n \theta^\top \bar{x}_T$. Simulation studies turn out hardly any difference between them in terms of empirical coverage and confidence interval lengths (see Table 3.2). Hence, we will use the rectangle-rule approximation to compute $\int_0^1 (\phi_T(r) - r\phi_T(1))^m dr (m = 2, 4, 6)$ in all experiments. Once the integral is computed and denoted by $\sigma_{m,T}$, inverting (3.65) produces the following the confidence interval

$$\theta^\top x^\star \in \left[\theta^\top \bar{x}_T - \frac{q_{\alpha,m}}{\sqrt{T}} \cdot \sigma_{m,T}, \theta^\top \bar{x}_T + \frac{q_{\alpha,m}}{\sqrt{T}} \cdot \sigma_{m,T}\right]. \tag{3.67}$$

However, $f_m(\phi_T^c)$ with an odd m can't be computed online efficiently. This is because there is no similar decomposition as (3.66) for the integral $\int_0^1 |\phi_T(r) - r\phi_T(1)|^{2k+1} dr$ due to its inner absolute value. More specially, computing (or approximating) $\int_0^1 (\phi_{T+1}(r) - r\phi_{T+1}(1))^2 dr$ necessitates the calculation of all the values $\{\phi_{n,T+1}\}_{n\in[T]}$, incurring $\mathcal{O}(T)$ computation cost. By contrast, as we illustrate in (3.66), the existence of a closed-form decomposition for the integration with an even m enables an incremental update to each decomposed term, incurring only $\mathcal{O}(1)$ computation cost per iteration. For completeness, we include three examples with $m = 1, 3, \infty$ for a fair comparison. When $m = \infty$, we have $f_\infty = \frac{\phi(1)}{\sup_{r\in[0,1]} |\phi(r) - r\phi(1)|}$.

3.5.3 A Qualitative Study

In previous subsections, we have proposed a family of the scale-invariant functional f_m which introduce the different asymptotic pivotal statistics $f_m(W)$. The choice of m's not only affects the critical value $q_{\alpha,m}$ in the confidence interval (3.65) but also the convergence of rejection probability. We measure the latter by $e(m,x)$ with the following definition

$$e(m,x) := |\mathbb{P}(|f_m(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}})| > x) - \mathbb{P}(|f_m(W)| > x)|, \tag{3.68}$$

which is the absolute error of the tail probability of $|f_m(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}})|$ against the tail probability of the limiting distribution $|f_m(W)|$.

Theorem 3.5.2. *Let $\varepsilon_{\mathrm{P}} = d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}, \boldsymbol{\theta}^\top \boldsymbol{\psi})$ denote the Lévy-Prokhorov distance.*[1] *Under the assumptions of Theorem 3.3.1, it follows that for any $x > 0$ and $z > 0$,*

$$e(m,x) \le 2\left[\mathrm{P}_m^{(0)}(x,z) \cdot \frac{\varepsilon_{\mathrm{P}}}{\omega} + \max\left\{\mathrm{P}_m^{(1)}(x,z), \mathrm{P}_m^{(2)}(x,z)\right\}\right] + o(\varepsilon_{\mathrm{P}}), \tag{3.69}$$

where

$$\begin{aligned}
\mathrm{P}_m^{(0)}(x,z) &= r(f_m(W),x) \cdot \frac{x+1}{z},\\
\mathrm{P}_m^{(1)}(x,z) &= \mathbb{P}\left(|f_m(W)| > x \textit{ and } h_m(W) \le z\right),\\
\mathrm{P}_m^{(2)}(x,z) &= \mathbb{P}\left(|f_m(W)| \le x \textit{ and } h_m(W) \le z\right).
\end{aligned}$$

In this context, $r(X,x)$ refers to the probability density function value of the random variable X at point x, while $\omega = \|\boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \boldsymbol{S}^{1/2}\|_2$ represents the unknown scale. Furthermore, we define $\mathrm{Prob}_m(x,z)$ *as follows, where $h_m(W)$ corresponds to the denominator of $f_m(W)$, Finally, the* $o(1)$ *term denotes an infinitesimal term (that might depend on x, z) when $\varepsilon_{\mathrm{P}} \to 0$.*

Proof of Theorem 3.5.2. Recall that $\phi_T = \boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}$ and $\psi = \boldsymbol{\theta}^\top \boldsymbol{\psi}$. Let $\boldsymbol{B}_x := \left\{\phi : |f_m(\phi)| > x\right\}$ and $\varepsilon_{\mathrm{P}} = d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}, \boldsymbol{\theta}^\top \boldsymbol{\psi})$. From Proposition 3.5.1, the functional f_m is continuous such that $\boldsymbol{B}_x$ is a measurable set in its Borel σ-field $\mathscr{C}_{[0,1],\mathbb{R}}$. By the definition of the Lévy-Prokhorov distance in $\mathsf{C}_{[0,1],\mathbb{R}}$, we have

$$\mathbb{P}(\phi_T \in \boldsymbol{B}_x) \le \mathbb{P}(\psi \in \boldsymbol{B}_x^{\varepsilon_{\mathrm{P}}}) + \varepsilon_{\mathrm{P}} \quad \text{and} \quad \mathbb{P}(\phi_T \in \boldsymbol{B}_x^{\mathrm{c}}) \le \mathbb{P}(\psi \in (\boldsymbol{B}_x^{\mathrm{c}})^{\varepsilon_{\mathrm{P}}}) + \varepsilon_{\mathrm{P}},$$

where $\boldsymbol{B}_x^{\varepsilon}$ is the ε-neighborhood of $\boldsymbol{B}_x$ defined as following and $\boldsymbol{B}_x^{\mathrm{c}}$ is the complementary set of $\boldsymbol{B}_x$,

$$\boldsymbol{B}_x^{\varepsilon} = \left\{\phi_1 \in \mathsf{C}_{[0,1],\mathbb{R}} : \exists \phi_2 \in \mathsf{C}_{[0,1],\mathbb{R}} \text{ such that } \left|\!\left|\!\left|\phi_1 - \phi_2\right|\!\right|\!\right| \le \varepsilon \text{ and } \phi_2 \in \boldsymbol{B}_x\right\}.$$

Then, it follows that

$$\mathbb{P}(|f_m(\phi_T)| > x) - \mathbb{P}(|f_m(\psi)| > x) \le \mathbb{P}(\psi \in \boldsymbol{B}_x^{\varepsilon_{\mathrm{P}}}, \psi \notin \boldsymbol{B}_x) + \varepsilon_{\mathrm{P}}, \tag{3.70}$$

[1] Since both $\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}}$ and W are continuous functions, the definition of the Lévy-Prokhorov distance for $\mathsf{C}_{[0,1],\mathbb{R}}$ is different than that for $\mathsf{D}_{[0,1],\mathbb{R}}$ in the topology (i.e., the uniform topology) and the Borel σ-field used.

$$\mathbb{P}(|f_m(\psi)| > x) - \mathbb{P}(|f_m(\phi_T)| > x) \leq \varepsilon_{\mathrm{P}} + \mathbb{P}(\psi \in B_x, \psi \notin (B_x^c)^{\varepsilon_{\mathrm{P}}}). \tag{3.71}$$

We rewrite $f_m(\phi) = \frac{\phi(1)}{h_m(\phi)}$ where $h_m(\phi) = \sqrt[m]{\int_0^1 (\phi(r) - r\phi(1))^m dr}$ denotes the integral functional. By the Minkowski inequality, we know that $h_m(\phi)$ is a 1-Lipschitz $|||\cdot|||$-continuous functional in the sense that for any $\phi_1, \phi_2 \in \mathscr{C}_{[0,1],\mathbb{R}}$, $|h_m(\phi_1) - h_m(\phi_2)| \leq \left|\left|\left|\phi_1 - \phi_2\right|\right|\right|$. Furthermore, we have

Proposition 3.5.4. *When $h_m(\phi_1) \geq z, h_m(\phi_2) \geq z$ and $|f_m(\phi_1)| \leq x$, one can show that*

$$|f_m(\phi_1) - f_m(\phi_2)| \leq \frac{1+x}{z}\left|\left|\left|\phi_1 - \phi_2\right|\right|\right|.$$

Proof of Proposition 3.5.4. It follows that

$$\begin{aligned}
|f_m(\phi_1) - f_m(\phi_2)| &\leq \left|\frac{\phi_1(1)}{h_m(\phi_1)} - \frac{\phi_1(1)}{h_m(\phi_2)}\right| + \left|\frac{\phi_2(1)}{h_m(\phi_2)} - \frac{\phi_2(1)}{h_m(\phi_2)}\right| \\
&\leq \left|\frac{\phi_1(1)}{h_m(\phi_1)}\right| \cdot \frac{|h_m(\phi_1) - h_m(\phi_2)|}{|h_m(\phi_2)|} + \frac{|\phi_2(1) - \phi_1(1)|}{|h_m(\phi_2)|} \\
&\leq \frac{x}{z}\left|\left|\left|\phi_1 - \phi_2\right|\right|\right| + \frac{1}{z}\left|\left|\left|\phi_1 - \phi_2\right|\right|\right| = \frac{1+x}{z}\left|\left|\left|\phi_1 - \phi_2\right|\right|\right|.
\end{aligned}$$

□

We then proceed to simplify (3.70). It follows that

$$\begin{aligned}
&\mathbb{P}(\psi \in B_x^{\varepsilon_{\mathrm{P}}}, \psi \notin B_x) = \mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } \exists\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, |f_m(\tilde{\psi})| > x\right) \\
&\quad\leq \mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } \exists\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, |f_m(\tilde{\psi})| > x, h_m(\psi) \geq z, h_m(\tilde{\psi}) \geq z\right) \\
&\qquad+ \mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } \forall\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, h_m(\tilde{\psi}) < z\right) \\
&\qquad+ \mathbb{P}(|f_m(\psi)| \leq x \text{ and } h_m(\psi) < z) \\
&\quad\leq \mathbb{P}\left(x - \frac{x+1}{z}\varepsilon_{\mathrm{P}} \leq |f_m(\psi)| \leq x\right) + 2\mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } h_m(\psi) \leq z\right) \\
&\quad= r(|f_m(\psi)|, x) \cdot \frac{x+1}{z} \cdot \varepsilon_{\mathrm{P}} + 2\mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } h_m(\psi) \leq z\right) + o(\varepsilon_{\mathrm{P}}),
\end{aligned}$$

where the second inequality uses Proposition 3.5.4 and the 1-Lipschitz $|||\cdot|||$-continuity of h_m and the last inequality uses the definition of differentiability.

By a similar argument, for any $z > \varepsilon_{\mathrm{P}}$, we simplify (3.71) to

$$\begin{aligned}
&\mathbb{P}(\psi \in B_x, \psi \notin (B_x^c)^{\varepsilon_{\mathrm{P}}}) = \mathbb{P}\left(|f_m(\psi)| > x \text{ and } \forall\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, |f_m(\tilde{\psi})| > x\right) \\
&\quad\leq \mathbb{P}\left(|f_m(\psi)| > x \text{ and } \forall\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, |f_m(\tilde{\psi})| > x, h_m(\psi) \geq z - \varepsilon_{\mathrm{P}}, h_m(\tilde{\psi}) \geq z - \varepsilon_{\mathrm{P}}\right) \\
&\qquad+ \mathbb{P}\left(|f_m(\psi)| > x \text{ and } \exists\tilde{\psi} \text{ satisfying } |||\tilde{\psi} - \psi||| \leq \varepsilon_{\mathrm{P}}, h_m(\tilde{\psi}) < z - \varepsilon_{\mathrm{P}}\right) \\
&\qquad+ \mathbb{P}(|f_m(\psi)| > x \text{ and } h_m(\psi) < z - \varepsilon_{\mathrm{P}})
\end{aligned}$$

$$\leq \mathbb{P}\left(x \leq |f_m(\psi)| \leq x + \frac{x+1}{z-\varepsilon_{\mathrm{P}}}\varepsilon_{\mathrm{P}}\right) + 2\mathbb{P}\left(|f_m(\psi)| > x \text{ and } h_m(\psi) \leq z\right)$$
$$= r(|f_m(\psi)|, x) \cdot \frac{x+1}{z-\varepsilon_{\mathrm{P}}} \cdot \varepsilon_{\mathrm{P}} + 2\mathbb{P}\left(|f_m(\psi)| > x \text{ and } h_m(\psi) \leq z\right) + o(\varepsilon_{\mathrm{P}}),$$
$$= r(|f_m(\psi)|, x) \cdot \frac{x+1}{z} \cdot \varepsilon_{\mathrm{P}} + 2\mathbb{P}\left(|f_m(\psi)| > x \text{ and } h_m(\psi) \leq z\right) + o(\varepsilon_{\mathrm{P}}).$$

Combing these bounds for (3.70) and (3.71), we have for any $z > 0$

$$|\mathbb{P}(|f_m(\phi_T)| > x) - \mathbb{P}(|f_m(\psi)| > x)| \leq r(|f_m(\psi)|, x) \cdot \frac{x+1}{z} \cdot \varepsilon_{\mathrm{P}} + 2 \cdot \mathrm{Prob}_m(x, z) + o(\varepsilon_{\mathrm{P}}).$$

where

$$\mathrm{Prob}_m(x, z) = \max\left\{\mathbb{P}\left(|f_m(\psi)| > x \text{ and } h_m(\psi) \leq z\right), \mathbb{P}\left(|f_m(\psi)| \leq x \text{ and } h_m(\psi) \leq z\right)\right\}.$$

Let $\omega = \|\boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \boldsymbol{S}^{1/2}\|_2$. We then have that $\psi \overset{d}{=} \omega W$. On one hand, we note that $r(|f_m(\psi)|, x) = 2 \cdot r(f_m(\psi), |x|) = 2 \cdot r(f_m(W), |x|)$ due to the symmetry of the probability density function of $f_m(\psi)$ and its scale-invariance, i.e., $f_m(\psi) \overset{d}{=} f_m(\omega W) = f_m(W)$. On the other hand, we have $h_m(\psi) \overset{d}{=} \omega h_m(W)$. Therefore,

$$\mathrm{Prob}_m(x, z) = \max\left\{\mathbb{P}\left(|f_m(W)| > x \text{ and } h_m(W) \leq \frac{z}{\omega}\right), \mathbb{P}\left(|f_m(W)| \leq x \text{ and } h_m(W) \leq \frac{z}{\omega}\right)\right\}.$$

Finally,we complete the proof by replacing z with $z\omega$ and still denote the last equation as $\mathrm{Prob}_m(x, z)$ with a slight abuse of notation. □

Theorem 3.5.2 shows that the absolute error $e(m, x)$ depends on three factors, namely the Lévy-Prokhorov distance ε_{P}, the probability density function values $r(f_m(W), x)$, and the joint probability $\mathrm{Prob}_m(x, z)$ where $h_m(W) = \sqrt[m]{\int_0^1 |W(r) - rW(1)|^m dr}$ is the denominator of $f_m(W)$. From Theorem 3.5.1, we know that $\varepsilon_{\mathrm{P}} \to 0$ as $T \to \infty$. A non-asymptotic bound for ε_{P} is accessible via a similar argument in proving Theorem 3.3.5 that makes the weak convergence bound for $d_{\mathrm{P}}(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi})$ explicit.

The bound (3.69) captures the convergence rate of rejection probability. The dependence of $e(m, q_{\alpha,m})$ on m is of interest because it provides practical instruction for selecting m. Let z_m be the number satisfying $\mathbb{P}\left(h_m(W) \leq z_m\right) = \frac{\varepsilon_{\mathrm{P}}}{\omega}$. Plugging $x = q_{\alpha,m}$ and $z = z_m$ into (3.69) yields that

$$e(m, q_{\alpha,m}) = \underbrace{2\mathrm{P}_m^{(0)}(q_{\alpha,m}, z_m) \cdot \frac{\varepsilon_{\mathrm{P}}}{\omega}}_{\text{decreasing in } m} + \underbrace{2\max\left\{\mathrm{P}_m^{(1)}(q_{\alpha,m}, z_m), \mathrm{P}_m^{(2)}(q_{\alpha,m}, z_m)\right\}}_{\text{increasing in } m} + o(\varepsilon_{\mathrm{P}}). \qquad (3.72)$$

When we set $\alpha = 0.975$ and $\frac{\varepsilon_{\mathrm{P}}}{\omega} = 0.05$, the first two terms in (3.72) are of comparable magnitude, but are still difficult to analyze. To understand the behavior of $e(m, q_{\alpha,m})$ as a function of m, we compute the individual components of the bound (3.72) and plot them in Figure 3.1(c). In Figure 3.1, we present the probability density functions for $r(f_m(W), x)$, which reveal that $r(f_m(W), x)$ decreases in m for a given $x \in (2.5, 10)$, an interval where most of the 97.5%-level asymptotic critic values $q_{0.975,m}$ are located. By contrast, Figure 3.1(c) demonstrates that $r(f_m(W), q_{\alpha,m})$ increases with m. By applying Hölder's inequality, we observe that $h_m(W)$ and z_m increase with m, whereas $|f_m(W)|$ and $q_{\alpha,m}$ decrease for any $\alpha > 0$ (Table 3.1 confirms this). Consequently, the term $\mathrm{P}_m^{(0)}(q_{\alpha,m}, z_m)$ decreases with m. Furthermore, Figure 3.1(c) illustrates that $\mathrm{P}_m^{(2)}(q_{\alpha,m}, z_m)$ increases with m and has a greater magnitude than both $\mathrm{P}_m^{(0)}(q_{\alpha,m}, z_m)$ and $\mathrm{P}_m^{(1)}(q_{\alpha,m}, z_m)$. Therefore, the final dependency of $e(m, q_{\alpha,m})$ on m is dominated by $\mathrm{P}_m^{(2)}(q_{\alpha,m}, z_m)$ and remains increasing. This trend is further supported by the experimental findings in Figure 3.2 and 3.3. It implies that, smaller m contributes to a faster convergence of $\mathbb{P}(|f_m(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T^{\mathrm{c}})| > q_{\alpha,m})$ and, in turn, a more rapid convergence of empirical coverage.

We then study the effect of m on the length of the asymptotic confidence interval. We denote the length by $L_{m,T} := \frac{2}{\sqrt{T}} \cdot q_{\alpha,m}\sigma_{m,T}$ according to (3.67). By Hölder's inequality, we know that $\sigma_{m,T}$ increases in m for any fixed T, while Table 3.1 shows that $q_{\alpha,m}$ decreases in m for most used α's. Numerical experiments turn out that the final monotone tendency of m on the length $L_{m,T}$ is still decreasing (see Table 3.2).

Finally, we comment that (3.69) can be further minimized by choosing an optimal z when an explicit formula of the growth rate in x of the head probability $\mathbb{P}(h_m(W) \le z)$ is available. The following corollary serves as an example.

Corollary 3.5.2. *Under the assumptions of Theorem 3.3.1, if there exist $a_m, b_m > 0$ such that $\mathbb{P}(h_m(W) \le z) = a_m \cdot z^{b_m} + o(z^{b_m})$ when $z \to 0$, then it follows that for any $x > 0$,*

$$e(m, x) = 4a_m^{\frac{1}{b_m+1}} \cdot \left(\frac{r(f_m(W), x) \cdot (x+1)}{w} \cdot \varepsilon_{\mathrm{P}} \right)^{\frac{b_m}{b_m+1}} + o\left(\varepsilon_{\mathrm{P}}^{\frac{b_m}{b_m+1}} \right).$$

Proof of Corollary 3.5.2. We omit the dependency on m for simplicity. The corollary follows by noting $\mathrm{Prob}_m(x, z) \le \mathbb{P}(h_m(W) \le z) = a_m \cdot z^{b_m} + o(z^{b_m})$ for any $x > 0$ and using the particular choice of $z = \left(\frac{r(f_m(W),x)\cdot(x+1)\omega^{b_m}}{a_m} \varepsilon_{\mathrm{P}} \right)^{\frac{1}{b_m+1}}$. □

T Method	400	2000	10000	50000	400	2000	10000	50000
f_1 (Both)	87.8 (1.464)	91.2 (1.267)	91.6 (1.241)	94.0 (1.062)	131.208 (75.616)	57.871 (29.175)	27.032 (12.445)	12.23 (5.583)
f_2 (Both)	87.6 (1.474)	90.8 (1.293)	92.6 (1.171)	94.4 (1.028)	126.916 (69.3)	56.731 (27.11)	26.424 (11.531)	11.961 (5.168)
f_3	86.4 (1.533)	90.2 (1.33)	92.4 (1.185)	94.8 (0.993)	122.709 (64.969)	55.44 (25.656)	25.827 (10.898)	11.718 (4.89)
f_4	86.2 (1.542)	89.6 (1.365)	91.8 (1.227)	94.2 (1.045)	118.943 (61.729)	54.179 (24.524)	25.274 (10.421)	11.497 (4.681)
f_6	85.2 (1.588)	89.2 (1.388)	91.8 (1.227)	93.6 (1.095)	114.021 (57.747)	52.534 (23.102)	24.597 (9.848)	11.242 (4.428)
f_∞	79.2 (1.815)	84.8 (1.606)	88.2 (1.443)	90.8 (1.293)	89.64 (42.106)	43.073 (17.296)	20.852 (7.631)	9.835 (3.465)
T Bootstrap	40	200	1000	5000	40	200	1000	5000
$B = 10$	39.0 (2.181)	65.4 (2.127)	73.8 (1.966)	78.2 (1.846)	17.501 (5.185)	25.63 (6.595)	19.352 (4.876)	9.404 (2.262)
$B = 50$	49.0 (2.236)	80.6 (1.768)	90.8 (1.293)	92.0 (1.213)	19.368 (4.409)	29.883 (4.434)	24.374 (3.206)	11.943 (1.57)
$B = 100$	47.8 (2.234)	79.0 (1.822)	92.6 (1.171)	95.0 (0.975)	19.672 (4.121)	31.176 (3.854)	25.191 (2.401)	12.473 (1.145)
$B = 200$	51.4 (2.235)	79.8 (1.796)	92.0 (1.213)	92.8 (1.156)	32.339 (8.637)	48.869 (6.128)	37.095 (3.056)	17.801 (1.184)

Table 3.2 Averaged coverage rates (%, left) and average lengths (10^{-2}, right) of different inference methods over 500 Monte-Carlo simulations. Standard deviations are reported inside the parentheses.

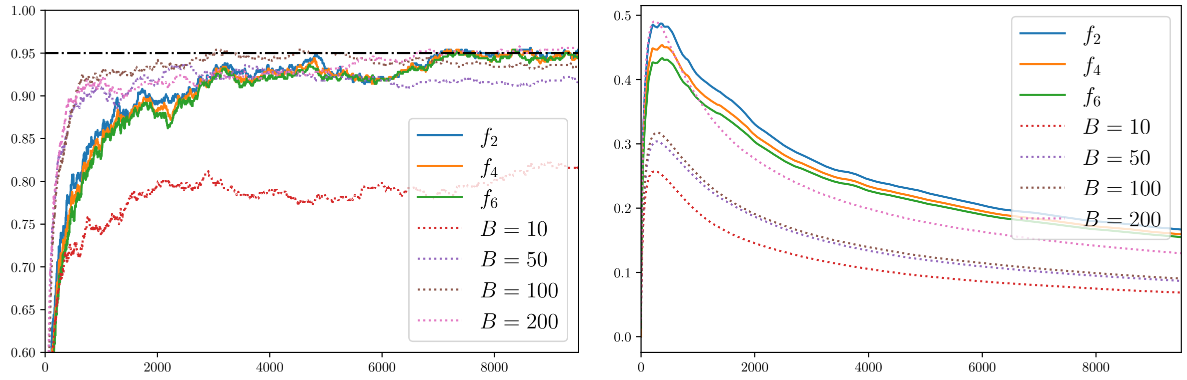
3.6 Numerical Experiments

In this numerical section, we not only conduct validation experiments to support the claims in the last section, but also investigate the empirical performance of the proposed inference procedures and their corresponding coverage rates for different examples introduced in Section 3.2.2.

3.6.1 Linear regression with autoregressive noises

In this experiment, we consider linear regression with autoregressive noises. In this linear problem, the observed data $\xi_t = (\boldsymbol{a}_t, y_t)$ is generated as the following manner

$$\boldsymbol{a}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d),\ y_t = \langle \boldsymbol{a}_t, \boldsymbol{x}^\star \rangle + \zeta_t,\ \zeta_t = \rho_\varepsilon \cdot \zeta_{t-1} + \varepsilon_t, \varepsilon_t \overset{i.i.d.}{\sim} \sqrt{d} \cdot \text{Uniform}(\mathbb{B}_{d-1}),$$



(a) Empirical coverage rates

(b) Averaged CI lengths

Figure 3.2 Performance of different inference methods for linear regression with autoregressive noises. (a) shows the empirical coverage rates based on 500 repeated experiments. The black dot line represents the nominal 95% coverage rate. (b) shows the averaged confidence interval (CI) lengths.

where the infused noise ζ_t is sampled from an autoregressive process and ρ_ε is the unknown coefficient. In this setup, one can find that all of the imposed assumptions are satisfied, the update (3.2) reduces to $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta_t \boldsymbol{a}_t(\langle \boldsymbol{a}_t, \boldsymbol{x} \rangle - y_t)$, and the confidence interval is given in (3.67). Here our target is to estimate and construct confidence intervals for $\boldsymbol{\theta}^\top \boldsymbol{x}^\star$ with $\boldsymbol{\theta} = (1, \cdots, 1)^\top/\sqrt{d} \in \mathbb{R}^d$ and $\boldsymbol{x}^\star$'s coordinates evenly spread in the interval $[0, 1]$. We test the performance of each f_m, where m takes values from $\{1, 2, 3, 4, 6, \infty\}$, and use two methods to calculate the integral in the denominator of f_1, f_2. Our benchmark is the online bootstrap inference method for linear SA with Markov data[43]. This method approximates the distribution of $\bar{\boldsymbol{x}}_T$ by maintaining and bootstrapping $B = 200$ perturbed SA iterates $\{\bar{\boldsymbol{x}}_T^b\}_{b\in[B]}$. The perturbations are made by computing $\boldsymbol{x}_{t+1}^b = \boldsymbol{x}_t^b - \eta_t W_t^b \boldsymbol{H}(\boldsymbol{x}_t^b, \xi_t)$ and $\bar{\boldsymbol{x}}_T^b = \frac{1}{T}\sum_{t\in[T]} \boldsymbol{x}_t^b$ where $\{W_t^b\}_{t\in[T], b\in[B]}$ is a bounded sequence of i.i.d. random variables with mean one and variance one.

We report the performance of confidence intervals with their average coverage rates and average lengths in Table 3.2 and Figure 3.2. We note the following findings from these results. Firstly, there is minimal difference in the average length and coverage rate between the exact computation and the rectangle-rule approximation for the denominators of f_1 and f_2. Therefore, for simplicity, we use the latter method in all future experiments. Secondly, as the iteration number T increases, all averaged coverage rates gradually grow towards 95% while the length of the intervals decreases. Finally, a larger value of m slightly reduces the average coverage rate but slightly decreases the length of the asymptotic confidence intervals. The impact of m on the performance is minimal, suggesting that f_2 could be used without further considerations.

The benchmark method, with a value of $B = 200$, reaches an average coverage rate of 95% after 5×10^3 iterations, while our method f_2 accomplishes a similar coverage rate in 10^4 iterations. At first glance, Figure 3.2 and Table 3.2 suggest that the online Bootstrap method is more sample efficient as it requires fewer iterations to achieve the nominal coverage rate of 95%. However, this efficiency is contingent on the availability of multiple oracles that can compute $\{\boldsymbol{H}(\boldsymbol{x}_t^b, \xi_t)\}_{b\in[B]}$ for different iterates $\{\boldsymbol{x}_t^b\}_{b\in[B]}$ at a given data ξ_t. In practical scenarios where one-trajectory sampling is performed, accessing multiple oracles is often not feasible due to limited control over the environment.① By contrast, our method does not require multiple oracles and even uses fewer gradient computations compared to the benchmark.② Table 3.2 demonstrates that given the same budget of gradient calls (e.g., 5×10^4), our method produces higher average coverage rates. Additionally, the bootstrap method is time-consuming, with the completion time of 5×10^3 updates taking approximately 1.5 hours for 500 repeated experiments, roughly equal to the time it takes for our method f_2 to finish 5×10^4 updates. Finally, an improperly chosen small value for $\boldsymbol{B}$ will reduce the performance, while a reasonably large value for $\boldsymbol{B}$ increases computation and memory demands. The difficulty of tuning a reasonable value for $\boldsymbol{B}$ contributes to the final disadvantage of the bootstrap method.

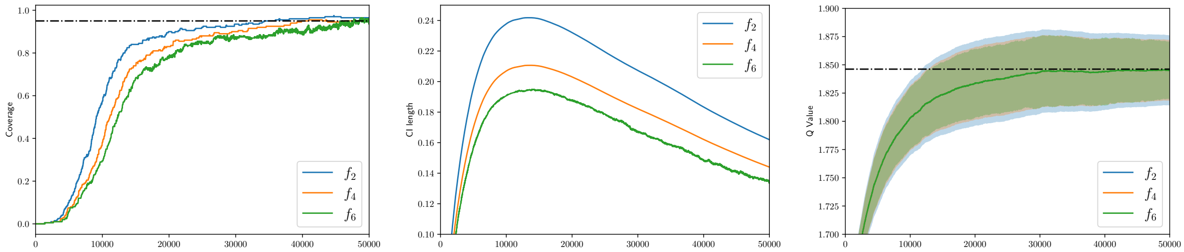
3.6.2 Asynchronous Q-Learning

In this experiment, we evaluate the performance in asynchronous Q-Learning with different methods (f_2, f_4, f_6) in a random MDP. The behavior policy is set to be uniformly random, and the target of the estimation is $\mathbb{E}_{(s,a)\sim\text{Uniform}(\mathcal{S}\times\mathcal{A})}Q^\star(s,a)$ where $Q^\star$ is the optimal Q-value function. We did not include the online bootstrap method of Ramprasad, Li, Yang, Wang, Sun, Cheng [43] in our comparison due to two reasons. Firstly, it is not theoretically guaranteed in nonlinear SA settings. Secondly, a direct application of the method resulted in unreasonable confidence intervals.

From the results shown in Figure 3.3, all of our methods reach the desired 95% coverage rate after approximately 4×10^4 iterations The length of the confidence intervals first increases and then decreases, which is due to the initialization of the length at zero, followed by the accumulation of errors, and finally the convergence. As expected, larger m values result in shorter confidence interval lengths, but slightly slower convergence of the empirical coverage.

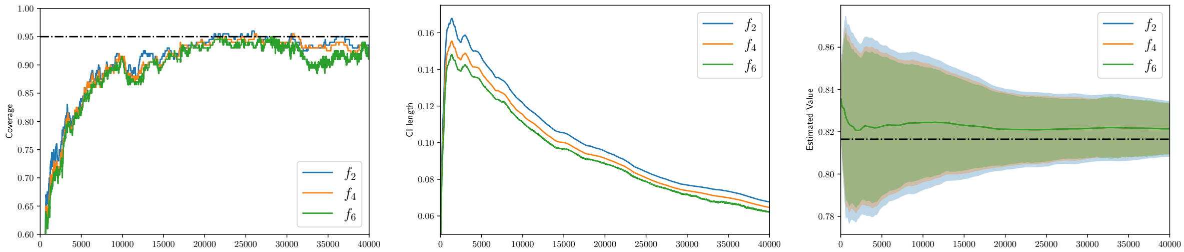
① Ramprasad, Li, Yang, Wang, Sun, Cheng [43] tested their algorithm in online game environments where rewards are deterministic and ξ_t is equal to the current state s_t of the underlying Markov chain. Hence, $\boldsymbol{H}(\boldsymbol{x}, \xi)$ is a deterministic function of $\boldsymbol{x}$ and the state s, making multiple gradient oracles accessible. However, in other applications, such as finance where rewards are random and Markov, accessing multiple oracles is not possible.

② It is worth noting that the online Bootstrap method requires $\boldsymbol{B} + 1$ gradient calls per iteration.



(a) Empirical coverage rates (b) Averaged CI lengths (c) The trajectory of averaged CIs

Figure 3.3 Performance of different inference methods for asynchronous Q-Learning. (a) shows the empirical coverage rates based on 200 repeated experiments. (b) shows the averaged confidence interval (CI) lengths therein. (c) shows the trajectory of averaged confidence intervals with shadows presenting their lengths. Black dot lines represent the nominal 95% coverage rate in (a) and the parameter of interest in (c).



(a) Empirical coverage rates (b) Averaged CI lengths (c) The trajectory of averaged CIs

Figure 3.4 Performance of different inference methods for logistic regression with Markovian data. (a) shows the empirical coverage rates based on 200 repeated experiments. (b) shows the averaged confidence interval (CI) lengths therein. (c) shows the trajectory of averaged confidence intervals with shadows presenting their lengths. Black dot lines represent the nominal 95% coverage rate in (a) and the target parameter in (c).

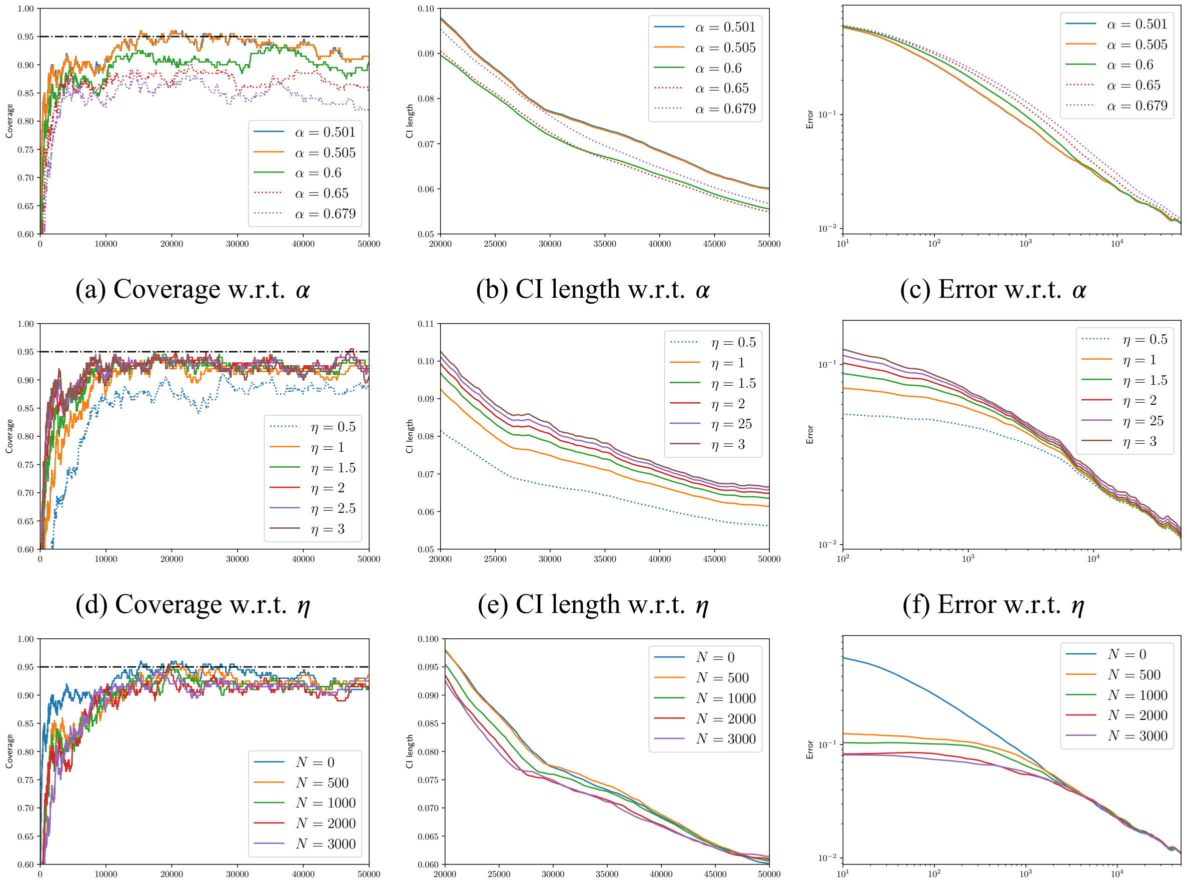
In Figure 3.3(c), we present the evolution of the averaged confidence intervals. After around 1.5×10^4 iterations, the averaged confidence interval starts to include the interest parameter with its center gradually increasing and converging to the interest parameter.

3.6.3 Logistic regression with Markovian data

In this experiment, we consider logistic regression with Markovian data. We take a similar simulation setup as Sun, Sun, Yin[122]. The observed data $\xi_t = (\boldsymbol{a}_t, y_t)$ is generated as the following manner

$$\boldsymbol{a}_t = \boldsymbol{A}\boldsymbol{a}_{t-1} + \boldsymbol{e}_1 W_t \text{ with } \boldsymbol{A}_{i,i-1} \overset{i.i.d.}{\sim} \text{Uniform}([0.8, 0.99]),\ W_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

$$y_t = \begin{cases} 1 & \text{with probability } S\left(\langle \boldsymbol{a}_t, \boldsymbol{x}^\star \rangle\right), \\ 0 & \text{with probability } 1 - S\left(\langle \boldsymbol{a}_t, \boldsymbol{x}^\star \rangle\right), \end{cases}$$



(a) Coverage w.r.t. α (b) CI length w.r.t. α (c) Error w.r.t. α

(d) Coverage w.r.t. η (e) CI length w.r.t. η (f) Error w.r.t. η

(g) Coverage w.r.t. N (h) CI length w.r.t. N (i) Error w.r.t. N

Figure 3.5 Sensitivity analysis for logistic regression with Markovian data. In these experiments, we chose f_2, set the step size to be $\eta_t = \eta t^{-\alpha}$ and treat $\boldsymbol{x}_N$ as the initial iterate for a warm-up. The perturbed parameters include α, η and N with the legend specifying the used values. (a) (d) (g) show the sensitivity of empirical coverage, (b) (e) (h) show the sensitivity of CI lengths, and (c) (f) (i) show the sensitivity of absolute errors.

where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is a subdiagonal matrix with only $\{\boldsymbol{A}_{i,i-1}\}_{1\le i\le d}$ non-zero, $\boldsymbol{e}_1$ is the first vector in the standard basis, and $S(x) = \frac{e^x}{1+e^x}$ is the sigmoid function. The target parameter in this experiment is $\boldsymbol{\theta}^\top \boldsymbol{x}^\star$, which is constructed similarly to the first experiment. By applying the update rule in Equation (3.2) to the negative log-likelihood objective, the experimental results are shown in Figure 3.4. All of our methods reach the desired 95% coverage rate, with f_2 having a slight advantage in terms of convergence speed. The confidence interval (CI) lengths decrease as the iteration progresses or as m increases. Figure 3.4(c) displays the trajectory of the average CI lengths, which start to include the target parameter from the very beginning.

Figure 3.5 displays the sensitivity of the results produced by our method to various parameters, including the step size parameter α, the step size scale η, and the warm-up iteration N. The empirical coverage rates and the averaged lengths of the confidence intervals are plotted with respect to each of these parameters. From Figure 3.5(a), it can be seen that the empirical coverage rates are relatively robust to changes in the step size parameter within the range $(0.5, 0.6)$. However, for larger values of α in the range $(0.6, 1)$, the empirical coverage rates begin to degrade. The optimal step size parameter predicted by Corollary 3.3.3 ($\alpha = 0.679$) is not seen to have an impact in this particular logistic regression experiment. This could be because the nonlinearity and Markovian data have a minimal impact, leading to $c_r \approx 0$ and $t_{\text{mix}} \approx 0$. In this case, the optimal α is close to 0.5, which is consistent with the results shown in Figure 3.5(a). Figures 3.5(b) and 3.5(c) provide insight into why smaller values of α result in faster convergence of the empirical coverage: for smaller α, the center of the confidence intervals converges more quickly, while the length of the intervals is even wider than for larger values of α. Additionally, from the middle and lowest row of Figure 3.5, both the absolute estimation error and the length of the confidence intervals converge more quickly for smaller values of η or larger values of N. However, these advantages are relatively small and our methods are robust to changes in the step size scale η and the warm-up iteration N.

3.7 Conclusion

From a methodological standpoint, in this chapter, we introduce a fully online statistical inference method for nonlinear stochastic approximation using a single trajectory of Markovian data. Our approach, motivated by the random scaling introduced in the last chapter, centers around constructing an asymptotic pivotal quantity through the application of a continuous scale-invariant functional f to the partial-sum process $\boldsymbol{\phi}_T$. To accomplish this, we propose a family of suitable functionals f_m that are indexed by $m \in \mathbb{N}$. In our simulations, we found that smaller values of m result in faster convergence of empirical coverage, although the confidence interval lengths may be slightly wider.

From a theoretical perspective, we demonstrate the validity of our approach through a functional central limit theorem and provide the first non-asymptotic upper bound on its weak convergence rate measured in the Lévy-Prokhorov metric. The asymptotic result in Equation (3.25) and the qualitative bound in Equation (3.26) for the coefficient-varying remainder process $\boldsymbol{\psi}_3$ can be leveraged in future studies on the weak convergence of iterative algorithms. Additionally, we present a semiparametric efficient lower bound to highlight the statistical efficiency of the partial-sum process $\boldsymbol{\phi}_T$. It is the most efficient RAL estimator among all RAL estimators with an asymptotic variance that attains the semiparametric efficient lower bound for all fractions $r \in [0, 1]$.

Chapter 4 Conclusion and Future Directions

4.1 Summary

In this dissertation, we investigated ways to conduct online statistical inference in federated learning (FL) and nonlinear stochastic approximation (SA), focusing on the Local SGD algorithm in FL and asynchronous Q-Learning in RL. Both of them are instances of nonlinear SAs, because they can be formulated as a stochastic iterative algorithm in the root finding problem of $\boldsymbol{g}(\boldsymbol{x}) := \int_{\Xi} \boldsymbol{H}(\boldsymbol{x}, \xi)\pi(d\xi) = \mathbf{0}$, where the root is denoted by $\boldsymbol{x}^\star$ that satisfies $\boldsymbol{g}(\boldsymbol{x}^\star) = \mathbf{0}$. The (possibly nonlinear) function $\boldsymbol{g}$ is the gradient of the aggregated global loss function in FL while being the Bellman equation in RL. Our target quantity is a linear functional of the true parameter $\boldsymbol{x}^\star$, which is $\boldsymbol{\theta}^\top \boldsymbol{x}^\star$ for a unit norm vector $\boldsymbol{\theta}$.

For Local SGD, we introduced two inference methods to construct confidence intervals: the plug-in method in Section 2.4.1 and the random scaling type method in Section 2.4.2. We establish either asymptotic normality or functional central limit theorem to support these methods. We compare these two methods in terms of their computational complexity and memory requirements. The plug-in method requires the access of noisy observations of the derivative of $\boldsymbol{g}(\boldsymbol{x})$ to estimate $\boldsymbol{G}$, i.e., the ability to evaluate $\nabla \boldsymbol{H}(\boldsymbol{x}, \xi)$, which satisfies $\mathbb{E}_{\xi\sim\pi} \nabla \boldsymbol{H}(\boldsymbol{x}, \xi) = \nabla \boldsymbol{g}(\boldsymbol{x})$. To obtain a consistent estimator for the asymptotic variance, the plug-in method needs to store both estimates of $\boldsymbol{G}$ and $\boldsymbol{S}$ and take the inverse of $\boldsymbol{G}$ at each iteration. This requires $\mathcal{O}(d^2)$ memory space and $\mathcal{O}(d^3)$ computation complexity. In contrast, the random scaling method does not attempt to estimate the asymptotic variance. It formulates an asymptotically pivotal statistic by utilizing the trajectory information, which is more computationally efficient and memory-friendly, requiring only $\mathcal{O}(d)$ memory space and $\mathcal{O}(d)$ computation complexity at each iteration.

For nonlinear SA, due to the lack of Hessian information, we propose a nonparametric inference following the spirit of random scaling in Section 3.5. Under the existence of Markovian data, we establish a functional central limit theorem for the partial-sum process $\boldsymbol{\phi}_T$. Furthermore, we propose a semeparametric efficient lower bound for the asymptotic variance and a non-parametric upper bound for weak convergence quantified by the Lévy-Prokhorov distance. By selecting any continuous scale-invariant functional f, this quantity $f(\boldsymbol{\phi}_T)$ becomes an asymptotic pivotal statistic, allowing us to construct an asymptotically valid confidence interval. We proposed a family of functionals f_m and analyze its several aspects including the rejection probability and confidence lengths. In the numerical part, we compare our method with another popular approach namely the online bootstrap method[43]. In general, despite its popularity, bootstrap is not suitable for trajectory data analysis where a complete control of the environment is lacked because it requires multiple oracles. Additionally, the memory and computation complexity of bootstrap methods are much more severe because they maintain multiple (say B) perturbed iterates and need to update them at each iteration. Hence, the complexity depends on the value of B. To ensure the estimated confidence intervals stable, B should be set sufficiently large, increasing the handwork of parameter tuning.

4.2 Future Directions

There are many other interesting issues presented in this dissertation that can be explored in future work.

Statistical analysis for decentralized data We first focus on the distributed learning setting. Recall that federated learning is a special case of distributed learning.

1. Weaker assumptions: One direction is to relax the current assumptions and consider Local SGD for more challenging optimization problems (e.g., non-smooth or non-convex problems). The quantile regressions would be an important application of non-smooth optimization. The use of neural networks forces us to step into the world of non-convex optimization.
2. Asymptotic analysis for other FL methods: Our theory shows that Local SGD enjoys statistical optimality in an asymptotic sense, and it is definitely not also optimal in finite-time convergence[60]. We can analyze the asymptotic normality of other state-of-the-art algorithms in FL. For example, Karimireddy, Kale, Mohri, Reddi, Stich, Suresh [84] proposed a new algorithm using control variates to remove the effect of data heterogeneity, which achieves a better non-asymptotic convergence rate.
3. Double efficient algorithms: From an theoretical perspective, it would be interesting to investigate algorithms that are efficient both asymptotically and non-asymptotically. The former means the produced estimate, say $\boldsymbol{x}_T$, enjoys an asymptotic normality where the asymptotic variance matrix nearly matches the Cramer-Rao lower bound, while the latter means the convergence rate of $\boldsymbol{x}_T$ is as tight as possible in terms of T and other instant-dependent quantities. This question has been answered partially[68, 103] in the context of the single-agent setting. It would be interesting to investigate similar double efficient algo-

rithms as well inference methods to handle the challenge in the big data era[3].

Random scaling for online statistical inference The idea of random scaling motivates the online inference method introduced in Chapter 3. Despite the progress made in our paper, several avenues for further research remain.

1. High-dimensional cases: It is important to extend our methods to high-dimensional scenarios. One possible solution is to use a proximal Robbins-Monro method[157] with ℓ_1 penalization in cases where the root $\boldsymbol{x}^\star$ is high-dimensional but sparse in its coordinates. The other possible method is stochastic mirror descent[158-160]. Although the last-iterate process of online ℓ_1 penalized problems has been analyzed[134], the partial-sum process of proximal methods has yet to be similarly studied.
2. Other stochastic optimization methods: We essentially establish a functional central limit theorem for SGD. Recent years witness many progresses in stochastic optimization and many efficient algorithms have been proposed. For example, the Nesterov accelerated gradient and proximal gradient descent for composite optimization, and variance reduced methods for finite-sum minimization[161-162]. It would be very interesting to establish similar FCLTs for these variants of SGD. In this way, we expect to achieve fast convergence and efficient statistical inference simultaneously. However, for these more delicate algorithms, our iterative analysis method should be modified, but we speculate the high-level picture is still similar.
3. Combination with other inference methods: Recent years have many nonparametric inference methods been proposed. The bootstrap replies on the multiple oracles[43], while the conformal inference methods depends on the exchangeability of observed data[163]. How to combine them with random scaling in an organic way so as to take their advantages for online statistical inference would be another interesting future direction.
4. Other efficient functional f: Although we propose a family of functionals f's, it is not clear whether there exist other functionals that can be efficiently computed online and also have improved empirical performance in terms of smaller confidence interval lengths and faster convergence of empirical coverage. Establishing similar weak convergence rates for different functionals would allow for their theoretical comparison.
5. Lower bound for weak convergence: The tightness of our upper bound for weak convergence remains uncertain. Determining the minimax lower bound for weak convergence and finding the optimal iterative procedure to match it are ongoing open problems.
6. Functional data analysis: In this work, we essentially consider data in the Euclidean space,

while statistical methods for analyzing functional data have been extensively developed in the past decades[164]. It is typically considered challenging to conduct statistical inference for streaming functional data. When data points are functions, it is more appropriate to consider stochastic approximation methods in Banach spaces. Recently, Mou, Khamaru, Wainwright, Bartlett, Jordan [103] studied the problem of estimating the fixed point of a contractive operator defined on a separable Banach space. They proposed a variance-reduced stochastic approximation method that achieves the local asymptotic minimax risk non-asymptotically. Xie, Shi, Sang, Shang, Jiang, Kong [165] proposed an online bootstrap resampling procedure to conduct inference for functional linear regression in a similar manner as Ramprasad, Li, Yang, Wang, Sun, Cheng [43] did in their online bootstrap linear SA paper. It is possible and would be interesting to extend the random scaling method for functional data.

References

[1] Shao J. Mathematical statistics[M]. [S.l.]: Springer Science & Business Media, 2003.

[2] Van der Vaart A W. Asymptotic statistics[M]. [S.l.]: Cambridge university press, 2000.

[3] Fan J, Ma C, Wang K, Zhu Z. Modern data modeling: Cross-fertilization of the two cultures[J]. Observational Studies, 2021, 7(1): 65-76.

[4] Sirur S, Nurse J R, Webb H. Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR)[C]//International Workshop on Multimedia Privacy and Security. [S.l. : s.n.], 2018: 88-95.

[5] McMahan B, Moore E, Ramage D, Hampson S, y Arcas B A. Communication-efficient learning of deep networks from decentralized data[C]//International Conference on Artificial Intelligence and Statistics. [S.l. : s.n.], 2017.

[6] Li T, Sahu A K, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[7] Kairouz P, McMahan H B, Avent B, Bellet A, Bennis M, Bhagoji A N, Bonawitz K, Charles Z, Cormode G, Cummings R, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1–2): 1-210.

[8] Anderson T. The theory and practice of online learning[M]. [S.l.]: Athabasca University Press, 2008.

[9] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. [S.l.]: MIT press, 2018.

[10] Stich S U. Local SGD converges fast and communicates little[J]. ArXiv preprint arXiv:1805.09767, 2018.

[11] Lin T, Stich S U, Patel K K, Jaggi M. Don’t use large mini-batches, use Local SGD[C]//International Conference on Learning Representations. [S.l. : s.n.], 2019.

[12] Li X, Yang W, Wang S, Zhang Z. Communication efficient decentralized training with multiple local updates[J]. ArXiv preprint arXiv:1910.09126, 2019.

[13] Bayoumi A K R, Mishchenko K, Richtárik P. Tighter theory for local SGD on identical and heterogeneous data[C]//International Conference on Artificial Intelligence and Statistics. [S.l. : s.n.], 2020: 4519-4529.

[14] Koloskova A, Loizou N, Boreiri S, Jaggi M, Stich S. A unified theory of decentralized SGD with changing topology and local updates[C]//International Conference on Machine Learning. [S.l. : s.n.], 2020: 5381-5393.

[15] Woodworth B, Patel K K, Stich S, Dai Z, Bullins B, Mcmahan B, Shamir O, Srebro N. Is local SGD better than minibatch SGD?[C]//International Conference on Machine Learning. [S.l. : s.n.], 2020: 10334-10343.

[16] Woodworth B E, Patel K K, Srebro N. Minibatch vs Local SGD for heterogeneous distributed learning[C]//Advances in Neural Information Processing Systems: vol. 33. [S.l. : s.n.], 2020: 6281-6292.

[17] Watkins C. Learning from delayed rewards[D]. King's College, Cambridge University, 1989.

[18] Qu G, Wierman A. Finite-time analysis of asynchronous stochastic approximation and Q-learning[C]//Conference on Learning Theory. [S.l. : s.n.], 2020: 3185-3205.

[19] Li G, Wei Y, Chi Y, Gu Y, Chen Y. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction[C]//Advances in neural information processing systems: vol. 33. [S.l. : s.n.], 2020: 7031-7043.

[20] Chen Z, Maguluri S T, Shakkottai S, Shanmugam K. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants[J]. ArXiv preprint arXiv:2102.01567, 2021.

[21] Li X, Yang W, Jiadong L, Zhang Z, Jordan M I. A statistical analysis of Polyak-Ruppert averaged Q-learning[C]//International Conference on Artificial Intelligence and Statistics: vol. 206. [S.l. : s.n.], 2023.

[22] Chen X, Xie M g. A split-and-conquer approach for analysis of extraordinarily large data[J]. Statistica Sinica, 2014: 1655-1684.

[23] Battey H, Fan J, Liu H, Lu J, Zhu Z. Distributed estimation and inference with statistical guarantees[J]. The Annals of Statistics, 2018: 1352-1382.

[24] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates[J]. Journal of Machine Learning Research, 2015, 16: 3299-3340.

[25] Lee J D, Liu Q, Sun Y, Taylor J E. Communication-efficient sparse regression[J]. The Journal of Machine Learning Research, 2017, 18(1): 115-144.

[26] Chang X, Lin S B, Wang Y, et al. Divide and conquer local average regression[J]. Electronic Journal of Statistics, 2017, 11(1): 1326-1350.

[27] Dobriban E, Sheng Y. Distributed linear regression by averaging[J]. The Annals of Statistics, 2021, 49(2): 918-943.

[28] Wang S. A sharper generalization bound for divide-and-conquer ridge regression[C]//AAAI Conference on Artificial Intelligence. [S.l. : s.n.], 2019.

[29] Blum J R. Approximation methods which converge with probability one[J]. The Annals of Mathematical Statistics, 1954: 382-386.

[30] Polyak B T, Juditsky A B. Acceleration of stochastic approximation by averaging[J]. SIAM journal on control and optimization, 1992, 30(4): 838-855.

[31] Anastasiou A, Balasubramanian K, Erdogdu M A. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT[C]//Conference on Learning Theory. [S.l. : s.n.], 2019: 115-137.

[32] Mou W, Li C J, Wainwright M J, Bartlett P L, Jordan M I. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration[C]//Conference on Learning Theory. [S.l. : s.n.], 2020: 2947-2997.

[33] Shamir O, Srebro N, Zhang T. Communication-efficient distributed optimization using an approximate newton-type method[C]//International conference on machine learning. [S.l. : s.n.], 2014: 1000-1008.

[34] Liu Z, Li M. Non-asymptotic analysis in kernel ridge regression[J]. ArXiv preprint arXiv:2006.01350, 2020.

[35] Jordan M I, Lee J D, Yang Y. Communication-efficient distributed statistical learning[J]. Stat, 2016, 1050: 25.

[36] Fan J, Guo Y, Wang K. Communication-efficient accurate statistical estimation[J]. ArXiv preprint arXiv:1906.04870, 2019.

[37] Zinkevich M, Weimer M, Li L, Smola A J. Parallelized stochastic gradient descent[C]//Advances in neural information processing systems. [S.l. : s.n.], 2010: 2595-2603.

[38] Chen H, Lu W, Song R. Statistical inference for online decision making via stochastic gradient descent[J]. Journal of the American Statistical Association, 2021, 116(534): 708-719.

[39] Zhu W, Chen X, Wu W B. Online covariance matrix estimation in stochastic gradient descent[J]. Journal of the American Statistical Association, 2021: 1-12.

[40] White M, White A. Interval estimation for reinforcement-learning algorithms in continuous-state domains[C]//Advances in Neural Information Processing Systems: vol. 23. [S.l. : s.n.], 2010.

[41] Hanna J P, Stone P, Niekum S. Bootstrapping with models: Confidence intervals for off-policy evaluation[C]//AAAI Conference on Artificial Intelligence. [S.l. : s.n.], 2017.

[42] Hao B, Ji X, Duan Y, Lu H, Szepesvari C, Wang M. Bootstrapping fitted Q-Evaluation for off-Policy inference[C]//International Conference on Machine Learning: vol. 139. [S.l.]: PMLR, 2021: 4074-4084.

[43] Ramprasad P, Li Y, Yang Z, Wang Z, Sun W W, Cheng G. Online bootstrap inference for policy evaluation in reinforcement learning[J]. Journal of the American Statistical Association, 2021.

[44] Li X, Wang S, Chen K, Zhang Z. Communication-efficient fistributed SVD via local power iterations[C]//International Conference on Machine Learning. [S.l. : s.n.], 2020.

[45] Grammenos A, Mendoza Smith R, Crowcroft J, Mascolo C. Federated principal component analysis[C]//Advances in Neural Information Processing Systems: vol. 33. [S.l. : s.n.], 2020: 6453-6464.

[46] Li X, Wang S, Chen K, Zhang Z. Communication-efficient distributed SVD via local power iterations[C]//International Conference on Machine Learning. [S.l. : s.n.], 2021: 6504-6514.

[47] Mohri M, Sivek G, Suresh A T. Agnostic federated learning[C]//International Conference on Machine Learning. [S.l. : s.n.], 2019: 4615-4625.

[48] Reisizadeh A, Farnia F, Pedarsani R, Jadbabaie A. Robust federated learning: The case of affine distribution shifts[C]//Advances in Neural Information Processing Systems: vol. 33. [S.l.]: Curran Associates, Inc., 2020: 21554-21565.

[49] Deng Y, Mahdavi M. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency[C]//International Conference on Artificial Intelligence and Statistics. [S.l. : s.n.], 2021: 1387-1395.

[50] Li X, Liang J, Chang X, Zhang Z. Statistical estimation and online inference via Local SGD[C]// Loh P L, Raginsky M. Conference on Learning Theory: vol. 178. [S.l.]: PMLR, 2022: 1613-1661.

[51] Li X, Huang K, Yang W, Wang S, Zhang Z. On the convergence of FedAvg on non-iid data[C]// International Conference on Learning Representations. [S.l. : s.n.], 2019.

[52] Borkar V S. Stochastic approximation: A dynamical systems viewpoint[M]. [S.l.]: Springer, 2009.

[53] Benveniste A, Métivier M, Priouret P. Adaptive algorithms and stochastic approximations[M]. [S.l.]: Springer Science & Business Media, 2012.

[54] Kushner H, Yin G G. Stochastic approximation and recursive algorithms and applications[M]. [S.l.]: Springer Science & Business Media, 2003.

[55] Li X, Liang J, Zhang Z. Online statistical inference for nonlinear stochastic approximation with Markovian data[J]. ArXiv preprint arXiv:2302.07690, 2023.

[56] Kearns M, Mansour Y, Ng A Y. A sparse sampling algorithm for near-optimal planning in large Markov decision processes[J]. Machine learning, 2002, 49(2): 193-208.

[57] Billingsley P. Convergence of probability measures[M]. [S.l.]: John Wiley & Sons, 2013.

[58] Ruppert D. Efficient estimations from a slowly convergent Robbins-Monro process[R]. [S.l.]: Cornell University Operations Research, 1988.

[59] Duchi J C, Ruan F. Asymptotic optimality in stochastic optimization[J]. The Annals of Statistics, 2021, 49(1): 21-48.

[60] Woodworth B E, Bullins B, Shamir O, Srebro N. The min-max complexity of distributed stochastic convex optimization with intermittent communication[C]//Conference on Learning Theory. [S.l. : s.n.], 2021: 4386-4437.

[61] Chen X, Lee J D, Tong X T, Zhang Y, et al. Statistical inference for model parameters in stochastic gradient descent[J]. The Annals of Statistics, 2020, 48(1): 251-273.

[62] Lee S, Liao Y, Seo M H, Shin Y. Fast and robust online inference with stochastic gradient descent via random scaling[C]//AAAI Conference on Artificial Intelligence: vol. 36. [S.l. : s.n.], 2022: 7381-7389.

[63] Kiefer N M, Vogelsang T J, Bunzel H. Simple robust testing of regression hypotheses[J]. Econometrica, 2000, 68(3): 695-714.

[64] Sun Y. Let’ s fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference[J]. Journal of Econometrics, 2014, 178: 659-677.

[65] Jordan M I, Lee J D, Yang Y. Communication-efficient distributed statistical inference[J]. Journal of the American Statistical Association, 2019, 114(526): 668-681.

[66] Chen X, Liu W, Zhang Y. First-order newton-type estimator for distributed estimation and inference[J]. Journal of the American Statistical Association, 2021: 1-40.

[67] Su W J, Zhu Y. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent[J]. ArXiv preprint arXiv:1802.04876, 2018.

[68] Li C J, Mou W, Wainwright M, Jordan M. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm[C]//Conference on Learning Theory. [S.l. : s.n.], 2022: 909-981.

[69] Kushner H J, Yang J. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes[J]. SIAM Journal on Control and Optimization, 1993, 31(4): 1045-1062.

[70] Wang Y, Wu S. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms[J]. The Journal of Machine Learning Research, 2020, 21(1): 8179-8281.

[71] Scott D W. Multivariate density estimation: Theory, practice, and visualization[M]. [S.l.]: John Wiley & Sons, 2015.

[72] Abadir K M, Paruolo P. Two mixed normal densities from cointegration analysis[J]. Econometrica: Journal of the Econometric Society, 1997: 671-680.

[73] Mania H, Pan X, Papailiopoulos D, Recht B, Ramchandran K, Jordan M I. Perturbed iterate analysis for asynchronous stochastic optimization[J]. SIAM Journal on Optimization, 2017, 27(4): 2202-2229.

[74] Stich S U, Cordonnier J B, Jaggi M. Sparsified SGD with memory[C]//Advances in Neural Information Processing Systems (NIPS). [S.l. : s.n.], 2018: 4447-4458.

[75] Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-iid data[J]. ArXiv preprint arXiv:1806.00582, 2018.

[76] Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge[C]//ICC 2019-2019 IEEE international conference on communications (ICC). [S.l. : s.n.], 2019: 1-7.

[77] Yuan H, Ma T. Federated accelerated stochastic gradient descent[J]. Advances in Neural Information Processing Systems, 2020, 33.

[78] Yuan H, Zaheer M, Reddi S. Federated composite optimization[C]//International Conference on Machine Learning. [S.l. : s.n.], 2021: 12253-12266.

[79] Zheng Q, Chen S, Long Q, Su W. Federated f-differential privacy[C]//International Conference on Artificial Intelligence and Statistics. [S.l. : s.n.], 2021: 2251-2259.

[80] Wang J, Kolar M, Srebro N, Zhang T. Efficient distributed learning with sparsity[C]//International Conference on Machine Learning. [S.l. : s.n.], 2017: 3636-3645.

[81] Jordan M I, Lee J D, Yang Y. Communication-efficient distributed statistical inference[J]. Journal of the American Statistical Association, 2018.

[82] Wang J, Joshi G. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms[J]. The Journal of Machine Learning Research, 2021, 22(1): 9709-9758.

[83] Haddadpour F, Kamani M M, Mahdavi M, Cadambe V. Local SGD with periodic averaging: Tighter qnalysis and adaptive synchronization[C]//Advances in Neural Information Processing Systems: vol. 32. [S.l. : s.n.], 2019: 11082-11094.

[84] Karimireddy S P, Kale S, Mohri M, Reddi S, Stich S, Suresh A T. SCAFFOLD: Stochastic controlled averaging for federated learning[C]//International Conference on Machine Learning. [S.l. : s.n.], 2020: 5132-5143.

[85] Liang X, Shen S, Liu J, Pan Z, Chen E, Cheng Y. Variance reduced Local SGD with lower communication complexity[J]. ArXiv preprint arXiv:1912.12844, 2019.

[86] Pathak R, Wainwright M J. FedSplit: An algorithmic framework for fast federated optimization[C]//Advances in Neural Information Processing Systems: vol. 33. [S.l. : s.n.], 2020: 7057-7066.

[87] Zhang X, Hong M, Dhople S, Yin W, Liu Y. FedPD: A federated learning framework with adaptivity to non-iid data[J]. IEEE Transactions on Signal Processing, 2021, 69: 6055-6070.

[88] Zhao T, Cheng G, Liu H. A partially linear framework for massive heterogeneous data[J]. Annals of statistics, 2016, 44(4): 1400.

[89] Wang B, Fang Y, Lian H, Liang H. Additive partially linear models for massive heterogeneous data[J]. Electronic Journal of Statistics, 2019, 13(1): 391-431.

[90] Liang J, Han Y, Li X, Zhang Z. Asymptotic behaviors of projected stochastic approximation: A jump diffusion perspective[C]//Advances in Neural Information Processing Systems. [S.l. : s.n.], 2022.

[91] Gu X, Lyu K, Huang L, Arora S. Why (and when) does Local SGD generalize better than SGD?[C]//International Conference on Learning Representations. [S.l. : s.n.], 2023.

[92] Deng W, Ma Y A, Song Z, Zhang Q, Lin G. On convergence of federated averaging Langevin dynamics[J]. ArXiv preprint arXiv:2112.05120, 2021.

[93] Godichon-Baggioni A. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms[J]. Journal of Statistical Planning and Inference, 2019, 203: 1-19.

[94] Glynn P W, Whitt W. Estimating the asymptotic variance with batch means[J]. Operations Research Letters, 1991, 10(8): 431-435.

[95] Fang Y, Xu J, Yang L. Online bootstrap confidence intervals for the stochastic gradient descent estimator[J]. The Journal of Machine Learning Research, 2018, 19(1): 3053-3073.

[96] Fang Y. Scalable statistical inference for averaged implicit stochastic gradient descent[J]. Scandinavian Journal of Statistics, 2019, 46(4): 987-1002.

[97] Li T, Liu L, Kyrillidis A, Caramanis C. Statistical inference using SGD[C]//The AAAI Conference on Artificial Intelligence: vol. 32. [S.l. : s.n.], 2018.

[98] Liang T, Su W J. Statistical inference for the population landscape via moment-adjusted stochastic gradients[J]. Journal of the Royal Statistical Society, 2019.

[99] Robbins H, Monro S. A stochastic approximation method[J]. The annals of mathematical statistics, 1951: 400-407.

[100] Moulines E, Bach F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning[C]//Advances in Neural Information Processing Systems: vol. 24. [S.l. : s.n.], 2011.

[101] Meyn S. Control systems and reinforcement learning[M]. [S.l.]: Cambridge University Press, 2022.

[102] Borkar V, Chen S, Devraj A, Kontoyiannis I, Meyn S. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning[J]. ArXiv preprint arXiv:2110.14427, 2021.

[103] Mou W, Khamaru K, Wainwright M J, Bartlett P L, Jordan M I. Optimal variance-reduced stochastic approximation in Banach spaces[J]. ArXiv preprint arXiv:2201.08518, 2022.

[104] Mou W, Pananjady A, Wainwright M, Bartlett P. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation[C]//Conference on Learning Theory. [S.l.]: PMLR, 2022: 2060-2061.

[105] Adomavicius G, Zhang J. Stability of recommendation algorithms[J]. ACM Transactions on Information Systems (TOIS), 2012, 30(4): 1-31.

[106] Lee S, Liao Y, Seo M H, Shin Y. Fast inference for quantile regression with tens of millions of observations[J]. Available at SSRN 4263158, 2022.

[107] Merlevède F, Peligrad M, Utev S. Functional Gaussian approximation for dependent structures[M]. [S.l.]: Oxford University Press, 2019.

[108] Krizmanic D. On functional weak convergence for partial sum processes[J]. Electronic Communications in Probability, 2014, 19: 1-12.

[109] Merlevède F, Peligrad M, Utev S. Functional CLT for martingale-like nonstationary dependent structures[J]. Bernoulli, 2019, 25(4B): 3203-3233.

[110] Haeusler E. An exact rate of convergence in the functional central limit theorem for special martingale difference arrays[J]. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 1984, 65(4): 523-534.

[111] Liang F. Trajectory averaging for stochastic approximation MCMC algorithms[J]. The Annals of Statistics, 2010, 38(5): 2823-2856.

[112] Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function[J]. The Annals of Mathematical Statistics, 1952: 462-466.

[113] Ljung L. Analysis of recursive stochastic algorithms[J]. IEEE transactions on automatic control, 1977, 22(4): 551-575.

[114] Ljung L. On positive real transfer functions and the convergence of some recursive schemes[J]. IEEE Transactions on Automatic Control, 1977, 22(4): 539-551.

[115] Ma D J, Makowski A M, Shwartz A. Stochastic approximations for finite-state Markov chains[J]. Stochastic Processes and Their Applications, 1990, 35(1): 27-45.

[116] Debavelaere V, Durrleman S, Allassonnière S. On the convergence of stochastic approximations under a subgeometric ergodic Markov dynamic[J]. Electronic Journal of Statistics, 2021, 15(1): 1583-1609.

[117] Mou W, Pananjady A, Wainwright M J. Optimal pracle inequalities for projected fixed-point equations, with applications to policy evaluation[J]. Mathematics of Operations Research, 2022.

[118] Konda V R, Tsitsiklis J N. Convergence rate of linear two-time-scale stochastic approximation[J]. The Annals of Applied Probability, 2004, 14(2): 796-819.

[119] Kaledin M, Moulines E, Naumov A, Tadic V, Wai H T. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise[C]//Conference on Learning Theory. [S.l. : s.n.], 2020: 2144-2203.

[120] Durmus A, Moulines E, Naumov A, Samsonov S. Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation[J]. ArXiv preprint arXiv:2207.04475, 2022.

[121] Duchi J C, Agarwal A, Johansson M, Jordan M I. Ergodic mirror descent[J]. SIAM Journal on Optimization, 2012, 22(4): 1549-1578.

[122] Sun T, Sun Y, Yin W. On Markov chain gradient descent[C]//Advances in Neural Information Processing Systems: vol. 31. [S.l. : s.n.], 2018.

[123] Doan T T, Nguyen L M, Pham N H, Romberg J. Finite-time analysis of stochastic gradient descent under Markov randomness[J]. ArXiv preprint arXiv:2003.10973, 2020.

[124] Nagaraj D, Wu X, Bresler G, Jain P, Netrapalli P. Least squares regression with markovian data: Fundamental limits and algorithms[J]. Advances in neural information processing systems, 2020, 33: 16666-16676.

[125] Polyak B T. New stochastic approximation type procedures[J]. Automat. i Telemekh, 1990, 7(98-107): 2.

[126] Hájek J. Local asymptotic minimax and admissibility in estimation[C]//Proceedings of the sixth Berkeley symposium on mathematical statistics and probability: vol. 1. [S.l. : s.n.], 1972: 175-194.

[127] Toulis P, Airoldi E M. Asymptotic and finite-sample properties of estimators based on stochastic gradients[J]. The Annals of Statistics, 2017, 45(4): 1694-1727.

[128] Shi C, Song R, Lu W, Li R. Statistical inference for high-dimensional models via recursive online-score estimation[J]. Journal of the American Statistical Association, 2021, 116(535): 1307-1318.

[129] Chen X, Lai Z, Li H, Zhang Y. Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods[J]. ArXiv e-prints, 2021: arXiv-2102.

[130] Tsitsiklis J N. Asynchronous stochastic approximation and Q-learning[J]. Machine learning, 1994, 16(3): 185-202.

[131] Even-Dar E, Mansour Y, Bartlett P. Learning rates for Q-learning.[J]. Journal of machine learning Research, 2003, 5(1).

[132] Shi C, Zhang S, Lu W, Song R. Statistical inference of the value function for reinforcement learning in infinite-horizon settings[J]. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 2021.

[133] Abounadi J, Bertsekas D P, Borkar V. Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms[J]. SIAM Journal on Control and Optimization, 2002, 41(1): 1-22.

[134] Chao S K, Cheng G. A generalization of regularized dual averaging and its dynamics[J]. ArXiv preprint arXiv:1909.10072, 2019.

[135] Negrea J, Yang J, Feng H, Roy D M, Huggins J H. Statistical inference with stochastic gradient algorithms[J]. ArXiv preprint arXiv:2207.12395, 2022.

[136] Xie C, Zhang Z. A statistical online inference approach in averaged stochastic approximation[C]//Advances in Neural Information Processing Systems. [S.l. : s.n.], 2022.

[137] Andrieu C, Moulines É, Priouret P. Stability of stochastic approximation under verifiable conditions[J]. SIAM Journal on control and optimization, 2005, 44(1): 283-312.

[138] Hairer M, Mattingly J C. Yet another look at Harris’ ergodic theorem for Markov chains[C]//Seminar on Stochastic Analysis, Random Fields and Applications VI. [S.l. : s.n.], 2011: 109-117.

[139] Glynn P W, Meyn S P. A Liapounov bound for solutions of the Poisson equation[J]. The Annals of Probability, 1996: 916-931.

[140] Gadat S, Panloup F. Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm[J]. Stochastic Processes and their Applications, 2023, 156: 312-348.

[141] Feigin P D, Tweedie R L. Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments[J]. Journal of time series analysis, 1985, 6(1): 1-14.

[142] Karimi B, Miasojedow B, Moulines E, Wai H T. Non-asymptotic analysis of biased stochastic approximation scheme[C]//Conference on Learning Theory. [S.l. : s.n.], 2019: 1944-1974.

[143] Chen Z, Maguluri S T, Shakkottai S, Shanmugam K. Finite-sample analysis of stochastic approximation using smooth convex envelopes[J]. ArXiv preprint arXiv:2002.00874, 2020.

[144] Gadat S. Stochastic optimization algorithms, non asymptotic and asymptotic behaviour[J]. Lecture notes, University of Toulouse, 2017.

[145] Greenwood P E, Wefelmeyer W. Efficiency of empirical estimators for Markov chains[J]. The Annals of Statistics, 1995: 132-143.

[146] Tsiatis A A. Semiparametric theory and missing data[M]. [S.l.]: Springer, 2006.

[147] Prokhorov Y V. Convergence of random processes and limit theorems in probability theory[J]. Theory of Probability & Its Applications, 1956, 1(2): 157-214.

[148] Gibbs A L, Su F E. On choosing and bounding probability metrics[J]. International statistical review, 2002, 70(3): 419-435.

[149] Borovkov A A. On the rate of convergence for the invariance principle[J]. Theory of Probability & Its Applications, 1974, 18(2): 207-225.

[150] Kubilius K. Rate of convergence in the functional central limit theorem for semimartingales[J]. Lithuanian Mathematical Journal, 1985, 25(1): 44-52.

[151] Wang X J, Hu S H. The Berry–Esseen bound for ρ-mixing random variables and its applications in nonparametric regression model[J]. Theory of Probability & Its Applications, 2019, 63(3): 479-499.

[152] Durieu O, Volnỳ D. Comparison between criteria leading to the weak invariance principle[C]// Annales de l'IHP Probabilités et statistiques. [S.l. : s.n.], 2008: 324-340.

[153] Gordin M, Peligrad M. On the functional central limit theorem via martingale approximation[J]. Bernoulli, 2011, 17(1): 424-440.

[154] Whitt W. Proofs of the martingale FCLT[J]. Probability Surveys, 2007, 4: 268-302.

[155] Greenwood P E, Wefelmeyer W. Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models[J]. Theory of Probability & Its Applications, 1998, 42(1): 103-111.

[156] Kartashov N. Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I[J]. Theory of Probability & Its Applications, 1986, 30(2): 247-259.

[157] Toulis P, Horel T, Airoldi E M, et al. The proximal Robbins–Monro method[J]. Journal of the Royal Statistical Society Series B, 2021, 83(1): 188-212.

[158] Duchi J C, Agarwal A, Wainwright M J. Dual averaging for distributed optimization: Convergence analysis and network scaling[J]. IEEE Transactions on Automatic control, 2011, 57(3): 592-606.

[159] Bekri Y, Ilandarideva S, Juditsky A B, Perchet V. Stochastic mirror descent for large-scale sparse recovery[J]., 2022.

[160] Juditsky A, Kulunchakov A, Tsyntseus H. Sparse recovery by reduced variance stochastic approximation[J]. Information and Inference: A Journal of the IMA, 2023, 12(2): 851-896.

[161] Gower R M, Schmidt M, Bach F, Richtárik P. Variance-reduced methods for machine learning[J]. Proceedings of the IEEE, 2020, 108(11): 1968-1983.

[162] Kulunchakov A. Stochastic optimization for large-scale machine learning: variance reduction and acceleration[D]. Université Grenoble Alpes [2020-....], 2020.

[163] Lei J, G’ Sell M, Rinaldo A, Tibshirani R J, Wasserman L. Distribution-free predictive inference for regression[J]. Journal of the American Statistical Association, 2018, 113(523): 1094-1111.

[164] Hsing T, Eubank R. Theoretical foundations of functional data analysis, with an introduction to linear operators[M]. [S.l.]: John Wiley & Sons, 2015.

[165] Xie J, Shi E, Sang P, Shang Z, Jiang B, Kong L. Scalable inference in functional linear regression with streaming data[J]. ArXiv preprint arXiv:2302.02457, 2023.

[166] Robbins H, Siegmund D. A convergence theorem for non negative almost supermartingales and some applications[G]//Optimizing methods in statistics. [S.l.]: Elsevier, 1971: 233-257.

[167] Hall P, Heyde C C. Martingale limit theory and its application[M]. [S.l.]: Academic press, 2014.

[168] Dharmadhikari S, Fabian V, Jogdeo K, et al. Bounds on the moments of martingales[J]. The Annals of Mathematical Statistics, 1968, 39(5): 1719-1723.

[169] Whitt W. Stochastic-process limits: An introduction to stochastic-process limits and their application to queues[M]. [S.l.]: Springer Science & Business Media, 2002.

[170] Chow Y. A martingale inequality and the law of large numbers[J]. Proceedings of the American Mathematical Society, 1960, 11(1): 107-111.

[171] Durrett R. Probability: Theory and examples (Edition 4.1)[M]. [S.l.]: Cambridge University Press, 2013.

[172] Burkholder D L. Sharp inequalities for martingales and stochastic integrals[J]. Astérisque, 1988, 157(158): 75-94.

[173] Chen H. Stochastic approximation and its applications[M]. [S.l.]: Springer Science & Business Media, 2002.

[174] Rosenblatt M. A central limit theorem and a strong mixing condition[J]. Proceedings of the national Academy of Sciences, 1956, 42(1): 43-47.

[175] Rio E. Inequalities and limit theorems for weakly dependent sequences[J]. HAL, 2013, 2013.

[176] Bradley R C. Basic properties of strong mixing conditions. A survey and some open questions[J]. Probability surveys, 2005, 2: 107-144.

Appendix A Omitted Proofs for Theorem 2.4.2

The proof idea of Theorem 2.4.2 has already been illustrated in Section 2.5. In this section, we provide the omitted proofs for the lemmas introduced therein.

A.1 Proof of Lemma 2.5.2

Proof of Lemma 2.5.2. Define $\mathscr{F}_t = \sigma(\{\xi_\tau^k\}_{1\le k\le K, 0\le \tau<t})$ by the natural filtration generated by ξ_τ^k's, so $\{\boldsymbol{x}_t^k\}_t$ is adapted to $\{\mathscr{F}_t\}_t$ and $\{\bar{\boldsymbol{x}}_{t_m}\}_m$ is adapted to $\{\mathscr{F}_{t_m}\}_m$. 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 $\mathbb{E}[\boldsymbol{h}_m|\mathscr{F}_{t_m}] = \nabla f(\bar{\boldsymbol{x}}_{t_m})$. 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 $\mathscr{F}_{t_m}$ in the last inequality gives

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

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

We then bound the last two terms in the right hand side of (A.1).

Step one For $\mathbb{E}[\|\boldsymbol{h}_m\|^2|\mathscr{F}_{t_m}]$, it follows that

$$\mathbb{E}[\|\boldsymbol{h}_m\|^2|\mathscr{F}_{t_m}] = \|\mathbb{E}[\boldsymbol{h}_m|\mathscr{F}_{t_m}]\|^2 + \mathbb{E}[\|\boldsymbol{h}_m - \mathbb{E}[\boldsymbol{h}_m|\mathscr{F}_{t_m}]\|^2|\mathscr{F}_{t_m}]$$

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

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 $\mathscr{F}_{t_m}$, so its conditional variance is E_m times smaller than the latter,

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

Lemma A.1.1. *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 3.2.2, it follows that*

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

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 3.2.2.

Proof of Lemma A.1.1. By Assumption 3.2.2, 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

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

Therefore, it follows that

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

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

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

□

With Lemma A.1.1, we have

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

Then, it follows that

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

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

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

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

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

where the first equality follows from the tower rule of conditional expectation and the second inequality follows from the expected L-smoothness in Assumption 2.3.1.

Combining the last two results, we have

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

where V_t is the residual error defined by

$$V_t = \sum_{k=1}^{K} p_k \mathbb{E}[\|\boldsymbol{x}_t^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathscr{F}_{t_m}]. \tag{A.3}$$

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 $\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2$.

Lemma A.1.2. *Under Assumptions 2.3.1 and 3.2.2, 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*

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

Proof of Lemma A.1.2. 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

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

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

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

where $C_3 = C_1 + 2\max_{k\in[K]}\|\nabla f_k(\boldsymbol{x}^\star)\|^2$ and $C_4 = 2C_2 + 4L^2$. The second inequality uses the L-smoothness to bound $\|\nabla f_k(\boldsymbol{x}_\tau^k) - \nabla f_k(\boldsymbol{x}^\star)\|$ and Lemma A.1.1 to bound $\mathbb{E}_{\xi_\tau^k}\|\varepsilon_k(\boldsymbol{x}_\tau^k)\|^2$ which yields

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

Therefore, by combing the last two results, we have

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

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

$$V_{t+1} = \sum_{k=1}^{K} p_k \mathbb{E}(\|\boldsymbol{x}_{t+1}^k - \bar{\boldsymbol{x}}_{t_m}\|^2|\mathscr{F}_{t_m}) \le \eta_m^2 E_m \sum_{\tau=t_m}^{t} \left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2 + C_4 V_\tau\right). \quad \text{(A.4)}$$

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

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

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

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

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. □

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

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

Note that $\gamma_m \to 0$ when m goes to infinity, which means there exists some m_0, such that for any $m \geq m_0$, we have $\gamma_m^2 C_4 \leq 1$ and $\gamma_m \leq \min\{\frac{1}{2L}, 1\}$. It implies that we can apply Lemma A.1.2 for sufficiently large m. Combining the two parts and plugging them into (A.1) yield for any $m \geq m_0$,

$$\begin{aligned}
\mathbb{E}[\Delta_{m+1} | \mathscr{F}_{t_m}] &\leq \Delta_m - \frac{\gamma_m}{2}\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \gamma_m^2 L \cdot \left[\|\nabla f(\bar{\boldsymbol{x}}_{t_m})\|^2 + \frac{C_1}{E_m} + \frac{C_2}{E_m}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2\right] \\
&\quad + \left(\frac{\gamma_m}{2} + \gamma_m^2 L\right)\gamma_m^2 L^2 \left(C_3 + C_4\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2\right) \\
&\leq \Delta_m - \gamma_m \mu \Delta_m + \gamma_m^2 L \cdot \left[\frac{C_1}{E_m} + \left(2L + \frac{2C_2}{\mu E_m}\right)\Delta_m\right] \\
&\quad + \left(\frac{\gamma_m}{2} + \gamma_m^2 L\right)\gamma_m^2 L^2 \left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&\leq \Delta_m - \gamma_m \mu \Delta_m + \gamma_m^2 L \cdot \left[C_1 + \left(2L + \frac{2C_2}{\mu}\right)\Delta_m\right] + \gamma_m^3 L^2 \left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&\leq \Delta_m - \gamma_m \mu \Delta_m + \gamma_m^2 L \cdot \left[C_1 + \left(2L + \frac{2C_2}{\mu}\right)\Delta_m\right] + \gamma_m^2 L^2 \left(C_3 + \frac{2C_4}{\mu}\Delta_m\right) \\
&= \left(1 + c_1\gamma_m^2\right)\Delta_m + c_2\gamma_m^2 - \mu\gamma_m\Delta_m, \qquad \text{(A.5)}
\end{aligned}$$

where

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

To conclude the proof, we need to apply the Robbins-Siegmund theorem[166].

Lemma A.1.3 (Robbins-Siegmund theorem). *Let $\{D_m, \beta_m, \alpha_m, \zeta_m\}_{m=0}^{\infty}$ be non-negative and adapted to a filtration $\{\mathscr{G}_m\}_{m=0}^{\infty}$, satisfying*

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

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 2.3.3, 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 (A.5), Lemma A.1.3 implies that $\Delta_m = f(\bar{\boldsymbol{x}}_{t_m}) - f(\boldsymbol{x}^\star)$ converges to a finite non-negative random variable Δ_∞ almost surely. Moreover, Lemma A.1.3 also ensures that

$$\mu \sum_{m=m_0}^{\infty} \gamma_m \Delta_m < \infty. \qquad \text{(A.6)}$$

If $\mathbb{P}(\Delta_m > 0) > 0$, then the left-hand side of (A.6) 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 $f(\bar{\boldsymbol{x}}_{t_m}) \to f(\boldsymbol{x}^\star)$ as well as $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^\star$ with probability one when m goes to infinity.

L_2 convergence We will obtain the L_2 convergence rate from (A.5). This part follows the same argument of Su, Zhu [67] (see Page 37-38 therein). For completeness, we conclude this section by presenting the proof of it. Taking expectation on both sides of (A.5),

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

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

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

Lemma A.1.4 (Lemma A.10 in Su, Zhu [67]). *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 $B_m \le \frac{\gamma_{m-1}(1-c_1\gamma_m)}{\gamma_m}B_{m-1} + c_2\gamma_m$, then* $\sup_m B_m < \infty$.

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

$$\sup_{0<m<\infty} \frac{\mathbb{E}\Delta_m}{\gamma_{m-1}} < C_5, \tag{A.7}$$

which immediately concludes that

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

□

A.2 Proof of Lemma 2.5.3

Proof of Lemma 2.5.3. 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 $\mathscr{F}_t = \sigma(\{\xi_\tau^k\}_{1\le k\le K, 0\le\tau<t})$ by the natural filtration generated by ξ_τ^k's and $\mathscr{G}_{m-1} = \mathscr{F}_{t_m}$. Then $\{\boldsymbol{\varepsilon}_m\}_{m=1}^\infty$ is a martingale difference with respect to $\{\mathscr{G}_m\}_{m=0}^\infty$ (for convention $\mathscr{G}_0 = \{\varnothing, \Omega\}$ if $\bar{\boldsymbol{x}}_0$ is deterministic, otherwise $\mathscr{G}_0 = \sigma(\bar{\boldsymbol{x}}_0)$): $\mathbb{E}[\boldsymbol{\varepsilon}_m|\mathscr{G}_{m-1}] = \boldsymbol{0}$.

The following lemma establishes an invariance principle which allows us to extend traditional martingale CLT. Interesting readers can find its proof in Hall, Heyde [167] (see Theorems 4.1, 4.2 and 4.4 therein).

Lemma A.2.1 (Invariance principles in the martingale CLT)**.** *Let* $\{S_n, \mathscr{G}_n\}_{n\geq 1}$ *be a zero-mean, square-integrable martingale with difference* $X_n = S_n - S_{n-1} (S_0 = 0)$. *Let* $U_n^2 = \sum_{m=1}^n \mathbb{E}[X_m^2 | \mathscr{G}_{m-1}]$ *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* $(0, 0)$, $(U_n^{-2}U_1^2, U_n^{-1}S_1)$, $(U_n^{-2}U_2^2, U_n^{-1}S_2)$, $\ldots$, $(1, U_n^{-1}S_n)$, *namely, for* $t \in [0, 1]$ *and* $0 \leq i \leq n-1$,

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

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

$$s_n^{-2}\sum_{m=1}^n \mathbb{E}[X_m^2 I(|X_m| \geq \varepsilon s_n)] \to 0, \tag{A.8}$$

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

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

Here $\boldsymbol{B}(t)$ *is the standard Brownian motion on* $[0, 1]$ *and* $\boldsymbol{C} = C[0, 1]$ *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 A.2.1 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 A.2.1 hold for such $\{X_m, \mathscr{G}_m\}_{m\geq 1}$.

The Linderberg condition For one thing, since $\bar{\boldsymbol{x}}_{t_m} \to \boldsymbol{x}^\star$ almost surely from Lemma 2.5.2, we have $\mathbb{E}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^{2+\delta_2} \lesssim 1$ from Assumption 3.2.2 when m is sufficiently large.

Lemma A.2.2 (Marcinkiewicz–Zygmund inequality and Burkholder inequality)**.** *If* $Z_1, \ldots, Z_n$ *are independent random vectors such that* $\mathbb{E}Z_m = 0$ *and* $\mathbb{E}|Z_m|^p < \infty$ *for* $1 \leq p < \infty$, *then*

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

where the C_p *are positive constants which depend only on p and not on the underlying distribution of the random variables involved. If* $Z_1, \ldots, Z_n$ *are martingale difference sequence, the above inequality still holds. It is named as Burkholder's inequality*[168].

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

$$\begin{aligned}\mathbb{E}|X_m|^{2+\delta_2} &\lesssim E_m^{-(1+\frac{\delta_2}{2})}\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^{n}\left|Z_m\right|^2\right)^{1+\frac{\delta_2}{2}} \lesssim E_m^{-(1+\frac{\delta_2}{2})}\mathbb{E}\left|Z_1\right|^{2+\delta_2}\\ &\lesssim E_m^{-(1+\frac{\delta_2}{2})}\|\boldsymbol{a}\|^{2+\delta_2}\mathbb{E}\|\varepsilon(\bar{\boldsymbol{x}}_{t_m})\|^{2+\delta_2} \lesssim E_m^{-1}. \end{aligned} \tag{A.9}$$

Moreover, from Assumption 3.2.2 and Lemma 2.5.2, we have that

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

Recall that $\sum_{m=1}^{T} E_m{}^{-1} \to \infty$ as $T \to \infty$. The Stolz–Cesàro theorem (Lemma A.2.3) implies that

$$\lim_{T\to\infty}\frac{s_T^2}{\sum_{m=1}^T \frac{1}{E_m}\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = \lim_{T\to\infty}\frac{\sum_{m=1}^T\frac{\boldsymbol{a}^\top\mathbb{E}\varepsilon(\bar{\boldsymbol{x}}_{t_m})\varepsilon(\bar{\boldsymbol{x}}_{t_m})^\top\boldsymbol{a}}{E_m}}{\sum_{m=1}^T \frac{1}{E_m}\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = \lim_{T\to\infty}\frac{\boldsymbol{a}^\top\mathbb{E}\varepsilon(\bar{\boldsymbol{x}}_{t_T})\varepsilon(\bar{\boldsymbol{x}}_{t_T})^\top\boldsymbol{a}}{\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a}} = 1. \tag{A.10}$$

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

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

The second condition We have established the divergence of $\{s_T^2\}_T$ in (A.10). Notice that

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

Therefore, from (A.10) and the Stolz–Cesàro theorem (Lemma A.2.3), it follows almost surely that

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

Lemma A.2.3 (Stolz–Cesàro theorem). *Let $\{a_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$ be two sequences of real numbers. Assume that $\{b_n\}_{n\ge1}$ 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 A.2.1 hold. Hence, by definition, $\zeta_T(r)\Rightarrow B(r)$ where

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

and $S_i=\sum_{m=1}^{i}X_m$. Since $s_T/U_T\to1$ almost surely and (A.10), it follows that

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

where $\boldsymbol{W}(r)$ is the d-dimensional standard Brownian motion. Recall that

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

Lemma A.2.4. *Under the same condition of Lemma 2.5.3, it follows that*

$$\sup_{r\in[0,1]}\left|\frac{\sqrt{t_T}}{T}U_T\zeta_T\left(\frac{U_{h(r,T)}^2}{U_T^2}\right)-\frac{\sqrt{t_T}}{T}U_T\zeta_T(r)\right|\to0\quad\textit{in probability.}$$

Hence,

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

By the arbitrariness of $\boldsymbol{a}$, it follows that[①]

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

[①] See the proof of Theorem 4.3.5. in Whitt[169] for more detail about how to argue multivariate weak convergence from univariate weak convergence along any direction.

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

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

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 $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{h(r,T)} \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}_m \Rightarrow \sqrt{v}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\boldsymbol{W}(r)$. □

A.3 Proof of Lemma A.2.4

Proof of Lemma A.2.4. From the Theorem A.2 of Hall, Heyde[167], 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 $\varepsilon > 0$, $\mathbb{P}(\sup_{|s-t|\le\delta}|\phi_n(s) - \phi_n(t)| \ge \varepsilon) \to 0$ uniformly in n as $\delta \to 0$. Since $\frac{\sqrt{t_T}}{T}U_T\zeta_T(r) \Rightarrow \sqrt{v}\boldsymbol{a}^\top\boldsymbol{S}^{1/2}\boldsymbol{W}(r)$, $\{\frac{\sqrt{t_T}}{T}U_T\zeta_T\}_T$ is tight. We denote the following notation for simplicity

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

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

$$\lim_{T\to\infty} p_T(r) = \lim_{T\to\infty}\frac{U_{h(r,T)}^2}{U_T^2} = \lim_{T\to\infty}\frac{s_{h(r,T)}^2}{s_T^2} = \lim_{T\to\infty}\frac{\sum_{m=1}^{h(r,T)}\frac{1}{E_m}}{\sum_{m=1}^{T}\frac{1}{E_m}} = r = p(r).$$

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.,

$$\lim_{T\to\infty}\sup_{r\in[0,1]}\left|p_T(r) - r\right| = 0. \tag{A.11}$$

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

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

With (A.11), for this δ, $\mathbb{P}(\sup_{r\in[0,1]}|p_T(r) - r| > \delta) \to 0$ as $T \to \infty$. It implies that

$$\limsup_{T\to\infty}\mathbb{P}\left(\sup_{r\in[0,1]}|\phi_T\left(p_T(r)\right) - \phi_T(r)| \ge \varepsilon\right)$$

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

Because η is arbitrary, we have shown that

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

□

A.4 Proof of Lemma 2.5.4

Proof of Lemma 2.5.4. Recall that $\boldsymbol{G} = \nabla^2 f(\boldsymbol{x}^\star)$, $\boldsymbol{s}_m = \bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star$ and

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

When $\|\boldsymbol{s}_m\| \le \delta_1$, by Assumption 2.3.1, $\|\nabla^2 f(s\boldsymbol{s}_m + \boldsymbol{x}^\star) - \nabla^2 f(\boldsymbol{x}^\star)\| \le sL'\|\boldsymbol{s}_m\|$, then

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

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

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

Hence,

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

By Lemma 2.5.2, $\boldsymbol{s}_m \to 0$ 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 $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\|\boldsymbol{s}_m\|^2 1_{\{\|\boldsymbol{s}_m\|\le\delta_1\}} = o_{\mathbb{P}}(1)$, which is implied by

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

It holds because $\frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\mathbb{E}\|\boldsymbol{s}_m\|^2 \lesssim \frac{\sqrt{t_T}}{T}\sum_{m=0}^{T}\gamma_m \to 0$ from Lemma 2.5.2 and Assumption 2.3.4.

□

A.5 Proof of Lemma 2.5.5

Proof of Lemma 2.5.5. In the proof of Lemma 2.5.2 (see the Part 2 therein), we have established for sufficiently large m,

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

where V_t is the residual error defined in (A.3) and $C_3, C_4 > 0$ are universal constants defined in Lemma A.1.2. Besides, Lemma 2.5.2 implies that $\mathbb{E}\|\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^\star\|^2 \lesssim \gamma_m \lesssim 1$. It follows that

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

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

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

from Lemma 2.5.2 and Assumption 2.3.4. □

A.6 Proof of Lemma 2.5.6

Proof of Lemma 2.5.6. If $\{E_m\}$ is uniformly bounded (i.e., there exists some C such that $1 \le E_m \le C$ for all m), the conclusion follows because

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

In the following, we instead assume E_m is non-decreasing in m (i.e., $1 \le E_m \le E_{m+1}$ 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 $m \ge N$, $0 \le H_m \le m\varepsilon$. Then

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

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

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

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

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

By the arbitrariness of ε, we complete the proof. □

A.7 Proof of Lemma 2.5.7

Proof of Lemma 2.5.7. 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 $h_i = \left[\frac{iT}{n}\right]$ $(i = 0, 1, \dots, n)$ 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 $\mathscr{A}$ whose complement is

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

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

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

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

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

which is implied by

$$\lim_{h_i\to\infty} \left\{ \sum_{m=0}^{h_i} \gamma_m^2 \left(\prod_{j=0}^{m} B_j \right)^{-2} \right\} \Big/ \left\{ \gamma_{h_i} \left(\prod_{j=0}^{h_i} B_j \right)^{-2} \right\}$$

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

as a result of the Stolz–Cesàro theorem (Lemma A.2.3). 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 $\mathscr{A}^c$ has diminishing probability, we focus on the event $\mathscr{A}$. We will prove that on the event $\mathscr{A}$ 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 \ ; \mathscr{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 \ ; \mathscr{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 \ ; \mathscr{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 \ ; \mathscr{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 \ ; \mathscr{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 \ ; \mathscr{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\}
\end{aligned}
$$

$$:= \sum_{i=0}^{n-1} \mathscr{P}_i,$$

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 $h \ge h_i = \left[\frac{iT}{n}\right] \to \infty$ as $T \to \infty$, $(1+o(\gamma_h))(1-\lambda\gamma_h) \le 1$ is solid and X_h^T is non-negative when T goes to infinity. Hence, each $\mathscr{P}_i$ 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 $\mathscr{P}_i$, we use Chow's inequality which is a generalization of Doob's inequality[170]. It follows that

$$\begin{aligned}
\mathscr{P}_i &= \mathbb{P}\left\{\frac{t_T^{1+\delta/2}}{T^{2+\delta}} \sup_{h\in[h_i,h_{i+1})} c_h Y_h \ge \varepsilon^{2+\delta}\right\} \\
&\le \frac{t_T^{1+\delta/2}}{\varepsilon^{2+\delta}T^{2+\delta}}\left\{c_{h_{i+1}-1}\mathbb{E}Y_{h_{i+1}-1} + \sum_{j=h_i+1}^{h_{i+1}-2}(c_i - c_{i+1})\mathbb{E}Y_j\right\}. \qquad \text{(A.12)}
\end{aligned}$$

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 A.2.2) to martingale differences[168]. 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 (A.9).

Plugging it into (A.12) yields that $\mathscr{P}_i$ is bounded by

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

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

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

where we use (ii) in Assumption 2.3.4 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

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

$$
\begin{aligned}
&\leq \limsup_{T\to\infty} \sum_{i=0}^{n-1} \mathscr{P}_i \\
&\lesssim \frac{1}{n^{\delta/2}}.
\end{aligned}
$$

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

Appendix B Omitted Proofs for Theorem 3.3.1

B.1 Proof of Lemma 3.4.1

Proof of Lemma 3.4.1. In the following, we use $a \lesssim b$ to denote $a \le Cb$ for an unimportant positive constant $C > 0$ that doesn't depends on p for simplicity. Let $C_{U,\boldsymbol{x}_t} = \kappa t_{\mathrm{mix}} \cdot (2L_H \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \sigma)$. By 2 in Lemma 3.2.2, we have $\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\| \le C_{U,\boldsymbol{x}_t}$.

1. By Assumption 3.2.1, it implies that

$$\begin{aligned}
\|\boldsymbol{r}_t\| &= \|\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{G}\boldsymbol{\Delta}_t\| \\
&\le \|\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{G}(\boldsymbol{x}_t - \boldsymbol{x}^\star)\| + \eta_t \|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\| \\
&\le \begin{cases} L_G \cdot \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \eta_t C_{U,\boldsymbol{x}_t} & \text{if } \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \le \delta_G \\ (L_H + \|\boldsymbol{G}\|) \cdot \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \eta_t C_{U,\boldsymbol{x}_t} & \text{if } \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \ge \delta_G \end{cases} \\
&\le \max\left\{ L_G, \frac{L_H + \|\boldsymbol{G}\|}{\delta_G} \right\} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \eta_t C_{U,\boldsymbol{x}_t}.
\end{aligned}$$

Since $\{\boldsymbol{x}_t\}_{t\ge0}$ satisfies the $(L^2, (1+\log t)\sqrt{\eta_t})$-consistency (from Assumption 3.2.6), $\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \lesssim \eta_t \log t$. As a result, when $T \to \infty$,

$$\frac{1}{\sqrt{T}} \sum_{t=0}^{T} \mathbb{E}\|\boldsymbol{r}_t\| \lesssim \frac{1}{\sqrt{T}} \sum_{t=0}^{T} \mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \frac{1}{\sqrt{T}} \sum_{t=0}^{T} \eta_t \lesssim \frac{\log T}{\sqrt{T}} \sum_{t=0}^{T} \eta_t \to 0.$$

2. By (3.17), we have $\mathbb{E}[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t)|\mathscr{F}_{t-1}] = \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})$ where $\mathscr{F}_t$ is the σ-field defined by $\mathscr{F}_t := \sigma(\{\xi_\tau\}_{0\le\tau\le t})$. Hence, $\{\boldsymbol{u}_t\}_{t\ge0}$ is a martingale difference sequence. By 2 in Lemma 3.2.2 and Assumption 3.2.2, $\sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_t$ is square integrable for all $r \in [0,1]$. By (3.17), we decompose $\boldsymbol{u}_t$ into two parts $\boldsymbol{u}_t = \boldsymbol{u}_{t,1} + \boldsymbol{u}_{t,2}$ where

$$\begin{aligned}
\boldsymbol{u}_{t,1} &= \left[\boldsymbol{U}(\boldsymbol{x}_t, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\right] - \left[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\right], \\
\boldsymbol{u}_{t,2} &= \left[\boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})\right].
\end{aligned} \tag{B.1}$$

It's clear that both $\{\boldsymbol{u}_{t,1}\}_{t\ge0}$ and $\{\boldsymbol{u}_{t,2}\}_{t\ge0}$ are also martingale difference sequences. We assert that $\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_t$ has the same asymptotic behavior as $\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_{t,2}$ due to

$$\mathbb{E} \sup_{r\in[0,1]} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_t - \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_{t,2} \right\|^2 = \mathbb{E} \sup_{r\in[0,1]} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \boldsymbol{u}_{t,1} \right\|^2 = o(1).$$

This is because from Doob's martingale inequality,

$$\mathbb{E}\sup_{r\in[0,1]}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_{t,1}\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{u}_{t,1}\|^2 \lesssim \frac{\log T}{T}\sum_{t=0}^{T}\eta_t \to 0,$$

where the last inequality uses the following result

$$\begin{aligned}\mathbb{E}\|\boldsymbol{u}_{t,1}\|^2 &\le 2\mathbb{E}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_t)-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\|^2+2\mathbb{E}\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2\\ &=2\mathbb{E}\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2+2\mathbb{E}\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2\\ &\overset{(a)}{\le} 4\mathbb{E}\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2\\ &\overset{(b)}{\le} 4\mathbb{E}(\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^p)^{\frac{2}{p}}\overset{(c)}{\le} 4L_U^2\cdot\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\overset{(d)}{\lesssim}\eta_t\cdot\log t.\end{aligned}$$

Here (a) follows from conditional Jensen's inequality, (b) follows from conditional Holder's inequality, (c) uses 4 in Lemma 3.2.2, and (d) uses $\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\lesssim\eta_t\cdot\log t$ from Assumption 3.2.6.

We then focus on the partial-sum process $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_{t,2}$. For one thing, by Assumption 3.2.2, $\{\boldsymbol{u}_{t,2}\}_{t\ge0}$ has uniformly bounded $p>2$ moments, which is because

$$\sup_{t\ge0}\mathbb{E}\|\boldsymbol{u}_{t,2}\|^p\le 2^{p-1}\sup_{t\ge0}\left[\mathbb{E}\|\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\|^p+\mathbb{E}\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^p\right]<\infty.$$

As a result, for any $\varepsilon>0$, as T goes to infinity,

$$\begin{aligned}\mathbb{E}\left\{\sum_{t=0}^{T}\mathbb{E}\left[\left\|\frac{\boldsymbol{u}_{t,2}}{\sqrt{T}}\right\|^2 1_{\{\|\boldsymbol{u}_{t,2}\|\ge\sqrt{T}\varepsilon\}}\middle|\mathscr{F}_{t-1}\right]\right\}&\le\frac{1}{\varepsilon^{\frac{p}{2}-1}T^{\frac{p}{2}}}\mathbb{E}\left\{\sum_{t=0}^{T}\mathbb{E}\left[\|\boldsymbol{u}_{t,2}\|^p\middle|\mathscr{F}_{t-1}\right]\right\}\\ &\le\frac{\sup_{t\ge0}\mathbb{E}\|\boldsymbol{u}_{t,2}\|^p}{\varepsilon^{\frac{p}{2}-1}T^{\frac{p}{2}-1}}\to0,\end{aligned}$$

which implies

$$\sum_{t=0}^{T}\mathbb{E}\left[\left\|\frac{\boldsymbol{u}_{t,2}}{\sqrt{T}}\right\|^2 1_{\{\|\boldsymbol{u}_{t,2}\|\ge\sqrt{T}\varepsilon\}}\middle|\mathscr{F}_{t-1}\right]\xrightarrow{p}0.$$

For another thing, we notice that

$$\begin{aligned}&\mathbb{E}\left[\left[\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\right]\left[\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\right]^\top\middle|\mathscr{F}_{t-1}\right]\\ =&\mathbb{E}\left[\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)^\top\middle|\mathscr{F}_{t-1}\right]-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})^\top\\ =&\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})^\top-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})^\top,\end{aligned}$$

which together with Birkhoff's ergodic theorem (Theorem 7.2.1 in[171]) implies

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left[\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})^\top-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})^\top\right]\\
&\quad\xrightarrow{p}\boldsymbol{S}=\mathbb{E}_{\xi\sim\pi}\left[\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top\right].
\end{aligned}$$

Because $\int_\Xi \mathscr{P}(\xi,\xi')\pi(d\xi)=\pi(\xi')$ by the definition of the stationary distribution π, we have

$$\mathbb{E}_{\xi\sim\pi}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top=\mathbb{E}_{\xi\sim\pi}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top.$$

Hence, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{u}_{t,2}\boldsymbol{u}_{t,2}^\top|\mathscr{F}_{t-1}]\xrightarrow{p}\boldsymbol{S}=\mathbb{E}_{\xi\sim\pi}\left[\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)^\top\right].$$

Hereto, we have shown $\{\boldsymbol{u}_{t,2}\}_{t\geq 0}$ satisfies the Lindeberg-Feller conditions for martingale central limit theorem. Then the martingale FCLT follows from Theorem 4.2 in[167] (or Theorem 8.8.8 in[171], or Theorem 2.1 in[154]). Therefore, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_{t,2}\xrightarrow{w}\boldsymbol{S}^{1/2}\boldsymbol{W}(r)\text{ and }\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_{t}\xrightarrow{w}\boldsymbol{S}^{1/2}\boldsymbol{W}(r).$$

Finally, by 4 in Lemma 3.2.2 and conditional Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{u}_{t,1}\|^p&\leq 2^{p-1}\left[\mathbb{E}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_t)-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\|^p+2\mathbb{E}\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t,\xi_{t-1})-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^p\right]\\
&\leq 2^p\mathbb{E}\|\boldsymbol{U}(\boldsymbol{x}_t,\xi_t)-\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\|^p\leq 2^pL_U^p\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^p.
\end{aligned}$$

As a result, we have $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{u}_{t,1}\|^p\lesssim\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^p<\infty$ from Assumption 3.2.6. Therefore, $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{u}_t\|^p\leq 2^{p-1}(\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{u}_{t,1}\|^p+\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{u}_{t,2}\|^p)<\infty$. By now, we complete the proof of this part.

3. By (3.18) and 4 in Lemma 3.2.2, we have

$$\begin{aligned}
\|\boldsymbol{v}_t\|&=\left\|\frac{\eta_{t+1}}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1},\xi_t)-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t,\xi_t)\right\|\\
&\leq\left\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1},\xi_t)-\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t,\xi_t)\right\|+\left\|\frac{\eta_{t+1}-\eta_t}{\eta_t}\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_{t+1},\xi_t)\right\|\\
&\leq L_U\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|+\left|\frac{\eta_{t+1}-\eta_t}{\eta_t}\right|\cdot C_{U,\boldsymbol{x}_{t+1}} \qquad\text{(B.2)}\\
&\lesssim L_U\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|+o(\eta_t)\cdot\left(\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|+\sigma\right).
\end{aligned}$$

From another hand, it follows that

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\| \le \eta_t\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)\| &\overset{(a)}{\le} \eta_t\left[\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\| + L_H\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|\right]\\ &\le \eta_t\left[L_H\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\| + \sup_{t\ge 0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\|\right] \overset{(b)}{\lesssim} \eta_t,\end{aligned}$$

where (a) uses the following result (which mainly follows from Assumption 3.2.3),

$$\begin{aligned}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)-\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\| &= \mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|\\ &\le \mathbb{E}(\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|^p)^{\frac{1}{p}}\\ &\le L_H\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|,\end{aligned}$$

and (b) uses the following two inequalities,

$$\begin{aligned}\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\| \le \sup_{t\ge 0}(\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\|^p)^{1/p} \lesssim 1,\\ \mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\| \le \sup_{t\ge 0}\sqrt[p]{\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^p} \lesssim 1.\end{aligned}$$

Finally, we have

$$\mathbb{E}\|\boldsymbol{v}_t\| \lesssim \eta_t \implies \frac{1}{\sqrt{T}}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{v}_t\| \lesssim \frac{1}{\sqrt{T}}\sum_{t=0}^{T}\eta_t \to 0 \text{ as } T\to\infty.$$

□

B.2 Proof of Lemma 3.4.2

Proof of Lemma 3.4.2. We analyze the four separate terms $\sup_{r\in[0,1]}\|\boldsymbol{\psi}_k(r)\|(0\le k\le 3)$ respectively.

For the partial-sum process of noises By 2 in Lemma 3.4.1, it follows that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}\boldsymbol{u}_t \overset{w}{\to} \boldsymbol{S}^{1/2}\boldsymbol{W}(r).$$

For $\boldsymbol{\psi}_0$ 2 in Lemma B.7.1 shows $\boldsymbol{A}_j^n$ is uniformly bounded. As $T\to\infty$,

$$\sup_{r\in[0,1]}\|\boldsymbol{\psi}_0(r)\| = \frac{1}{\sqrt{T}\eta_0}\sup_{r\in[0,1]}\|\boldsymbol{A}_0^{\lfloor Tr\rfloor}\boldsymbol{B}_0\boldsymbol{\Delta}_0\| \le \frac{C_0}{\sqrt{T}\eta_0}\|\boldsymbol{B}_0\boldsymbol{\Delta}_0\| \to 0.$$

For $\boldsymbol{\psi}_1$ Recall that $\boldsymbol{\psi}_1(r) = \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{A}_t^{\lfloor Tr \rfloor}(\boldsymbol{r}_t + \boldsymbol{v}_t)$. Since $\|\boldsymbol{A}_j^n\| \le C_0$ for any $n \ge j \ge 0$, it follows that as $T \to \infty$,

$$\mathbb{E} \sup_{r \in [0,1]} \|\boldsymbol{\psi}_1(r)\| \le \frac{C_0}{\sqrt{T}} \mathbb{E} \sum_{t=0}^{T} \left(\|\boldsymbol{r}_t\| + \|\boldsymbol{v}_t\|\right) \to 0,$$

where the last inequality uses 1 and 3 in Lemma 3.4.1.

For $\boldsymbol{\psi}_2$ Recall that $\boldsymbol{\psi}_2(r) = \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left(\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\right) \boldsymbol{u}_t$ with $\boldsymbol{u}_t$ a martingale difference. In the following, we set $\boldsymbol{z}_t = \boldsymbol{\psi}_2(t/T)$ (indexed by $t \in [T]$) for simplicity. It is clear that $\{\boldsymbol{z}_t, \mathscr{F}_t\}_{t \in [T]}$ forms a square integrable martingale difference sequence. As a result $\{\|\boldsymbol{z}_t\|_2, \mathscr{F}_t\}_{t \in [T]}$ is a submartingale due to $\mathbb{E}[\|\boldsymbol{u}_t\|_2 | \mathscr{F}_{t-1}] \ge \|\mathbb{E}[\boldsymbol{u}_t | \mathscr{F}_{t-1}]\|_2 = \|\boldsymbol{u}_{t-1}\|_2$ from conditional Jensen's inequality. By Doob' s maximum inequality for submartingales (which we use to derive the following $(*)$ inequality),

$$\begin{aligned}
\mathbb{E} \sup_{r \in [0,1]} \|\boldsymbol{\psi}_2(r)\|_2^2 = \mathbb{E} \sup_{t \in [T]} \|\boldsymbol{z}_t\|_2^2 &\overset{(*)}{\le} 4\mathbb{E}\|\boldsymbol{z}_T\|_2^2 \\
&= \frac{4}{T} \sum_{t=0}^{T} \mathbb{E}\| \left(\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\right) \boldsymbol{u}_t\|_2^2 \\
&\le 4 \sup_{t \ge 0} \mathbb{E}\|\boldsymbol{u}_t\|_2^2 \cdot \frac{1}{T} \sum_{t=0}^{T} \|\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\|_2^2 \to 0,
\end{aligned}$$

where we use 2 in Lemma B.7.1 and the fact $\|\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\|$ is uniformly bounded by $C_0 + \|\boldsymbol{G}^{-1}\|$. Due to the norm equivalence in $\mathbb{R}^d$, $\|\cdot\|$ is equivalent to $\|\cdot\|_2$ up to universal constants.

For $\boldsymbol{\psi}_3$ Recall that $\boldsymbol{\psi}_3(r) = \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \left(\boldsymbol{A}_t^{\lfloor Tr \rfloor} - \boldsymbol{A}_t^T\right) \boldsymbol{u}_t$ with $\boldsymbol{u}_t$ a martingale difference. Notice that for any $n \in [T]$

$$\begin{aligned}
\sum_{t=0}^{n} (\boldsymbol{A}_t^T - \boldsymbol{A}_t^n)\boldsymbol{u}_t &= \sum_{t=0}^{n} \sum_{j=n+1}^{T} \left(\prod_{i=t+1}^{j} \boldsymbol{B}_i\right) \eta_t \boldsymbol{u}_t = \sum_{j=n+1}^{T} \sum_{t=0}^{n} \left(\prod_{i=t+1}^{j} \boldsymbol{B}_i\right) \eta_t \boldsymbol{u}_t \\
&= \sum_{j=n+1}^{T} \left(\prod_{i=n+1}^{j} \boldsymbol{B}_i\right) \sum_{t=1}^{n} \left(\prod_{i=t+1}^{n} \boldsymbol{B}_i\right) \eta_t \boldsymbol{u}_t \qquad \text{(B.3)} \\
&= \frac{1}{\eta_{n+1}} \boldsymbol{A}_{n+1}^T \boldsymbol{B}_{n+1} \sum_{t=0}^{n} \left(\prod_{i=t+1}^{n} \boldsymbol{B}_i\right) \eta_t \boldsymbol{u}_t.
\end{aligned}$$

From 2 in Lemma B.7.1, $\|\boldsymbol{A}_{n+1}^T \boldsymbol{B}_{n+1}\| \le C_0(1 + \|\boldsymbol{G}\|)$ for any $T \ge n \ge 0$. Hence,

$$\sup_{r\in[0,1]} \|\boldsymbol{\psi}_3(r)\| = \sup_{n\in[T]} \left\| \frac{1}{\sqrt{T}} \sum_{t=0}^{n} \left(\boldsymbol{A}_t^n - \boldsymbol{A}_t^T \right) \boldsymbol{u}_t \right\| \tag{3.22}$$

$$\lesssim \sup_{n\in[T]} \left\| \frac{1}{\sqrt{T}} \frac{1}{\eta_{n+1}} \sum_{t=0}^{n} \left(\prod_{i=t+1}^{n} \boldsymbol{B}_i \right) \eta_t \boldsymbol{u}_t \right\| = o_{\mathbb{P}}(1), \tag{B.4}$$

where the last inequality uses Lemma 3.4.3. We then complete the proof. □

B.3 Proof of Lemma 3.4.3

For the proof in the part, we will consider random variables (or matrices) in the complex field $\mathbb{C}$. Hence, we will introduce new notations for them. For a vector $\boldsymbol{v} \in \mathbb{C}$ (or a matrix $\boldsymbol{U} \in \mathbb{C}^{d\times d}$), we use $\boldsymbol{v}^{\mathrm{H}}$ (or $\boldsymbol{U}^{\mathrm{H}}$) to denote its Hermitian transpose or conjugate transpose. For any two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{C}$, with a slight abuse of notation, we use $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \boldsymbol{v}^{\mathrm{H}}\boldsymbol{u}$ to denote the inner product in $\mathbb{C}$. For simplicity, for a complex matrix $\boldsymbol{U} \in \mathbb{C}^{d\times d}$, we use $\|\boldsymbol{U}\|$ to denote the its operator norm introduced by the complex inner product $\langle\cdot,\cdot\rangle$. When $\boldsymbol{U} \in \mathbb{R}^{d\times d}$, $\|\boldsymbol{U}\|$ is reduced to the spectrum norm.

Proof of Lemma 3.4.3. We provide the proof only for the asymptotic result; for the weak convergence rate see Lemma D.2.2 and its proof. To simplify notation, we say a random sequence $\{\boldsymbol{y}_t\}_{t\ge0}$ is *uniformly ignorable* if $\frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \overset{p}{\to} 0$ when $T \to \infty$. Our target is equivalent to show the defined $\{\boldsymbol{y}_t\}_{t\ge0}$ is uniformly ignorable.

We are going to prove the lemma in two steps. In the first step, we prove a weaker version in Lemma B.3.1 under an additional assumption that requires $\boldsymbol{G}$ is diagonalizable. The proof of Lemma B.3.1 is deferred in Section B.4. Then, in the second step, we remove the added assumption via a refined analysis that relies on induction to reduce the general Hurwitz case to the established diagonalizable case by using the Jordan decomposition of $\boldsymbol{G}$.

Lemma B.3.1. *Under the same condition of Lemma 3.4.3, if we additionally assume $\boldsymbol{G}$ is diagonalizable, then*

$$\frac{1}{\sqrt{T}} \sup_{0\le t\le T} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \overset{p}{\to} 0.$$

Lemma B.3.2 serves a bridge to connect the general Hurwitz case with the diagonalizable case. Its proof is provided in Section B.6.

Lemma B.3.2. *Let $\{\eta_t\}_{t\geq 0}$ be the step size satisfying Assumption 3.2.5 and $\lambda \in \mathbb{C}$ be a complex number with positive real part* $\mathrm{Re}(\lambda) > 0$*. Let $\{\omega_t\}_{t\geq 0} \subseteq \mathbb{C}$ be a sequence of random variables taking value in the complex field and $\frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{|\omega_t|}{\eta_t} \xrightarrow{p} 0$ as $T \to \infty$. Consider the sequence $\{z_t\}_{t\geq 0}$ defined recursively as following: $z_0 = 0$ and*

$$z_{t+1} = z_t - \lambda\eta_t z_t + \eta_t\omega_t.$$

Then when $T \to \infty$, we have $\{z_t\}_{t\geq 0}$ is also uniformly ignorable, i.e.,

$$\frac{1}{\sqrt{T}} \sup_{0\leq t\leq T} \frac{|z_{t+1}|}{\eta_{t+1}} \xrightarrow{p} 0.$$

By viewing $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ as a complex matrix, it has the Jordan decomposition with the Jordan canonical form denoted by $\boldsymbol{G} = \boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^{-1} = \boldsymbol{V}\mathrm{diag}\{\boldsymbol{J}_1, \cdots, \boldsymbol{J}_r\}\boldsymbol{V}^{-1}$, where $\boldsymbol{V}$ is the non-singular matrix and $\{\boldsymbol{J}_i\}_{1\leq i\leq r}$ collects all Jordan blocks. Recall that $\{\boldsymbol{y}_t\}_{t\geq 0}$ is defined in (3.24). Let $\tilde{\boldsymbol{y}}_t = \boldsymbol{V}^{-1}\boldsymbol{y}_t$, $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{V}^{-1}\boldsymbol{\varepsilon}_t$ be transformed vectors. Then the recursion formula (3.24) becomes

$$\tilde{\boldsymbol{y}}_{t+1} = (\boldsymbol{I} - \eta_t\boldsymbol{J})\tilde{\boldsymbol{y}}_t + \eta_t\tilde{\boldsymbol{\varepsilon}}_t.$$

Without loss of generality, we assume that $\boldsymbol{J}$ consists of only one Jordan block, i.e.

$$\boldsymbol{J} = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \tag{B.5}$$

with $\lambda \in \mathbb{C}$ and $\mathrm{Re}(\lambda) > 0$.

Let $\left(\tilde{\boldsymbol{y}}_t\right)_k$ denote the k-th coordinate of the vector $\tilde{\boldsymbol{y}}_t$ and so does $\left(\tilde{\boldsymbol{\varepsilon}}_t\right)_k$. Then, in order to prove that $\{\tilde{\boldsymbol{y}}_t\}_{t\geq 0}$ is uniformly ignorable, we only needs to prove that each of its coordinates $\{\left(\tilde{\boldsymbol{y}}_t\right)_k\}_{t\geq 0}(1 \leq k \leq d)$ is uniformly ignorable. Notice that the last coordinate process evolves as $(\tilde{\boldsymbol{y}}_{t+1})_d = (1-\eta_t\lambda)(\tilde{\boldsymbol{y}}_t)_d + \eta_t(\tilde{\boldsymbol{\varepsilon}}_t)_d$. Lemma B.3.1 implies that $\{(\tilde{\boldsymbol{y}}_t)_d\}_{t\geq 1}$, as a one-dimensional process, is uniformly ignorable. We are going to finish the proof by induction. Suppose for the coordinates $k, k+1, \cdots, d$, we already have $\{(\tilde{\boldsymbol{y}}_t)_k\}_{t\geq 0}$ is uniformly ignorable, now we are going to prove $\{(\tilde{\boldsymbol{y}}_t)_{k-1}\}_{t\geq 0}$ is also uniformly ignorable.

Using the structure of $\boldsymbol{J}$ in (B.5), we have

$$(\tilde{\boldsymbol{y}}_{t+1})_{k-1} = (1 - \lambda\eta_t)(\tilde{\boldsymbol{y}}_t)_{k-1} - \eta_t(\tilde{\boldsymbol{y}}_t)_k + \eta_t(\tilde{\boldsymbol{\varepsilon}}_t)_{k-1}. \tag{B.6}$$

To facilitate analysis, we construct a surrogate sequence $\{(\hat{\boldsymbol{y}}_t)_{k-1}\}$ defined by

$$(\hat{\boldsymbol{y}}_{t+1})_{k-1} = (1 - \lambda\eta_t)(\hat{\boldsymbol{y}}_t)_{k-1} + \eta_t(\tilde{\boldsymbol{\varepsilon}}_t)_{k-1}. \tag{B.7}$$

Again, by Lemma B.3.1, $\{(\hat{\boldsymbol{y}}_t)_{k-1}\}_{t\geq 0}$ is uniformly ignorable. Let $\tilde{\boldsymbol{\Delta}}_t := (\tilde{\boldsymbol{y}}_t)_{k-1} - (\hat{\boldsymbol{y}}_t)_{k-1}$ be their difference. From (B.6) − (B.7), it follows that

$$\tilde{\boldsymbol{\Delta}}_{t+1} = (1 - \lambda\eta_t)\tilde{\boldsymbol{\Delta}}_t - \eta_t(\tilde{\boldsymbol{y}}_t)_k.$$

Thanks to Lemma B.3.2 and our hypothesis, $\{\tilde{\boldsymbol{\Delta}}_t\}_{t\geq 0}$ is uniformly ignorable. Finally, putting the pieces together, we have

$$\begin{aligned}\frac{1}{\sqrt{T}}\sup_{t\in[0,T]}\frac{\left|(\tilde{\boldsymbol{y}}_{t+1})_{k-1}\right|}{\eta_{t+1}} &= \frac{1}{\sqrt{T}}\sup_{t\in[0,T]}\frac{\left|(\hat{\boldsymbol{y}}_{t+1})_{k-1} + \tilde{\boldsymbol{\Delta}}_{t+1}\right|}{\eta_{t+1}} \\ &\leq \frac{1}{\sqrt{T}}\sup_{t\in[0,T]}\frac{\left|(\hat{\boldsymbol{y}}_{t+1})_{k-1}\right|}{\eta_{t+1}} + \frac{1}{\sqrt{T}}\sup_{t\in[0,T]}\frac{\left|\tilde{\boldsymbol{\Delta}}_{t+1}\right|}{\eta_{t+1}} \xrightarrow{p} 0.\end{aligned}$$

□

B.4 Proof of Lemma B.3.1

Proof of Lemma B.3.1. The proof is divided into three steps.

Step one: Divide the time interval Given a positive integer n, we separate the time interval $[0, T]$ uniformly into n portions with $h_k = \left[\frac{k}{n}(T+1)\right]$ $(k = 0, 1, \ldots, n)$ the k-th endpoint. The choice of n is independent of T, which implies that $\lim_{T\to\infty} h_k = \infty$ for any $k \geq 1$. Let $c_0' := c_0\exp(c\eta_0)$ with the constants c, c_0 defined in 1 in Lemma B.7.1. For any $\varepsilon > 0$, we define an event $\mathscr{A}$ whose complement is

$$\mathscr{A}^c := \left\{\exists 0 \leq k \leq n \text{ s.t. } \frac{c_0'}{\sqrt{T}}\left\|\frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}}\right\| \geq \varepsilon\right\}. \tag{B.8}$$

We claim that $\limsup\limits_{T\to\infty}\mathbb{P}(\mathscr{A}^c) = 0$. For one thing,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{y}_{t+1}\|^2 &= \mathbb{E}\left\|\sum_{j=0}^{t}\left(\prod_{i=j+1}^{t}\left(\boldsymbol{I} - \eta_i\boldsymbol{G}\right)\right)\eta_j\boldsymbol{\varepsilon}_j\right\|^2 = \sum_{j=0}^{t}\mathbb{E}\left\|\boldsymbol{X}_{j+1}^{t}\eta_j\boldsymbol{\varepsilon}_j\right\|^2 \\ &\leq \sup_{j\geq 0}\mathbb{E}\|\boldsymbol{\varepsilon}_j\|^2\cdot\sum_{j=0}^{t}\eta_j^2\exp\left(-c\sum_{t=j+1}^{t}\eta_i\right)\overset{(a)}{\leq}\sup_{j\geq 0}\mathbb{E}\|\boldsymbol{\varepsilon}_j\|^2\cdot c_1\eta_t\overset{(b)}{\leq} c_2\eta_t,\end{aligned} \tag{B.9}$$

where (a) follows from 4 in Lemma B.7.1 and (b) follows by setting $c_2 = c_1 \cdot \sup_{j\ge 0} \mathbb{E}\|\boldsymbol{\varepsilon}_j\|^2$. For another thing, by the union bound and Markov's inequality,

$$\mathbb{P}(\mathscr{A}^c) \le \sum_{k=0}^{n} \mathbb{P}\left(\frac{c_0'}{\sqrt{T}}\left\|\frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}}\right\| \ge \varepsilon\right) \le \frac{(c_0')^2}{T\varepsilon^2}\sum_{k=0}^{n}\mathbb{E}\left\|\frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}}\right\|^2 \overset{(a)}{\le} \frac{c_2(c_0')^2}{T\varepsilon^2}\sum_{k=0}^{n}\frac{1}{\eta_{h_k}} \overset{(b)}{\le} \frac{c_2(c_0')^2(n+1)}{\eta_T T\varepsilon^2}. \tag{B.10}$$

Here (a) uses the inequality (B.9) and (b) is because $t\eta_t \to \infty$. So, when $T \to \infty$, $\mathbb{P}(\mathscr{A}^c) \to 0$ due to $T\eta_T \to \infty$.

Using the notation in Lemma B.7.1, we denote $\boldsymbol{X}_j^n := \prod_{i=j}^{n}\left(\boldsymbol{I} - \eta_i \boldsymbol{G}\right)$. Clearly $\boldsymbol{X}_j^n$'s are exchangeable since they are polynomials of the same matrix $\boldsymbol{G}$. Hence, $\boldsymbol{X}_{j+1}^t = (\boldsymbol{X}_t^T)^{-1}\boldsymbol{X}_{j+1}^T$. From (3.24), if $t \in [h_k, h_{k+1})$ for some $k \in [n]$, we then have

$$\begin{aligned}
\boldsymbol{y}_{t+1} &= \sum_{j=0}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j = (\boldsymbol{X}_{t+1}^T)^{-1} \sum_{j=0}^{t} \boldsymbol{X}_{j+1}^T \eta_j \boldsymbol{\varepsilon}_j \\
&= (\boldsymbol{X}_{t+1}^T)^{-1}\left[\sum_{j=0}^{h_k-1} \boldsymbol{X}_{j+1}^T \eta_j \boldsymbol{\varepsilon}_j + \sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^T \eta_j \boldsymbol{\varepsilon}_j\right] \\
&= \boldsymbol{X}_{h_k}^t \boldsymbol{y}_{h_k} + \sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j.
\end{aligned}$$

When T is sufficiently large ($T \ge nj_0$ is sufficient with j_0 defined in 3 of Lemma B.7.1), we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{0\le t\le T}\frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \ge 2\varepsilon; \mathscr{A}\right) \\
&\le \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{t\in[h_k,h_{k+1})}\frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \ge 2\varepsilon; \mathscr{A}\right) \\
&\le \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left[\left\|\boldsymbol{X}_{h_k}^t \boldsymbol{y}_{h_k}\right\| + \left\|\sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j\right\|\right] \ge 2\varepsilon; \mathscr{A}\right) \\
&\le \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{t\in[h_k,h_{k+1})}\frac{\eta_{h_k}}{\eta_{t+1}}\left\|\boldsymbol{X}_{h_k}^t\right\|\left\|\frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}}\right\| \ge \varepsilon; \mathscr{A}\right) \\
&\qquad + \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}}\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left\|\sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j\right\| \ge \varepsilon; \mathscr{A}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\le} \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}} \left\| \sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j \right\| \ge \varepsilon; \mathscr{A} \right) \\
&\le \sum_{k=0}^{n-1} \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}} \left\| \sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j \right\| \ge \varepsilon \right) := \sum_{k=0}^{n-1} \mathscr{P}_k. \qquad \text{(B.11)}
\end{aligned}
$$

Here (a) uses the following result. When $k \ge 1$, due to $h_k \ge \lceil \frac{k}{n}(T+1) \rceil$, h_k could be arbitrarily large with increasing T and fixed n. From 1 and 3 of Lemma B.7.1, when $h_k \ge j_0$,

$$
\frac{\eta_{h_k}}{\eta_{t+1}} \left\| \boldsymbol{X}_{h_k}^t \right\| \le \exp\left(\frac{c}{2} \sum_{t=h_k}^{t+1} \eta_t \right) \cdot c_0 \exp\left(-c \sum_{t=h_k}^{t} \eta_t \right) \le c_0 \exp(c\eta_0) = c_0',
$$

which implies $\sup_{t\in[h_k,h_{k+1})} \frac{\eta_{h_k}}{\eta_{t+1}} \left\| \boldsymbol{X}_{h_k}^t \right\| \le c_0'$ for any $k \ge 1$. Notice $\frac{c_0'}{\sqrt{T}} \left\| \frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}} \right\| \ge \varepsilon$ is impossible on the event $\mathscr{A}$. We then have

$$
\mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{t\in[h_k,h_{k+1})} \frac{\eta_{h_k}}{\eta_{t+1}} \left\| \boldsymbol{X}_{h_k}^t \right\| \left\| \frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}} \right\| \ge \varepsilon; \mathscr{A} \right) \le \mathbb{P}\left(\frac{c_0'}{\sqrt{T}} \left\| \frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}} \right\| \ge \varepsilon; \mathscr{A} \right) = 0.
$$

When $k = 0$, the above probability is clearly zero since $\boldsymbol{y}_0 = \boldsymbol{0}$.

Step two: Bound each $\mathscr{P}_k$ The proof of Lemma B.4.1 can be found in Section B.5.

Lemma B.4.1. *Assume $T \ge n$. For each $k \in [n]$,*

$$
\mathscr{P}_k := \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}} \left\| \sum_{j=h_k}^{t} \boldsymbol{X}_{j+1}^t \eta_j \boldsymbol{\varepsilon}_j \right\| \ge \sqrt{T}\varepsilon \right) \le p^p C_3^p \cdot n^{-\frac{p}{2}} \varepsilon^{-p}
$$

where C_3 is a positive constant depending on the step sizes, $\boldsymbol{G}, d$ and $\sup_{t\ge0} \sqrt[p]{\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^p}$. In short, C_3 has nothing to do with p.

Step three: Put pieces together Therefore,

$$
\begin{aligned}
\mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{0\le t\le T} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \ge 2\varepsilon \right) &\le \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{0\le t\le T} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \ge 2\varepsilon; \mathscr{A} \right) + \mathbb{P}(\mathscr{A}^c) \\
&\overset{(a)}{\le} \sum_{k=0}^{n-1} \mathscr{P}_k + \frac{c_2 (c_0')^2 (n+1)}{\eta_T T \varepsilon^2} \overset{(b)}{\lesssim} p^p C_3^p \varepsilon^{-p} n^{-\frac{p}{2}+1} + \frac{n}{\eta_T T \varepsilon^2},
\end{aligned}
$$

where (a) uses (B.11) and (B.10) and (b) uses Lemma B.4.1. As a result, for any $\varepsilon > 0$,

$$
\limsup_{T\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{T}} \sup_{0\le t\le T} \frac{\|\boldsymbol{y}_{t+1}\|}{\eta_{t+1}} \ge 2\varepsilon \right) \lesssim p^p C_3^p \varepsilon^{-p} n^{-\frac{p}{2}+1}.
$$

Since $p > 2$ and the probability of the left-hand side has nothing to do with n, letting $n \to \infty$ concludes the proof. □

B.5 Proof of Lemma B.4.1

Proof of Lemma B.4.1. Readers should keep in mind that we only have $p > 2$ in this part. Without loss of generality, we fix $k \in [n]$.

Step one: Diagonalization Since $\boldsymbol{G}$ is diagonalizable, there exist two non-singular matrices $\boldsymbol{U}, \boldsymbol{D} \in \mathbb{C}^{d\times d}$ that satisfy $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ and $\boldsymbol{D}$ is a diagonal matrix with each entry the eigenvalue of $\boldsymbol{G}$.[1] Further, $\boldsymbol{D} = \operatorname{diag}(\{\lambda_i(\boldsymbol{G})\}_{i\in[d]})$ and $\operatorname{Re}\lambda_i(\boldsymbol{G}) > 0$. Therefore, denote $\tilde{\boldsymbol{X}}_j^t := \prod_{i=j}^{t}\left(\boldsymbol{I} - \eta_i\boldsymbol{D}\right)$ and thus we have

$$\boldsymbol{U}^{-1}\sum_{j=h_k}^{t}\boldsymbol{X}_{j+1}^t\eta_j\boldsymbol{\varepsilon}_j = \sum_{j=h_k}^{t}\left(\prod_{i=j+1}^{t}\left(\boldsymbol{I} - \eta_i\boldsymbol{D}\right)\right)\eta_j\boldsymbol{U}^{-1}\boldsymbol{\varepsilon}_j = \sum_{j=h_k}^{t}\tilde{\boldsymbol{X}}_{j+1}^t\eta_j\boldsymbol{U}^{-1}\boldsymbol{\varepsilon}_j.$$

Hence,

$$\begin{aligned}
\mathscr{P}_k &= \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left\|\sum_{j=h_k}^{t}\boldsymbol{X}_{j+1}^t\eta_j\boldsymbol{\varepsilon}_j\right\| \ge \sqrt{T}\varepsilon\right)\\
&\le \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left\|\sum_{j=h_k}^{t}\tilde{\boldsymbol{X}}_{j+1}^t\eta_j\boldsymbol{U}^{-1}\boldsymbol{\varepsilon}_j\right\| \ge \frac{\sqrt{T}\varepsilon}{\|\boldsymbol{U}\|}\right)\\
&\overset{(a)}{\le} \sum_{i=1}^{d}\mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left|\left(\sum_{j=h_k}^{t}\tilde{\boldsymbol{X}}_{j+1}^t\eta_j\boldsymbol{U}^{-1}\boldsymbol{\varepsilon}_j\right)_i\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)\\
&\overset{(b)}{=} \sum_{i=1}^{d}\mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})}\frac{1}{\eta_{t+1}}\left|\sum_{j=h_k}^{t}(\tilde{\boldsymbol{X}}_{j+1}^t)_{i,i}\eta_j\left(\boldsymbol{U}^{-1}\boldsymbol{\varepsilon}_j\right)_i\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right) := \sum_{i=1}^{d}\mathscr{P}_{k,i},
\end{aligned}$$

where (a) uses the notation $(\boldsymbol{v})_i$ denotes the i-th coordinate of the vector $\boldsymbol{v}$ and $|\cdot|$ denotes the norm for complex numbers and (b) uses the fact that $\boldsymbol{G}$ is a diagonal matrix. The above analysis shows we only need to focus on each coordinate thanks to diagonalization.

Step two: Establish tail probability bound for each coordinate Without loss of generality, we fix any coordinate $i \in [d]$. Let $\lambda := \lambda_i(\boldsymbol{G})$ denotes the i-th eigenvalue for short (only in

[1] In this proof, with a slight abuse of notation, we use $\boldsymbol{U}$ to denote a non-singular complex matrix. Readers should distinguish it from the bivariate function $\boldsymbol{U}(\boldsymbol{x}, \xi)$ defined in Lemma 3.2.2.

this part). With a little abuse of notation, we set $X_{j+1}^t := (\tilde{\boldsymbol{X}}_{j+1}^t)_{i,i}$ and $\varepsilon_j = (\boldsymbol{U}^{-1}\varepsilon_j)_i$, both complex numbers and $X_{j+1}^t = \prod\limits_{i=j+1}^{t} \left(\boldsymbol{I} - \eta_i\lambda\right)$. Hence, $\mathscr{P}_k \le \sum_{i=1}^d \mathscr{P}_{k,i}$ where

$$\begin{aligned}
\mathscr{P}_{k,i} &= \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}}\left|\sum_{j=h_k}^{t} X_{j+1}^t\eta_j\varepsilon_j\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)\\
&= \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}}\left|(X_{t+1}^T)^{-1}\sum_{j=h_k}^{t} X_{j+1}^T\eta_j\varepsilon_j\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)\\
&\overset{(a)}{=} \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\eta_{t+1}}\left|(X_{t+1}^T)^{-1}\right|\left|\sum_{j=h_k}^{t} X_{j+1}^T\eta_j\varepsilon_j\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)\\
&\overset{(b)}{=} \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\left|\eta_{t+1}X_{t+1}^T\right|}\left|\sum_{j=h_k}^{t} X_{j+1}^T\eta_j\varepsilon_j\right| \ge \frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)\\
&= \mathbb{P}\left(\sup_{t\in[h_k,h_{k+1})} \frac{1}{\left|\eta_{t+1}X_{t+1}^T\right|^p}\left|\sum_{j=h_k}^{t} X_{j+1}^T\eta_j\varepsilon_j\right|^p \ge \left(\frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)^p\right),
\end{aligned} \tag{B.12}$$

where (a) follows from $|ab| = |a|\cdot|b|$ for any $a, b \in \mathbb{C}$; and (b) follows from $|a^{-1}|\cdot|a| = 1$ for any $a \ne 0 \in \mathbb{C}$.

Lemma B.5.1 (Chow's inequality[170]). *Let $\{Y_t\}_{t\ge0} \subseteq \mathbb{R}$ be a sub-martingale and $\{b_t\}_{t\ge0}$ be a non-increasing sequence. Denote $Y_t^+ = \max(0, Y_t)$. Then for any $\varepsilon > 0$, we have*

$$\varepsilon\cdot\mathbb{P}\left(\sup_{0\le t\le T} b_tY_t \ge \varepsilon\right) \le \sum_{t=0}^{T-1}(b_t - b_{t+1})\mathbb{E}Y_t^+ + b_T\mathbb{E}Y_T^+.$$

Lemma B.5.2 (Burkholder's inequalities[172]). *Fix any $p \ge 2$. For $\mathbb{C}$-valued martingale difference $X_1, \cdots, X_T$, each with finite L^p-norm, one has*

$$\mathbb{E}\left|\sum_{t=1}^{T} X_t\right|^p \le B_p^p\mathbb{E}\left(\sum_{t=1}^{T}|X_t|^2\right)^{\frac{p}{2}}$$

where $B_p = \max\left\{p-1, \frac{1}{p-1}\right\}$ is a universal constant depending only on p. It together with Jensen's inequality implies

$$\mathbb{E}\left|\sum_{t=1}^{T} X_t\right|^p \le B_p^pT^{\frac{p}{2}-1}\sum_{t=1}^{T}\mathbb{E}|X_t|^p. \tag{B.13}$$

Based on (B.12), we will use Chow's inequality to bound each $\mathscr{P}_{k,i}$'s. We first check (B.12) satisfies the conditions in Lemma B.5.1. First, $\eta_{t+1}\left|X_{t+1}^T\right|$ is non-decreasing for when t is sufficiently large. This is because

$$\eta_t\left|X_t^T\right| = \frac{\eta_t}{\eta_{t+1}}\left|1-\eta_t\lambda\right|\cdot\eta_{t+1}\left|X_{t+1}^T\right| \le \eta_{t+1}\left|X_{t+1}^T\right|, \tag{B.14}$$

for which we use

$$\frac{\eta_t}{\eta_{t+1}}\left|1-\eta_t\lambda\right| = (1+o(\eta_t))\sqrt{(1-\eta_t\mathrm{Re}\lambda)^2+\eta_t^2(\mathrm{Im}\lambda)^2} = (1+o(\eta_t))\sqrt{1-2\eta_t\mathrm{Re}\lambda+O(\eta_t^2)} \le 1,$$

when η_t is sufficiently small, or equivalently, t is sufficiently large, say larger than t_0'. Hence, $b_t := \left|\eta_{t+1}X_{t+1}^T\right|^{-p}$ is non-increasing. Second, let $Y_t := \left|\sum_{j=h_k}^{t} X_{j+1}^T\eta_j\varepsilon_j\right|^p$. It is easy to check Y_t is a sub-martingale satisfying $\mathbb{E}[Y_t|\mathscr{F}_{t-1}] \ge Y_{t-1}$. What's more, (B.13) implies $\mathbb{E}Y_t$ is bounded by

$$\begin{aligned}
\mathbb{E}Y_t &\le B_p^p(t-h_k+1)^{\frac{p}{2}-1}\sum_{j=h_k}^{t}\mathbb{E}\left|X_{j+1}^T\eta_j\varepsilon_j\right|^p \\
&\le B_p^p\left(\frac{T}{n}\right)^{\frac{p}{2}-1}\sum_{j=h_k}^{t}\left|X_{j+1}^T\eta_j\right|^p\mathbb{E}\left|\varepsilon_j\right|^p \\
&\le p^pC_3^p\cdot\left(\frac{T}{n}\right)^{\frac{p}{2}-1}\sum_{j=h_k}^{t}\left|X_{j+1}^T\eta_{j+1}\right|^p = C_3^p\cdot\left(\frac{T}{n}\right)^{\frac{p}{2}-1}\sum_{j=h_k}^{t}\frac{1}{b_j},
\end{aligned} \tag{B.15}$$

where $C_3 := 2\cdot\sup_{t\ge0,i\in[d]}\sqrt[p]{\mathbb{E}|\varepsilon_t|^p}$ is a constant depending only on $\boldsymbol{U}$ and $\sup_{t\ge0}\sqrt[p]{\mathbb{E}\|\varepsilon_t\|^p}$.

Hence, by Lemma B.5.1 and (B.15), it follows that

$$\begin{aligned}
\mathscr{P}_{k,i} &\le \left(\frac{\varepsilon}{\|\boldsymbol{U}\|}\sqrt{\frac{T}{d}}\right)^{-p}\cdot p^pC_3^p\left(\frac{T}{n}\right)^{\frac{p}{2}-1}\left[\sum_{t=h_k}^{h_{k+1}-2}\left(b_t-b_{t+1}\right)\sum_{j=h_k}^{t}\frac{1}{b_j}+b_{h_{k+1}-1}\sum_{j=h_k}^{h_{k+1}-1}\frac{1}{b_j}\right] \\
&= \left(\frac{\sqrt{d}\|\boldsymbol{U}\|}{\sqrt{T}\varepsilon}\right)^p\cdot p^pC_3^p\left(\frac{T}{n}\right)^{\frac{p}{2}-1}\left[\sum_{j=h_k}^{h_{k+1}-2}\frac{1}{b_j}\sum_{t=j}^{h_{k+1}-2}\left(b_t-b_{t+1}\right)+b_{h_{k+1}-1}\sum_{j=h_k}^{h_{k+1}-1}\frac{1}{b_j}\right] \\
&\le \left(\frac{\sqrt{d}\|\boldsymbol{U}\|}{\sqrt{T}\varepsilon}\right)^p\cdot p^pC_3^p\left(\frac{T}{n}\right)^{\frac{p}{2}} = p^pC_3^p\left(\frac{\sqrt{d}\|\boldsymbol{U}\|}{\sqrt{n}\varepsilon}\right)^p,
\end{aligned}$$

which implies that for any $k\ge1$,

$$\mathscr{P}_k \le \sum_{i\in[d]}\mathscr{P}_{k,i} \le p^pC_3^pd^{1+\frac{p}{2}}\cdot\left(\frac{\|\boldsymbol{U}\|}{\sqrt{n}\varepsilon}\right)^p.$$

For $k=0$, in order to establish (B.14), we can follow the same argument of bounding each

$\mathscr{P}_k$'s by noticing

$$
\begin{aligned}
\mathscr{P}_0 &= \mathbb{P}\left(\sup_{t\in[h_0,h_1)} \frac{1}{\eta_{t+1}} \left\| \sum_{j=h_0}^{t} \boldsymbol{X}_{j+1}^{t} \eta_j \boldsymbol{\varepsilon}_j \right\| \geq \sqrt{T}\varepsilon \right) \\
&\leq \mathbb{P}\left(\sup_{t\in[0,t_0')} \frac{1}{\eta_{t+1}} \left\| \sum_{j=0}^{t} \boldsymbol{X}_{j+1}^{t} \eta_j \boldsymbol{\varepsilon}_j \right\| \geq 0.5\sqrt{T}\varepsilon \right) \\
&\quad + \mathbb{P}\left(\left\| \boldsymbol{X}_0^{t_0'} \right\| \sup_{t\in[t_0',h_1)} \frac{1}{\eta_{t+1}} \left\| \sum_{j=t_0'}^{t} \boldsymbol{X}_{j+1}^{t} \eta_j \boldsymbol{\varepsilon}_j \right\| \geq 0.5\sqrt{T}\varepsilon \right) \\
&\leq \frac{2^p}{T^{\frac{p}{2}}\varepsilon^p} \cdot \mathbb{E}\left[\sup_{t\in[0,t_0')} \frac{1}{\eta_{t+1}} \left\| \sum_{j=0}^{t} \boldsymbol{X}_{j+1}^{t} \eta_j \boldsymbol{\varepsilon}_j \right\| \right]^p + p^p C_3^p 2^p \left\| \boldsymbol{X}_0^{t_0'} \right\|^p d^{1+\frac{p}{2}} \cdot \left(\frac{\|\boldsymbol{U}\|}{\sqrt{n}\varepsilon} \right)^p \\
&\leq p^p C_3^p n^{-\frac{p}{2}} \varepsilon^{-p},
\end{aligned}
$$

where the last inequality redefines C_3 by enlarging the original C_3 and $T \geq n$. Note that the moment quantity in the first term, $\|\boldsymbol{X}_0^{t_0'}\|, t_0', \|\boldsymbol{U}\|$ depends on $\boldsymbol{G}, \{\eta_t\}_{t\geq 0}$. C_3 is a quantity that depends on $\boldsymbol{G}, d$ and $\sup_{t\geq 0} \sqrt[p]{\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^p}$.

□

B.6 Proof of Lemma B.3.2

Proof of Lemma B.3.2. By definition, we have that

$$
z_{t+1} = \sum_{j=0}^{t} \left(\prod_{i=j+1}^{t} (1-\lambda\eta_i) \right) \eta_j \omega_j .
$$

The last equation implies that

$$
\begin{aligned}
\frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{|z_{t+1}|}{\eta_t} &= \frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{1}{\eta_t} \left| \sum_{j=0}^{t} \prod_{i=j+1}^{t} (1-\lambda\eta_j)\eta_j\omega_j \right| \\
&\leq \frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \prod_{i=j+1}^{t} \eta_j |1-\lambda\eta_j| |\omega_j| \\
&= \sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \eta_j^2 \prod_{i=j+1}^{t} |1-\lambda\eta_j| \times \frac{|\omega_j|}{\eta_j\sqrt{T}} \\
&\leq \sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \eta_j^2 \prod_{i=j+1}^{t} |1-\lambda\eta_j| \times \sup_{\tau\in[0,t]} \frac{|\omega_\tau|}{\eta_\tau\sqrt{T}}
\end{aligned}
$$

$$\leq \left(\sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \eta_j^2 \prod_{i=j+1}^{t} |1-\lambda\eta_j|\right) \times \left(\sup_{\tau\in[0,T]} \frac{|\omega_\tau|}{\eta_\tau\sqrt{T}}\right).$$

The fact that λ has a positive real part implies when t is sufficiently large, we have

$$|1-\lambda\eta_t| = \sqrt{(1-\mathrm{Re}(\lambda)\eta_t)^2 + \mathrm{Im}(\lambda)^2\eta_t^2} = \sqrt{1-2\mathrm{Re}(\lambda)\eta_t + |\lambda|^2\eta_t^2} \lesssim 1-\mathrm{Re}(\lambda)\eta_t \lesssim \exp\left(-\mathrm{Re}(\lambda)\eta_t\right).$$

By 4 in Lemma B.7.1, there exists $c_1 > 0$ such that

$$\sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \eta_j^2 \prod_{i=j+1}^{t} |1-\lambda\eta_j| \leq c_1.$$

As a result, we have

$$\frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{|z_{t+1}|}{\eta_t} \leq \left(\sup_{t\in[0,T]} \frac{1}{\eta_t} \sum_{j=0}^{t} \eta_j^2 \prod_{i=j+1}^{t} |1-\lambda\eta_j|\right) \times \left(\sup_{t\in[0,T]} \frac{|\omega_t|}{\eta_t\sqrt{T}}\right) \leq c_1 \times \left(\sup_{t\in[0,T]} \frac{|\omega_t|}{\eta_t\sqrt{T}}\right).$$

We complete the proof by using the condition that $\frac{1}{\sqrt{T}} \sup_{t\in[0,T]} \frac{|\omega_t|}{\eta_t} \xrightarrow{p} 0$ as $T \to \infty$ and $\eta_t - \eta_{t+1} = \eta_t o(\eta_t)$. □

B.7 Properties of Recursion Matrices

Lemma B.7.1. *Recall that $\boldsymbol{B}_i := \boldsymbol{I} - \eta_i \boldsymbol{G}$ and $-\boldsymbol{G}$ is Hurwitz (i.e., $\mathrm{Re}\lambda_i(\boldsymbol{G}) > 0$ for all $i \in [d]$). For any $n \geq j$, define $\boldsymbol{X}_j^n$ and $\boldsymbol{A}_j^n$ as*

$$\boldsymbol{X}_j^n := \prod_{i=j}^{n} \boldsymbol{B}_i \tag{B.16}$$

$$\boldsymbol{A}_j^n := \sum_{t=j}^{n} \boldsymbol{X}_{j+1}^t \eta_j = \sum_{t=j}^{n} \left(\prod_{i=j+1}^{t} \boldsymbol{B}_i\right) \eta_j. \tag{B.17}$$

When $\{\eta_t\}_{t\geq 0}$ satisfies Assumption 3.2.5, it follows that

1. *There exist constant $c_0, c > 0$ such that for any $n \geq j \geq 0$,*

$$\|\boldsymbol{X}_j^n\| \leq c_0 \exp\left(-c \sum_{t=j}^{n} \eta_t\right).$$

2. *There exist C_0 such that $\boldsymbol{A}_j^n$ is uniformly bounded with respect to both j and n for*

$0 \le j \le k$ *(i.e.,* $\|\boldsymbol{A}_j^n\| \le C_0$ *for any* $n \ge j \ge 0$*), and*

$$\frac{1}{n}\sum_{j=0}^{n} \|\boldsymbol{A}_j^n - \boldsymbol{G}^{-1}\| \to 0 \text{ as } n \to \infty.$$

3. *For the c given in 1, there exists* $j_0 > 0$ *such that any* $n \ge j \ge j_0$*,*

$$\frac{\eta_{j-1}}{\eta_n} \le \exp\left(\frac{c}{2}\sum_{t=j-1}^{n} \eta_t\right).$$

4. *Let c be a positive constant, then there exists another constant* c_1 *such that for any* $n \ge 1$*,*

$$\sum_{j=1}^{n} \eta_{j-1}^2 \exp\left(-c\sum_{t=j}^{n} \eta_t\right) \le c_1\eta_n.$$

Proof of Lemma B.7.1. 1. The proof can be found in the proof of Lemma 3.1.1 in[173].

2. The proof can be found in the proof of Lemma 3.4.1 in[173].

3. Due to $\frac{\eta_{t-1}-\eta_t}{\eta_{t-1}} = o(\eta_{t-1})$, it follows that

$$\frac{\eta_{j-1}}{\eta_n} = \prod_{t=j}^{n-1} \frac{\eta_{t-1}}{\eta_t} = \prod_{t=j}^{n-1}(1 + o(1)\eta_t) \le \exp\left(o(1)\sum_{t=j}^{n-1} \eta_t\right) \le \exp\left(o(1)\sum_{t=j}^{n} \eta_t\right),$$

where $o(1)$ denotes a magnitude that tends to zero as $j \to \infty$ and the last inequality follows from $\eta_n \to 0$. We then find $j_0 > 0$ such that any $n \ge j \ge j_0$, we have $\frac{\eta_{j-1}}{\eta_n} \le \exp\left(\frac{c}{2}\sum_{t=j-1}^{n} \eta_t\right)$ with c given in 1.

4. Lemma 3.3.2 in[173] implies that for any $c > 0$, $\sum_{j=1}^{n} \eta_{j-1} \exp\left(-c\sum_{t=j}^{n} \eta_t\right)$ is uniformly bounded for $n \ge 1$. When $n \to \infty$, as a result of $n\eta_n \to \infty$, we have $c\sum_{t=1}^{n} \eta_t + \ln \eta_n \to \infty$ and thus $\eta_n^{-1} \exp\left(-c\sum_{t=1}^{n} \eta_t\right) \to 0$. Therefore, we can find $n_0, c_3 > 0$ such that any $n \ge n_0$ we have $\sum_{j=1}^{j_0} \eta_{j-1}^2 \exp\left(-c\sum_{t=j}^{n} \eta_t\right) \le c_3\eta_n$. Then as long as $n \ge \max\{j_0, n_0\}$, it follows that

$$\begin{aligned}\sum_{j=1}^{n} \eta_{j-1}^2 \exp\left(-c\sum_{t=j}^{n} \eta_t\right) &= \sum_{j=1}^{j_0} \eta_{j-1}^2 \exp\left(-c\sum_{t=j}^{n} \eta_t\right) + \sum_{j=j_0}^{n} \eta_{j-1}^2 \exp\left(-c\sum_{t=j}^{n} \eta_t\right)\\ &= c_3\eta_n + \eta_n\sum_{j=j_0}^{n} \eta_{j-1} \exp\left(-\frac{c}{2}\sum_{t=j}^{n} \eta_t\right) \le c_1\eta_n.\end{aligned}$$

For $n < \max\{j_0, n_0\}$, since there is only a finite number of cases, we can enlarge c_1 in order to cover all $n \ge 1$.

□

Appendix C Omitted Proofs for Theorem 3.3.2

C.1 Proof of Lemma 3.4.5

Proof of Lemma 3.4.5.

- Under Assumption 3.2.5, we have $\eta_t \downarrow 0$ and $t\eta_t \uparrow \infty$ as $t \to \infty$. Hence, for any fixed $m > 0$, we have $t\eta_t \geq m$ for sufficiently large t. Then,

$$a_t = \lceil \log_\rho \frac{\eta_t}{\sigma\kappa} \rceil = \lceil \log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_t} \rceil \leq \lceil \log_{\frac{1}{\rho}} \frac{\sigma\kappa t}{m} \rceil \implies a_t = \mathcal{O}(\log t).$$

- Since $a_t = \mathcal{O}(\log t)$, for sufficiently large t, there exists $\mu > 0$ such that $a_t \leq \mu \log t$ and thus

$$a_t \eta_{t-a_t} \log t \leq \mu \log^2 t \cdot \eta_{t-\mu\log t} = \frac{\mu \log^2 t}{\log^2(t-\mu\log t)} \cdot \eta_{t-\mu\log t} \log^2(t-\mu\log t) = o(1),$$

where we use $\eta_t \log^2 t = o(1)$ when t goes to infinity due to Assumption 3.2.5.

- It follows that

$$\frac{\eta_{t-a_t}}{\eta_t} = \prod_{\tau=t-a_t}^{t-1} \frac{\eta_\tau}{\eta_{\tau+1}} = \prod_{\tau=t-a_t}^{t-1} (1+o(\eta_t)) \leq \exp\left(o(1) \sum_{\tau=t-a_t}^{t-1} \eta_\tau \right) \leq \exp(o(1) a_t \eta_{t-a_t}) = \mathcal{O}(1).$$

- By $\eta_{t+1} = \eta_t(1+o(\eta_t))$, it follow that

$$\log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_{t+1}} = \log_{\frac{1}{\rho}} \frac{\sigma\kappa(1+o(\eta_t))}{\eta_t} = \log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_t} + \log_{\frac{1}{\rho}}(1+o(\eta_t)) = \log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_t} + o(\eta_t).$$

For sufficiently large t, we will have $o(\eta_t) \leq 0.5$ and thus

$$a_{t+1} = \left\lceil \log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_{t+1}} \right\rceil \leq \left\lceil \log_{\frac{1}{\rho}} \frac{\sigma\kappa}{\eta_t} \right\rceil + 1 = a_t + 1.$$

It is clearly that we have $a_t \leq a_{t+1}$ due to $\eta_t \downarrow 0$.

□

C.2 Proof of Lemma 3.4.6

Proof of Lemma 3.4.6. The conclusion is obvious if $a_t = 0$. Without loss of generality, we assume $\rho > 0$ and $a_t \geq 1$. Recall that the update rule is $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)$. Hence, under Assumption 3.3.2,

$$|\|\boldsymbol{x}_{t+1}\| - \|\boldsymbol{x}_t\|| \leq \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = \eta_t \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\| \leq M\eta_t(\|\boldsymbol{x}_t\| + g(\xi_t)),$$

which implies that

$$\|\boldsymbol{x}_{t+1}\| \le (1 + M\eta_t)\|\boldsymbol{x}_t\| + M\eta_t g(\xi_t).$$

For simplicity, we denote $\eta_{s,t} := \sum_{l=s}^{t} \eta_l$ is $s \le t$ otherwise $\eta_{s,t} := 0$ for $s > t$. Iterating the last inequality yields for any $t - a_t - 1 \le \tau \le t - 1$ with $t \ge K$,

$$\begin{aligned}
\|\boldsymbol{x}_{\tau+1}\| &\le \prod_{s=t-a_t}^{\tau} (1 + M\eta_s)\|\boldsymbol{x}_{t-a_t}\| + M \sum_{s=t-a_t}^{\tau} \eta_s g(\xi_s) \prod_{l=s+1}^{\tau} (1 + M\eta_l) \\
&\le \exp\left(M\eta_{t-a_t,\tau}\right) \|\boldsymbol{x}_{t-a_t}\| + M \sum_{s=t-a_t}^{\tau} \eta_s g(\xi_s) \exp\left(M\eta_{s+1,\tau}\right) \\
&\overset{(a)}{\le} \exp\left(M a_t \eta_{t-a_t}\right) \left[\|\boldsymbol{x}_{t-a_t}\| + M \sum_{s=t-a_t}^{\tau} \eta_s g(\xi_s) \right] \\
&\overset{(b)}{\le} 2\left(\|\boldsymbol{x}_{t-a_t}\| + M a_t \eta_{t-a_t} g_{t-1} \right),
\end{aligned}$$

where (a) uses $\eta_{t-a_t,\tau} \le a_t \eta_{t-a_t}$ by definition and (b) uses Lemma 3.4.5 and the definition of g_{t-1} in (3.36).

As a result,

$$\begin{aligned}
\|\boldsymbol{x}_t - \boldsymbol{x}_{t-a_t}\| &\le \sum_{\tau=t-a_t}^{t-1} \|\boldsymbol{x}_{\tau+1} - \boldsymbol{x}_\tau\| \le \sum_{\tau=t-a_t}^{t-1} M\eta_\tau(\|\boldsymbol{x}_\tau\| + g(\xi_\tau)) \\
&\le \sum_{\tau=t-a_t}^{t-1} M\eta_\tau(2\|\boldsymbol{x}_{t-a_t}\| + 2M a_t \eta_{t-a_t} g_{t-1}) + M\eta_{t-a_t,t-1} g_{t-1} \\
&\le 2M\eta_{t-a_t,t-1}(\|\boldsymbol{x}_{t-a_t}\| + g_{t-1}) \le 2M a_t \eta_{t-a_t}(\|\boldsymbol{x}_{t-a_t}\| + g_{t-1}),
\end{aligned}$$

where the last inequality uses Lemma 3.4.5 and $M a_t \eta_{t-a_t} \le \log 2 \le \frac{1}{2}$.

Therefore, using $\log 2 \le \frac{1}{3}$, we have

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-a_t}\| \le 2M a_t \eta_{t-a_t}(\|\boldsymbol{x}_{t-a_t} - \boldsymbol{x}_t\| + \|\boldsymbol{x}_t\| + g_{t-1}) \le \frac{2}{3}\|\boldsymbol{x}_{t-a_t} - \boldsymbol{x}_t\| + 2M a_t \eta_{t-a_t}(\|\boldsymbol{x}_t\| + g_{t-1}),$$

which implies

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-a_t}\| \le 6M a_t \eta_{t-a_t}(\|\boldsymbol{x}_t\| + g_{t-1}) \le 2(\|\boldsymbol{x}_t\| + g_{t-1}).$$

□

C.3 Proof of Lemma 3.4.7

Proof of Lemma 3.4.7. Our target is to prove

$$\mathbb{E} \sup_{t-a_t \le \tau \le t-1} |g(\xi_\tau)| \le \left(\mathbb{E} \sup_{t-a_t \le \tau \le t-1} |g(\xi_\tau)|^{\frac{p}{2}} \right)^{\frac{2}{p}} = \mathcal{O}(a_t).$$

The left inequality follows from Jensen's inequality. We then focus on the right equality.

The fact $p > 2$ implies $\frac{2}{p} < 1$. Then $(x+y)^{\frac{2}{p}} \le x^{\frac{2}{p}} + y^{\frac{2}{p}}$ for any $x, y \ge 0$. Therefore,

$$\begin{aligned} \left(\mathbb{E} \sup_{t-a_t \le \tau \le t-1} |g(\xi_\tau)|^{\frac{p}{2}} \right)^{\frac{2}{p}} &\le \left(\mathbb{E} \sum_{t-a_t \le \tau \le t-1} |g(\xi_\tau)|^{\frac{p}{2}} \right)^{\frac{2}{p}} \\ &\le \sum_{t-a_t \le \tau \le t-1} \left(\mathbb{E} |g(\xi_\tau)|^{\frac{p}{2}} \right)^{\frac{2}{p}} \le a_t \cdot \sup_{t \ge 0} \left(\mathbb{E} |g(\xi_t)|^{\frac{p}{2}} \right)^{\frac{2}{p}} \lesssim a_t. \end{aligned}$$

□

C.4 Proof of Lemma 3.4.8

Proof of Lemma 3.4.8. By homogeneity, we only need to prove for the case of $A = 1$.

- When $\alpha \in (0, 1]$, we let $f(x) = 1 + (1+\alpha)x + |x|^{1+\alpha} - (1+x)^{1+\alpha}$ and its derivative is $f'(x) = (1+\alpha)\left(1 + |x|^{\alpha}\mathrm{sign}(x) - (1+x)^{\alpha}\right)$. When $1 \ge \alpha > 0$, we have $(1+x)^{\alpha} \le x^{\alpha} + 1$ for $x \ge 0$ and $1 \le (1-x)^{\alpha} + x^{\alpha}$ for $x \in [0, 1]$. It implies that $f'(x) \ge 0$ for $x \ge 0$ and $f'(x) < 0$ for $-1 \le x < 0$. Hence, $f(x) \ge f(0) = 0$ for any $x \ge -1$.
- When $\alpha \in [1, \infty)$, we let $f(x) = 1 + (1+\alpha)x + \frac{c_\alpha(1+\alpha)}{2}x^2 + c_\alpha|x|^{1+\alpha} - (1+x)^{1+\alpha}$ and its derivative is $f'(x) = (1+\alpha)\left[1 + c_\alpha(x + |x|^{\alpha}\mathrm{sign}(x)) - (1+x)^{\alpha}\right]$. Similarly, we are going to show $f'(x) \ge 0$ for $x \ge 0$ and $f'(x) < 0$ for $-1 \le x < 0$. These two conditions is equivalent to

  $$c_\alpha \ge \begin{cases} \frac{(1+x)^{\alpha}-1}{x+x^{\alpha}} & \text{if } 0 \le x < \infty; \\ \frac{1-(1-x)^{\alpha}}{x+x^{\alpha}} & \text{if } 0 \le x \le 1. \end{cases}$$

  The last inequality is satisfied when we set

  $$c_\alpha := \sup_{x \ge 0} \frac{(1+x)^{\alpha} - 1}{x + x^{\alpha}}.$$

  We explain the reason in the following. Since $(1-x)^r \ge 1 - rx$ for any $x \in [0, 1]$ and $r \ge 1$, we have $\sup_{x \in [0,1]} \frac{1-(1-x)^{\alpha}}{x+x^{\alpha}} \le c_\alpha \le \sup_{x \in [0,1]} \frac{x\alpha}{x+x^{\alpha}} = c_\alpha \le \sup_{x \in [0,1]} \frac{\alpha}{1+x^{\alpha-1}} = \alpha$. Let $h(x) = \frac{(1+x)^{\alpha}-1}{x+x^{\alpha}}$, then $c_\alpha = \sup_{x \in [0,\infty)} h(x)$. One can easily show that, on the interval

$(0, \infty)$, $h(x)$ is a continuous function with $\lim_{x \to 0^+} h(x) = \alpha$ and $\min_{x \to \infty} h(x) = 1$. As a result, we know that $\sup_{x \in [0,\infty)} h(x)$ is finite and no smaller than $h(0) := \alpha$.

We complete the proof by showing $c_\alpha \le^\alpha$ in the following. If $x \ge 1$, we have

$$\left(\frac{(1+x)^\alpha - 1}{x + x^\alpha} \right)^{\frac{1}{\alpha}} \le \left(\frac{(1+x)^\alpha}{x^\alpha} \right)^{\frac{1}{\alpha}} \le \frac{1+x}{x} \le 2.$$

If $0 \le x \le 1$, using $(1+x)^\alpha - 1 \le \alpha x (1+x)^{\alpha-1}$, we have that for any $\alpha \ge 1$,

$$\left(\frac{(1+x)^\alpha - 1}{x + x^\alpha} \right)^{\frac{1}{\alpha}} \le \left(\frac{\alpha(1+x)^{\alpha-1}}{1 + x^{\alpha-1}} \right)^{\frac{1}{\alpha}} \le (\alpha 2^{\alpha-1})^{\frac{1}{\alpha}} \le 3.$$

□

C.5 Proof of Lemma 3.4.9

Proof of Lemma 3.4.9. The main idea is to decompose $\mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \delta_t$ into three terms and then bound each term respectively. It follows that

$$\begin{aligned} \mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \delta_t = {} & \mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle \\ & + \mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle \\ & - \frac{(\bar{p}-1)\lambda\eta_t}{2} \cdot \mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^2. \end{aligned} \tag{C.1}$$

For the first term From (3.30) and (3.31), it follows that

$$\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle \ge A_3 M(\boldsymbol{x}_t - \boldsymbol{x}^\star). \tag{C.2}$$

For the second term Similar to (3.33), we have

$$\begin{aligned} & |\mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star)), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle| \\ & \quad \le |\mathbb{E} M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star)) - \nabla M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star)), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle| \\ & \quad\quad + |\mathbb{E} M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star)^\alpha \langle \nabla M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star)), \mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-a_t-1}) \rangle| \\ & \quad\quad + |\mathbb{E} \left(M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha - M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star)^\alpha \right) \langle \nabla M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star)), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle| \\ & \quad := T_1 + T_2 + T_3. \end{aligned}$$

We are going to analyze the three terms separately. By (3.37), we have

$$|\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star) - \nabla M(\boldsymbol{x}_{t-a_t} - \boldsymbol{x}^\star), \mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1}) \rangle|$$

$$\leq 6\sigma M(\bar{p}-1)u_{\bar{p}}^2\lambda \cdot a_t\eta_{t-a_t}\left(\left(1+\frac{\lambda}{l_{\bar{p}}^2}\right)M(\boldsymbol{x}_t-\boldsymbol{x}^\star)+\|\boldsymbol{x}^\star\|+g_{t-1}+1\right),$$

which implies the first term T_1 satisfies

$$\begin{aligned}
T_1 &\lesssim a_t\eta_{t-a_t}(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha}+\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{\alpha}(1+g_{t-1}))\\
&\overset{(a)}{\lesssim} a_t\eta_{t-a_t}(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha}+(1+(\mathbb{E}g_{t-1}^{\frac{p}{2}})^{\frac{2}{p}})(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}})\\
&\overset{(b)}{\lesssim} a_t\eta_{t-a_t}(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha}+a_t\cdot(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}),
\end{aligned}$$

where (a) uses Holder's inequality and $2(1+\alpha)=p$ for simplicity and (b) uses $(\mathbb{E}h_{t-1}^{\frac{p}{2}})^{\frac{2}{p}}\lesssim \log t$ due to Lemma 3.4.7. By (a) in (3.38), we have

$$\begin{aligned}
&|\langle\nabla M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star),\mathscr{P}^{a_t+1}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-a_t-1})\rangle|\\
&\quad\leq \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda\cdot\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|\\
&\quad\leq \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda\cdot\frac{\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|^2+1}{2}\\
&\quad\leq \eta_t(\bar{p}-1)u_{\bar{p}}^2\lambda\cdot\left(\left(1+\frac{\lambda}{l_{\bar{p}}^2}\right)M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)+1\right),
\end{aligned}$$

which implies the second term T_2 satisfies

$$\begin{aligned}
T_2 &\lesssim \eta_t(\mathbb{E}M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)^{1+\alpha}+\mathbb{E}M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)^{\alpha})\\
&\lesssim \eta_t(\mathbb{E}M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)^{1+\alpha}+(\mathbb{E}M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}).
\end{aligned}$$

Finally, as for the third term T_3, by a similar argument of the last inequality, we have

$$|\langle\nabla M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star),\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\rangle|\leq\sigma(\bar{p}-1)u_{\bar{p}}^2\lambda\cdot\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|.$$

On the other hand, noticing that $|\|\boldsymbol{x}\|_M^{2\alpha}-\|\boldsymbol{y}\|_M^{2\alpha}|\leq 2\alpha\|\boldsymbol{x}-\boldsymbol{y}\|_M\cdot\max\{\|\boldsymbol{x}\|_M,\|\boldsymbol{y}\|_M\}^{2\alpha-1}$, we have

$$|M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{\alpha}-M(\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star)^{\alpha}|\leq\frac{\alpha}{2^{\alpha-1}}\|\boldsymbol{x}_t-\boldsymbol{x}_{t-a_t}\|_M\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha-1}.$$

As a result, we have

$$\begin{aligned}
T_3 &\lesssim \mathbb{E}\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|\|\boldsymbol{x}_t-\boldsymbol{x}_{t-a_t}\|_M\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha-1}\\
&\lesssim \mathbb{E}\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\|\boldsymbol{x}_t-\boldsymbol{x}_{t-a_t}\|\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha-1}\\
&\lesssim a_t\eta_{t-a_t}\mathbb{E}(\|\boldsymbol{x}_t\|+g_{t-1})\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha-1}\\
&\lesssim a_t\eta_{t-a_t}\mathbb{E}(\|\boldsymbol{x}_t\|_M+g_{t-1})\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha}
\end{aligned}$$

$$
\begin{aligned}
&\lesssim a_t\eta_{t-a_t}\mathbb{E}(\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}+g_{t-1})\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha}\\
&\lesssim a_t\eta_{t-a_t}\mathbb{E}(\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^2+g_{t-1}+1)\cdot\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}^{2\alpha}\\
&\overset{(a)}{\lesssim} a_t\eta_{t-a_t}\mathbb{E}(b_t^2+g_{t-1}+1)\cdot b_t^{2\alpha}\lesssim a_t\eta_{t-a_t}\left[\mathbb{E}b_t^{2(1+\alpha)}+\mathbb{E}(g_{t-1}+1)b_t^{2\alpha}\right]\\
&\overset{(b)}{\lesssim} a_t\eta_{t-a_t}\left[\mathbb{E}b_t^{2(1+\alpha)}+(1+(\mathbb{E}g_{t-1}^{\frac{p}{2}})^{\frac{2}{p}})(\mathbb{E}b_t^{2(1+\alpha)})^{\frac{\alpha}{1+\alpha}}\right]\\
&\overset{(c)}{\lesssim} a_t\eta_{t-a_t}\left(\mathbb{E}b_t^{2(1+\alpha)}+a_t\cdot(\mathbb{E}b_t^{2(1+\alpha)})^{\frac{\alpha}{1+\alpha}}\right)\\
&\overset{(d)}{\lesssim} a_t\eta_{t-a_t}\left(d_t+a_t\cdot d_t^{\frac{\alpha}{1+\alpha}}\right),
\end{aligned}
$$

where (a) follows the notation $b_t=\max\{\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|_M,\|\boldsymbol{x}_{t-a_t}-\boldsymbol{x}^\star\|_M\}$, (b) uses Holder's inequality and $2(1+\alpha)=p$, (c) uses $(\mathbb{E}h_{t-1}^{\frac{p}{2}})^{\frac{2}{p}}\lesssim a_t$ due to Lemma 3.4.7, and (d) uses the notation $d_t:=\max_{t-a_t\le\tau\le t}\mathbb{E}M(\boldsymbol{x}_\tau-\boldsymbol{x}^\star)^{1+\alpha}$ and $d_t\ge\mathbb{E}b_t^{2(1+\alpha)}$ by definition.

Combing the bounds for T_1, T_2 and T_3, we have

$$
|\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^\alpha\langle\nabla M(\boldsymbol{x}_t-\boldsymbol{x}^\star)),\mathscr{P}\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\rangle|\lesssim a_t\eta_{t-a_t}\left(d_t+a_t\cdot d_t^{\frac{\alpha}{1+\alpha}}\right). \tag{C.3}
$$

For the last term Finally, we analyze the last term of (C.1). It follows from Hölder's inequality that

$$
\begin{aligned}
\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^\alpha\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)\|_{\bar{p}}^2&\le(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}(\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)\|_{\bar{p}}^{2(1+\alpha)})^{\frac{1}{1+\alpha}}\\
&\lesssim(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}(\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)\|^{2(1+\alpha)})^{\frac{1}{1+\alpha}}\\
&\lesssim(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}(\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})\|^p)^{\frac{1}{1+\alpha}}.
\end{aligned}
$$

By Assumption 3.2.3 and 3.2.2, we have

$$
\begin{aligned}
\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})\|^p&\lesssim\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|^p+\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|^p\\
&\lesssim\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_{t-1})-\boldsymbol{H}(\boldsymbol{x}^\star,\xi_{t-1})\|^p+\sup_{t\ge0}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi_t)\|^p\\
&\lesssim\mathbb{E}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^p+1\lesssim\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha}+1.
\end{aligned} \tag{C.4}
$$

Using $(x+1)^{\frac{1}{1+\alpha}}\le x^{\frac{1}{1+\alpha}}+1$ for $x\ge0$, we have

$$
\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^\alpha\|\boldsymbol{H}(\boldsymbol{x}_t,\xi_t)\|_{\bar{p}}^2\lesssim\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha}+(\mathbb{E}M(\boldsymbol{x}_t-\boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha}{1+\alpha}}. \tag{C.5}
$$

Plugging (C.2), (C.3) and (C.5) into (C.1), we complete the proof. □

C.6 Proof of Lemma 3.4.10

Proof of Lemma 3.4.10. By the definition of δ_t, it follows that

$$\mathbb{E}|\delta_t|^{1+\alpha} \lesssim \mathbb{E}|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^{1+\alpha} + \eta_t^{1+\alpha}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^{2(1+\alpha)}$$

For one thing, by Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^{1+\alpha} &\le \mathbb{E}\|\nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_{\bar{q}}^{1+\alpha} \cdot \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^{1+\alpha} \\ &\lesssim \mathbb{E}\|\nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_{\bar{q}}^{2(1+\alpha)} + \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^{2(1+\alpha)}.\end{aligned}$$

Since $\boldsymbol{M}(\boldsymbol{x})$ is smooth w.r.t. the norm $\|\cdot\|_{\bar{p}}$ (due to 1 in Lemma 3.4.4) and $\nabla \boldsymbol{M}(\boldsymbol{0}) = \boldsymbol{0}$, we have

$$\mathbb{E}\|\nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_{\bar{q}}^{2(1+\alpha)} = \mathbb{E}\|\nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)\|_{\bar{q}}^{p} \lesssim \mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_{\bar{p}}^{p} \lesssim \mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^{p} \lesssim \mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha},$$

where the last inequality uses the fact $\|\cdot\|$ is equivalent to $\|\cdot\|_M$ up to constant factors and $\boldsymbol{M}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_M^2$. For another thing,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^{2(1+\alpha)} &\lesssim \mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|^{2(1+\alpha)} = \mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1})\|^{p} \\ &\lesssim \mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} + 1,\end{aligned}$$

where the last inequality follows from (C.4).

Putting two pieces together, we have

$$\mathbb{E}|\delta_t|^{1+\alpha} \lesssim \mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} + \eta_t^{1+\alpha}.$$

□

C.7 Proof of Lemma 3.4.11

Proof of Lemma 3.4.11. By definition of δ_t in (3.41), it follows that

$$|\delta_t|^2 \le 2|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^2 + \frac{(\bar{p}-1)^2\lambda^2\eta_t^2}{2}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^4,$$

by which, $\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\delta_t|^2$ can be further divided into two terms

$$\begin{aligned}\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\delta_t|^2 &\le 2\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\langle \nabla \boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^2 \\ &\quad + \frac{(\bar{p}-1)^2\lambda^2\eta_t^2}{2}\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^4.\end{aligned}$$

For the first term We first note that by a similar argument of (a) in (3.38), we have

$$|\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle| \le (\bar{p}-1)u_{\bar{p}}^2 \lambda \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \cdot \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|.$$

Second, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|^2 | \mathscr{F}_{t-1}] &= \mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1})\|^2 \\
&\le 2\left[\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1}) - \boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\|^2 + \mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\|^2\right] \\
&\overset{(a)}{\le} 2L_H^2\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + 2\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star, \xi_{t-1})\|^2 \\
&\overset{(b)}{\le} 2L_H^2\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + 4M^2(\|\boldsymbol{x}^\star\|^2 + \mathscr{P}g^2(\xi_{t-1})),
\end{aligned}$$

where (a) uses Assumption 3.2.3 and (b) uses Assumption 3.3.2. As a result

$$\begin{aligned}
&\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}|\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^2 \\
&\quad\lesssim \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}\left[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \cdot \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|^2\right] \\
&\quad= \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \cdot \mathbb{E}[\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|^2 | \mathscr{F}_{t-1}] \\
&\quad\lesssim \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \cdot \left[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{x}^\star\|^2 + \mathscr{P}g^2(\xi_{t-1})\right] \\
&\quad\overset{(a)}{\lesssim} \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha+1} + \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha} + \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha}g^2(\xi_t) \\
&\quad\overset{(b)}{\le} \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha+1} + \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha} + \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}\right)^{\frac{\alpha}{\alpha+1}},
\end{aligned}$$

where (a) uses 2 in Lemma 3.4.4 and (b) uses the following inequality (proved by Hölder's inequality) and $p = 2(1+\alpha)$,

$$\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha}g^2(\xi_t) \le \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha\frac{p}{p-2}}\right)^{1-\frac{2}{p}}(\mathbb{E}|g(\xi_t)|^p)^{\frac{2}{p}} \lesssim \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}\right)^{\frac{\alpha}{\alpha+1}}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}&|\langle \nabla M(\boldsymbol{x}_t - \boldsymbol{x}^\star), \boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\rangle|^2 \\
&\lesssim \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha+1} + \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha} + \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}\right)^{\frac{\alpha}{\alpha+1}} \\
&\le \mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha+1} + \left(\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha}\right)^{\frac{\alpha}{\alpha+1}}.
\end{aligned}$$

For the second term It follows from Hölder's inequality that

$$\begin{aligned}
\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{\alpha-1}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^4 &\le (\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha-1}{1+\alpha}}(\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^{2(1+\alpha)})^{\frac{2}{1+\alpha}} \\
&\lesssim (\mathbb{E}M(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha-1}{1+\alpha}}(\mathbb{E}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|^{2(1+\alpha)})^{\frac{2}{1+\alpha}}
\end{aligned}$$

$$\lesssim (\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha-1}{1+\alpha}} (\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1})\|^p)^{\frac{2}{1+\alpha}}.$$

By (C.4), we have $\mathbb{E}\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}_t, \xi_{t-1})\|^p \lesssim \mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} + 1$. Using $(x+1)^{\frac{1}{1+\alpha}} \le x^{\frac{1}{1+\alpha}} + 1$ for $x \ge 0$, we again have

$$\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^\alpha \|\boldsymbol{H}(\boldsymbol{x}_t, \xi_t)\|_{\bar{p}}^2 \lesssim \mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha} + (\mathbb{E}\boldsymbol{M}(\boldsymbol{x}_t - \boldsymbol{x}^\star)^{1+\alpha})^{\frac{\alpha-1}{\alpha+1}}. \tag{C.5}$$

Combing these two parts, we complete the proof. □

Appendix D Omitted Proofs for Theorem 3.3.5

D.1 Proof of Corollary 3.3.3

Proof of Corollary 3.3.3. We proceed with the proof by discussing two scenarios.

Linear SA with i.i.d. data We first consider a simple case that is linear SA with i.i.d. data. In this case, $t_{\mathrm{mix}} = 0$ and $c_r = 0$ so that the second term in (3.14) disappear and the bound (3.14) becomes

$$\tilde{\mathcal{O}}\left(T^{-J_1(\alpha)} + T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta}\wedge\frac{1}{3}\right]} + t_{\mathrm{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}}\right) = \tilde{\mathcal{O}}\left(T^{-h_1(\alpha)}\right),$$

where $J_1(\cdot)$ is defined in (3.60) and

$$h_1(\alpha) = \min\left\{J_1(\alpha), (1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]\right\}. \tag{D.1}$$

In the following, we maximize (D.1) by considering different values of δ. To ensure the optimal α^* is achievable, we consider $\alpha \in [0.5+\varepsilon, 1)$ for a very small $\varepsilon > 0$. Note that $J_1(\alpha)$ strictly increases in α and has a unique intersection point with the straight line $\frac{\delta(1-\alpha)}{3+2\delta}$ on the interval $[0, 1]$.

1. If $\delta \in \left(0, \frac{1-2\varepsilon}{1+2\varepsilon}\right]$, we then have $0.5 + \varepsilon \le \frac{1}{1+\delta}$. (D.1) becomes $\min\left\{\frac{\delta(1-\alpha)}{3+2\delta}, \frac{\alpha(1+\delta)}{3+2\delta}\right\}$. One can show that, $\frac{\alpha(1+\delta)}{3+2\delta}$ intersect with $\frac{\delta(1-\alpha)}{3+2\delta}$ at $\alpha_1 := \frac{\delta}{2\delta+1}$ which doesn't lie in the interval we consider. Hence, $\max_{\alpha\in[0.5+\varepsilon,1)} h_1(\alpha) = \max_{\alpha\in[0.5+\varepsilon,1)} \frac{\delta(1-\alpha)}{3+2\delta} = \frac{\delta(1-2\varepsilon)}{6+4\delta}$.
2. If $\delta \in \left[\frac{1-2\varepsilon}{1+2\varepsilon}, 3\right]$, we then have $\frac{1}{1+\delta} \le 0.5 + \varepsilon$. (D.1) becomes $\min\left\{\frac{\delta(1-\alpha)}{3+2\delta}, \frac{\alpha}{2+\alpha}\right\}$. Denote the intersection point between $\frac{\delta(1-\alpha)}{3+2\delta}$ and $\frac{\alpha}{2+\alpha}$ by α_2. Direct calculation yields $\alpha_2 = \frac{\sqrt{9(1+\delta)^2+8\delta^2}-3(1+\delta)}{2\delta}$ and $\alpha_2 < 0.5$ for all $\delta \le 3$. It implies the two segments doesn't intersect at the given interval. Hence, $\max_{\alpha\in[0.5+\varepsilon,1)} h_1(\alpha) = \max_{\alpha\in[0.5+\varepsilon,1)} \frac{\delta(1-\alpha)}{3+2\delta} = \frac{\delta(1-2\varepsilon)}{6+4\delta}$.
3. If $\delta \in [3, \infty)$,(D.1) becomes $\min\left\{\frac{1-\alpha}{3}, \frac{\alpha}{2+\alpha}\right\}$. Denote the intersection point between $\frac{1-\alpha}{3}$ and $\frac{\alpha}{2+\alpha}$ by α_3. Direct calculation yields $\alpha_3 = \sqrt{6} - 2 < 0.5$. It implies the two curves doesn't intersect at the given interval. So, $\max_{\alpha\in[0.5+\varepsilon,1)} h_1(\alpha) = \max_{\alpha\in[0.5+\varepsilon,1)} \frac{1-\alpha}{3} = \frac{1-2\varepsilon}{6}$.

Putting pieces together, we have

$$\min_{\alpha\in[0.5+\varepsilon,1)} d_{\mathrm{P}}\boldsymbol{P}\left(\boldsymbol{\theta}^\top\boldsymbol{\phi}_T, \boldsymbol{\theta}^\top\boldsymbol{\psi}\right) = \tilde{\mathcal{O}}\left(T^{-f_1(\delta)}\right),$$

where

$$f_1(\delta) = \left[\frac{\delta}{6+4\delta} \wedge \frac{1}{6}\right](1-2\varepsilon). \tag{D.2}$$

Other cases One can show that $J_1(\alpha) \geq J_2(\alpha)$ for any $\alpha \in (0.5, 1)$. Once $t_{\text{mix}} > 0$ or $c_r > 0$, the bound (3.14) becomes

$$\widetilde{\mathcal{O}}\left((c_r + t_{\text{mix}})^{\frac{p}{2+p}} \cdot T^{-J_2(\alpha)} + T^{-(1-\alpha)\left[\frac{\delta}{3+2\delta}\wedge\frac{1}{3}\right]} + t_{\text{mix}}^{\frac{1}{6}} T^{-\frac{1}{6}}\right).$$

where $J_2(\cdot)$ is defined in (3.61) and

$$h_2(\alpha) = \min\left\{J_2(\alpha), (1-\alpha)\left[\frac{\delta}{3+2\delta} \wedge \frac{1}{3}\right]\right\}. \tag{D.3}$$

One can find that $J_2(\alpha)$ and $\frac{\delta(1-\alpha)}{3+2\delta}$ intersect at a unique point. Denote a polynomial function ℓ by $\ell(\delta) := 4\delta^3 + 7\delta^2 - 2\delta - 3$. One can find that (i) $\ell(\delta)$ strictly increases in $\delta \in (0.5, \infty)$ and (ii) there exists a unique $\delta_0 \in (0.5, 1)$ such that $\ell(\delta_0) = 0$.

In the following, we maximize (D.3) by considering different values of δ.

1. If $\delta \in (0, \delta_0]$, denote the solution of $\frac{\delta(1-\alpha)}{3+2\delta} = (\alpha - 0.5)\frac{1+\delta}{2+\delta}$ by α_1. Direct calculation yields $\alpha_1 = \frac{4\delta^2+9\delta+3}{6\delta^2+14\delta+6}$. One can show that $0.5 < \alpha_1 \leq \frac{1}{1+\delta}$. The right-hand side inequality is equivalent to $\ell(\delta) := 4\delta^3 + 7\delta^2 - 2\delta - 3 \leq 0$, which is true because $\delta \leq \delta_0$. Hence, $\max_{\alpha\in(0.5,1)} h_2(\alpha) = \frac{\delta(2\delta^2+5\delta+3)}{2(3+2\delta)(3\delta^2+7\delta+3)} = \frac{\delta(\delta+1)}{2(3\delta^2+7\delta+3)}$.
2. If $\delta \in [\delta_0, 3]$, denote the solution of $\frac{\delta(1-\alpha)}{3+2\delta} = \frac{\alpha-0.5}{\alpha+1}$ by α_2. Direct calculation yields $\alpha_2 = \frac{\sqrt{(3+2\delta)^2+2\delta(3+4\delta)}-(3+2\delta)}{2\delta} > 0.5$. Once can show that $\alpha_2 \geq \frac{1}{\delta+1}$. This is because the inequality is equivalent to $\ell(\delta) := 4\delta^3 + 7\delta^2 - 2\delta - 3 \geq 0$ which is true because $\delta \geq \delta_0$. Hence, $\max_{\alpha\in(0.5,1)} h_2(\alpha) = \frac{3+4\delta-\sqrt{(3+2\delta)^2+2\delta(3+4\delta)}}{2(3+2\delta)}$.
3. If $\delta \in [3, \infty)$, denote the solution of $\frac{1-\alpha}{3} = \frac{\alpha-0.5}{\alpha+1}$ by α_3. Direct calculation yields $\alpha_3 = \frac{\sqrt{19}-3}{2}$ and $1 > \alpha_3 > 0.5 > \frac{1}{1+\delta}$. Hence, $\max_{\alpha\in(0.5,1)} h_2(\alpha) = \frac{5-\sqrt{19}}{6}$

Putting pieces together, we have

$$\min_{\alpha\in(0.5,1)} d_{\text{P}}\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \boldsymbol{\psi}\right) = \mathcal{O}\left(\left[(c_r + t_{\text{mix}})^{\frac{p}{2+p}} + 1\right] \cdot T^{-f_2(\delta)}\right),$$

where

$$f_2(\delta) = \begin{cases} \frac{\delta(2\delta^2+5\delta+3)}{2(3+2\delta)(3\delta^2+7\delta+3)} & \text{if } \delta \in (0, \delta_0], \\ \frac{3+4\delta-\sqrt{(3+2\delta)^2+2\delta(3+4\delta)}}{2(3+2\delta)} & \text{if } \delta \in [\delta_0, 3], \\ \frac{5-\sqrt{19}}{6} & \text{if } \delta \in [3, \infty). \end{cases} \tag{D.4}$$

□

D.2 Proof of Lemma 3.4.13

Proof of Lemma 3.4.13. We will analyze each term in (3.50) respectively. For simplicity, we define

$$C_{U,\boldsymbol{x}_t} = \kappa t_{\text{mix}} \cdot (2L_H \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \sigma).$$

For $\boldsymbol{\phi}_T - \tilde{\boldsymbol{\phi}}_T$ Recall $p = 2(1+\delta)$ and let $m \in [0, 2\delta+1]$ such that $1 \le 1 + m \le p$. It follows that

$$\begin{aligned}
d_{\mathrm{P}}\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_T, \boldsymbol{\theta}^\top \tilde{\boldsymbol{\phi}}_T\right) &\overset{(a)}{\le} d_{\mathrm{P}}\left(\boldsymbol{\phi}_T, \tilde{\boldsymbol{\phi}}_T\right) \overset{(b)}{\le} \tilde{d}\left(\frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \eta_t \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\right) \\
&\overset{(c)}{\le} \left(\mathbb{E} \sup_{r\in[0,1]} \left\| \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \eta_t \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1}) \right\|^{m+1}\right)^{\frac{1}{m+2}} \\
&\overset{(d)}{\le} \left(T^{\frac{m+1}{2}} \mathbb{E} \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t C_{U,\boldsymbol{x}_t}\right)^{1+m}\right)^{\frac{1}{m+2}} \\
&\overset{(e)}{\lesssim} \left(T^{\frac{m+1}{2}} t_{\text{mix}}^{m+1} \cdot \frac{1}{T} \sum_{t=0}^{T} \eta_t^{m+1} (\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^{m+1} + \sigma^{m+1})\right)^{\frac{1}{m+2}} \\
&\overset{(f)}{\lesssim} \mathcal{O}\left(T^{\frac{m+1}{m+2}} \cdot t_{\text{mix}}^{\frac{m+1}{m+2}} \cdot \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t^{m+1}\right)^{\frac{1}{m+2}}\right),
\end{aligned}$$

where (a) uses Proposition 3.4.2, (b) follows from $\tilde{\boldsymbol{\phi}}_T(r) = \boldsymbol{\phi}_T(r) - \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \eta_t \mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})$ and Proposition 3.4.1, (c) follows from Proposition 3.4.3, (d) uses the fact that $\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}_t, \xi_{t-1})\| \le C_{U,\boldsymbol{x}_t}$ from Lemma 3.2.2, (e) uses Jensen's inequality, and (f) uses Assumption 3.2.6 .

For $\boldsymbol{\psi}_0$ From Lemma B.7.1, we know that $\boldsymbol{A}_j^n$ is uniformly bounded. Hence, as $T \to \infty$,

$$|||\boldsymbol{\psi}_0||| = \sup_{r\in[0,1]} \|\boldsymbol{\psi}_0(r)\| = \frac{1}{\sqrt{T}\eta_0} \sup_{r\in[0,1]} \|\boldsymbol{A}_0^{\lfloor Tr \rfloor} \boldsymbol{B}_0 \boldsymbol{\Delta}_0\| = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Because $\boldsymbol{\psi}_0(\cdot)$ is a deterministic process (given $\mathscr{F}_0$), it's easy to show that $\tilde{d}(\boldsymbol{\psi}_0) = \mathcal{O}(T^{-1/2})$ (by letting $p \to \infty$ in Proposition 3.4.3).

For $\boldsymbol{\psi}_{1,1}$ Notice that $\boldsymbol{\psi}_{1,1}(r) = \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{A}_t^{\lfloor Tr \rfloor} \boldsymbol{r}_t$. From the proof of Lemma 3.4.1, we know there exists $c_r := \max \left\{ L_G, \frac{L_H + \|\boldsymbol{G}\|}{\delta_G} \right\}$ such that

$$\begin{aligned} \|\boldsymbol{r}_t\| &\le c_r \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \eta_t C_{U,\boldsymbol{x}_t} \\ &\le (c_r + \kappa t_{\mathrm{mix}}) \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \eta_t \kappa t_{\mathrm{mix}} (\sigma + L_H^2) \\ &:= \tilde{c}_r \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \eta_t C_U. \end{aligned}$$

Lemma B.7.1 implies that $\boldsymbol{A}_j^n$ is uniformly bounded by a universal constant C_0 in the sense that $\|\boldsymbol{A}_j^n\| \le C_0$ for all $j \le n$. Let λ denote any positive number satisfying $0 \le \lambda \le \delta$. By the $(L^p, (1 + \log t)\sqrt{\eta_t})$-consistency assumption with $p = 2 + 2\delta$, we upper bound the $(1 + \lambda)$-th moment of $\|\boldsymbol{\psi}_1\|_\infty$ by Jensen's inequality as following

$$\begin{aligned} \mathbb{E} \left\| \left| \boldsymbol{\psi}_{1,1} \right| \right\|^{1+\lambda} &= \mathbb{E} \sup_{0 \le r \le 1} \|\boldsymbol{\psi}_{1,1}(r)\|^{1+\lambda} = \mathbb{E} \sup_{0 \le r \le 1} \left\| \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{A}_t^{\lfloor Tr \rfloor} \boldsymbol{r}_t \right\|^{1+\lambda} \\ &\le \mathbb{E} \left(\frac{1}{\sqrt{T}} \sum_{t=0}^{T} C_0 \|\boldsymbol{r}_t\| \right)^{1+\lambda} \le T^{\frac{1+\lambda}{2}} C_0^{1+\lambda} \mathbb{E} \left(\frac{1}{T} \sum_{t=0}^{T} \|\boldsymbol{r}_t\| \right)^{1+\lambda} \\ &\le T^{\frac{1+\lambda}{2}} C_0^{1+\lambda} 2^\lambda \left[\tilde{c}_r^{1+\lambda} \mathbb{E} \left(\frac{1}{T} \sum_{t=0}^{T} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \right)^{1+\lambda} + C_U^{1+\lambda} \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t \right)^{1+\lambda} \right] \\ &\le T^{\frac{1+\lambda}{2}} C_0^{1+\lambda} 2^\lambda \left[\tilde{c}_r^{1+\lambda} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E} \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^{2(1+\lambda)} + C_U^{1+\lambda} \frac{1}{T} \sum_{t=0}^{T} \eta_t^{1+\lambda} \right] \\ &\le T^{\frac{1+\lambda}{2}} C_0^{1+\lambda} 2^\lambda \left[\tilde{c}_r^{1+\lambda} \frac{1}{T} \sum_{t=0}^{T} \eta_t^{1+\lambda} C_p \log^p T + C_U^{1+\lambda} \frac{1}{T} \sum_{t=0}^{T} \eta_t^{1+\lambda} \right]. \end{aligned}$$

As a result of Proposition 3.4.3 with $p = 1 + \lambda$, we get

$$\begin{aligned} \tilde{d}(\boldsymbol{\psi}_{1,1}) \le (\mathbb{E} \left\| \left| \boldsymbol{\psi}_{1,1} \right| \right\|^{1+\lambda})^{\frac{1}{2+\lambda}} &= \tilde{\mathcal{O}} \left(\left(\tilde{c}_r + C_U \right)^{\frac{1+\lambda}{2+\lambda}} \cdot T^{\frac{1+\lambda}{2(2+\lambda)}} \cdot \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t^{1+\lambda} \right)^{\frac{1}{2+\lambda}} \right) \\ &= \tilde{\mathcal{O}} \left(\left(c_r + t_{\mathrm{mix}} \right)^{\frac{1+\lambda}{2+\lambda}} \cdot T^{\frac{1+\lambda}{2(2+\lambda)}} \cdot \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t^{1+\lambda} \right)^{\frac{1}{2+\lambda}} \right). \end{aligned}$$

For $\psi_{1,2}$ Recall that $\psi_{1,2}(r) = \frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor} A_t^{\lfloor Tr\rfloor} v_t$. Note that $\frac{\eta_t-\eta_{t+1}}{\eta_t} = 1-\left(1-\frac{1}{t+1}\right)^\alpha \le 1-\exp(-\frac{\alpha}{t}) \le \frac{\alpha}{t} \le \alpha\eta_t$. By the definition of v_t in (3.18) and the inequality (B.2), we have

$$\begin{aligned}\|v_t\| &\le L_U\|x_{t+1}-x_t\| + \left|\frac{\eta_{t+1}-\eta_t}{\eta_t}\right| \cdot C_{U,x_{t+1}} \\ &\le L_U\|x_{t+1}-x_t\| + \alpha\kappa t_{\text{mix}}\eta_t \cdot (2L_H\|x_{t+1}-x^\star\| + \sigma) \\ &\le (L_U+\alpha\kappa t_{\text{mix}})\left(\|x_{t+1}-x_t\| + \eta_t\|x_{t+1}-x^\star\|\right) + \eta_t \cdot \alpha\kappa\sigma t_{\text{mix}} \end{aligned} \tag{D.5}$$

where $L_U = \mathcal{O}(L_H(1+\kappa t_{\text{mix}}))$ is given in Lemma 3.2.2. Let $0 \le m \le 2\delta+1$ be any real number.

We assert that there exists a positive constant $C_{1,2} > 0$ so that

$$\mathbb{E}\left(\|x_{t+1}-x_t\| + \eta_t\|x_{t+1}-x^\star\|\right)^{1+m} \lesssim C_{1,2}^{1+m}\eta_t^{1+m}. \tag{D.6}$$

We prove this statement in the following. First, from Assumption 3.3.3, $x_t - x^\star$ has uniformly bounded p-th order moments and thus $\sup_{t\ge0}(\mathbb{E}\|x_t-x^\star\|^p)^{\frac{1}{p}} < \infty$. Second, it follows that $\mathbb{E}\|x_{t+1}-x_t\|^{1+m} = \eta_t^{1+m}\mathbb{E}\|H(x_t,\xi_t)\|^{1+m}$ and

$$\begin{aligned}\mathbb{E}\|H(x_t,\xi_t)\|^{1+m} &\le 2^m\left[\mathbb{E}\|H(x_t,\xi_t)-H(x^\star,\xi_t)\|^{1+m} + \mathbb{E}\|H(x^\star,\xi_t)\|^{1+m}\right] \\ &\overset{(a)}{\le} 2^m\left[L_H^{1+m}\mathbb{E}\|x_t-x^\star\|^{1+m} + \mathbb{E}\|H(x^\star,\xi_t)\|^{1+m}\right] \\ &\overset{(b)}{\le} 2^m\left[L_H^{1+m}\left(\sup_{t\ge0}\mathbb{E}\|x_t-x^\star\|^p\right)^{\frac{1+m}{p}} + \sup_{t\ge0}\left(\mathbb{E}\|H(x^\star,\xi_t)\|^p\right)^{\frac{1+m}{p}}\right].\end{aligned}$$

where (a) uses Assumption 3.2.3, (b) uses Jensen's inequality, and (c) uses Assumption 3.2.2. Combing the last two points, we know that (D.6) with $C_{1,2}$ depending on universal constants as well as $\sup_{t\ge0}\left(\mathbb{E}\|H(x^\star,\xi_t)\|^p\right)^{\frac{1}{p}}$, $\sup_{t\ge0}\left(\mathbb{E}\|x_t-x^\star\|^p\right)^{\frac{1}{p}}$, and L_H.

Therefore,

$$\begin{aligned}\mathbb{E}\|v_t\|^{1+m} &\le 2^{1+m}(L_U+\alpha\kappa t_{\text{mix}})^{1+m}C_{1,2}^{1+m}\eta_t^{1+m} + \left(2\alpha\kappa\sigma t_{\text{mix}}\cdot\eta_t\right)^{1+m} \\ &\le (2C_{1,2}(L_U+\alpha\kappa t_{\text{mix}}) + 2\alpha\kappa\sigma t_{\text{mix}})^{1+m}\eta_t^{1+m} \\ &\le (t_{\text{mix}}+1)^{1+m}C_{1,3}^{1+m}\eta_t^{1+m}\end{aligned}$$

where the first inequality follows from applying Jensen's inequality to (D.5) and plugging (D.6) into it and the last inequality enlarges $C_{1,2}$ to $C_{1,3}$ to simplify notation. Using the last inequality, we have

$$
\begin{aligned}
\mathbb{E}\left|\!\left|\!\left|\boldsymbol{\psi}_{1,2}\right|\!\right|\!\right|^{1+m} &= \mathbb{E}\sup_{0\le r\le 1}\|\boldsymbol{\psi}_{1,2}(r)\|^{1+m} = \mathbb{E}\sup_{0\le r\le 1}\left\|\frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor Tr\rfloor}\boldsymbol{A}_t^{\lfloor Tr\rfloor}\boldsymbol{v}_t\right\|^{1+m}\\
&\le \mathbb{E}\left(\frac{1}{\sqrt{T}}\sum_{t=0}^{T}C_0\|\boldsymbol{v}_t\|\right)^{1+m}\\
&\le T^{\frac{1+m}{2}}C_0^{1+m}\mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T}\|\boldsymbol{v}_t\|\right)^{1+m}\\
&\le T^{\frac{1+m}{2}}C_0^{1+m}\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{v}_t\|^{1+m}\\
&\lesssim T^{\frac{1+m}{2}}(1+t_{\mathrm{mix}})^{1+m}\cdot\frac{1}{T}\sum_{t=0}^{T}\eta_t^{1+m}.
\end{aligned}
$$

As a result of Proposition 3.4.3 with $p = 1 + m$, we get

$$
\tilde{d}(\boldsymbol{\psi}_{1,2}) \le (\mathbb{E}\|\boldsymbol{\psi}_{1,2}\|_\infty^{1+m})^{\frac{1}{2+m}} = \tilde{\mathcal{O}}\left(\left(1+t_{\mathrm{mix}}\right)^{\frac{1+m}{2+m}}\cdot T^{\frac{1+m}{2(2+m)}}\cdot\left(\frac{1}{T}\sum_{t=0}^{T}\eta_t^{1+m}\right)^{\frac{1}{2+m}}\right).
$$

For $\boldsymbol{\psi}_2$ Notice that $\sum t = 0^k(\boldsymbol{A}_t^T - \boldsymbol{G}^{-1})\boldsymbol{u}_t$ is a martingale with the natural filtration $\mathscr{F}_k$. By Doob's inequality,

$$
\begin{aligned}
\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_2(r)\|_2^2 &\le \frac{4}{T}\sum_{t=0}^{T}\mathbb{E}\|(\boldsymbol{A}_t^T-\boldsymbol{G}^{-1})\boldsymbol{u}_t\|_2^2\\
&\le 4\sup_{t\ge 0}\mathbb{E}\|\boldsymbol{u}_t\|_2^2\cdot\frac{1}{T}\sum_{t=0}^{T}\|\boldsymbol{A}_t^T-\boldsymbol{G}^{-1}\|_2^2.
\end{aligned}
$$

We then need to analyze the order of $\|\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\|_2$. To that end, we introduce another quantity

$$
\boldsymbol{D}_t^n := \sum_{j=t}^{n}\eta_{j+1}\prod_{i=t}^{j}(\boldsymbol{I}-\eta_i\boldsymbol{G}). \tag{D.7}
$$

Lemma D.2.1. *There exists two constants $c, c_0 > 0$ such that for all $n \ge t \ge 0$,*

$$
\|\boldsymbol{D}_t^n - \boldsymbol{G}^{-1}\|_2 \le \eta_t + c_0\exp\left\{-c\sum_{i=t}^{n+1}\eta_i\right\}\|\boldsymbol{G}^{-1}\|_2 \quad \text{and} \quad \left\|\boldsymbol{A}_{t-1}^n - \boldsymbol{D}_t^n\right\|_2 = \mathcal{O}(t^{\alpha-1})
$$

where we hide dependence α and other universal constant factors.

The proof of Lemma D.2.1 is provided in Section D.3. By Lemma D.2.1 and triangular inequality, we have

$$\|\boldsymbol{A}_t^n - \boldsymbol{G}^{-1}\|_2 = \mathcal{O}\left(t^{\alpha-1} + \exp\left\{-c\sum_{i=t}^{n+1}\eta_i\right\}\right)$$

Therefore,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T}\|\boldsymbol{A}_t^T - \boldsymbol{G}^{-1}\|_2^2 &= \frac{\mathcal{O}(1)}{T}\sum_{t=0}^{T}\left(t^{2\alpha-2} + \exp\left\{-2c\sum_{i=t}^{T+1}\eta_i\right\}\right)\\ &= \mathcal{O}\left(T^{2\alpha-2} + \frac{1}{T\eta_T}\cdot\eta_T\sum_{t=1}^{T}\exp\left\{-2c\sum_{i=t}^{T}\eta_i\right\}\right)\\ &= \mathcal{O}\left(T^{2\alpha-2} + \frac{1}{T\eta_T}\right) = \mathcal{O}(T^{\alpha-1}),\end{aligned}$$

where the last equation holds by the fact $\eta_T\sum_{t=1}^T\exp\left\{-2c\sum_{i=t+1}^T\eta_i\right\} \to 1$ as $T \to \infty$ and thus is uniformly bounded. One can prove it by using Stolz–Cesàro theorem. Thus, by setting $p = 2$ in Proposition 3.4.3, we know that

$$\tilde{d}(\boldsymbol{\psi}_2) \le (\mathbb{E}|||\boldsymbol{\psi}_2|||^2)^{1/3} \precsim (\mathbb{E}\sup_{r\in[0,1]}\|\boldsymbol{\psi}_2(r)\|_2^2)^{1/3} = \mathcal{O}\left(T^{\frac{\alpha-1}{3}}\right).$$

For $\boldsymbol{\psi}_3$ In our previous asymptotic result, we establish Lemma B.3.1 to analyze the term $\boldsymbol{\psi}_3$. In order to provide an quantitative result, we need to capture the exact convergence rate in Lemma B.3.1, which is equivalent to analyze the moments of decomposed errors therein. Thanks to our technique developed therein, it is possible to do that. In the following, we will use the same notation in the proof of Lemma B.3.1 for the sake of consistency and quick understanding.

Lemma D.2.2. *Assume the same assumptions in Lemma 3.4.3 and let* $\{\boldsymbol{y}_t\}_{t\ge0}$ *be defined in* (3.24)*. If we set* $\eta_t = t^{-\alpha}$*, then for any* $0 \le l \le \delta$ *where* $p = 2 + 2\delta$*, we have*

$$\tilde{d}\left(\bar{\boldsymbol{y}}_T\right) = \mathcal{O}\left((1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}}\right) \quad \textit{where} \quad \bar{\boldsymbol{y}}_T(r) = \frac{\boldsymbol{y}_{\lfloor(T+1)r\rfloor}}{\sqrt{T}\eta_{\lfloor(T+1)r\rfloor}} \quad \textit{for} \quad r\in[0,1]. \tag{3.26}$$

The proof of Lemma D.2.2 can be found in Section D.4. With Lemma D.2.2, we are ready to bound $\tilde{d}(\boldsymbol{\psi}_3)$. By (3.22), we have

$$|||\boldsymbol{\psi}_3||| = \sup_{r\in[0,1]}\|\boldsymbol{\psi}_3(r)\| \precsim \sup_{n\in[T]}\left\|\frac{1}{\sqrt{T}}\frac{1}{\eta_{n+1}}\sum_{t=0}^{n}\left(\prod_{i=t+1}^{n}\boldsymbol{B}_i\right)\eta_t\boldsymbol{u}_t\right\| := \sup_{r\in[0,1]}\|\bar{\boldsymbol{y}}_T(r)\| = |||\bar{\boldsymbol{y}}_T|||,$$

where we define $\bar{\boldsymbol{y}}_T(r) = \frac{\boldsymbol{y}_{\lfloor (T+1)r \rfloor}}{\sqrt{T}\eta_{\lfloor (T+1)r \rfloor}}$ and $\boldsymbol{y}_{n+1} = \sum_{t=0}^{n} \left(\prod_{i=t+1}^{n} \boldsymbol{B}_i \right) \eta_t \boldsymbol{u}_t$ in the last inequality. By the definition of $\tilde{d}(\cdot)$ and Lemma D.2.2,

$$\tilde{d}(\boldsymbol{\psi}_3) = \inf_{\varepsilon} \varepsilon \vee \mathbb{P}(|||\boldsymbol{\psi}_3||| > \varepsilon) \lesssim \inf_{\varepsilon} \varepsilon \vee \mathbb{P}(|||\bar{\boldsymbol{y}}_T||| > \varepsilon) = \tilde{d}(\bar{\boldsymbol{y}}_T) = \tilde{\mathcal{O}}\left((1+l) \cdot T^{-\frac{l(1-\alpha)}{3+2l}} \right).$$

For $\boldsymbol{\psi}_{4,1}$ Let k by any real number satisfying $1 \le k \le 2 + 2\delta$. By Burkholder-Davis-Gundy inequality, it follows that

$$\begin{aligned}
\mathbb{E} \sup_{0 \le t \le T} \left\| \frac{1}{\sqrt{T}} \sum_{j=0}^{t} \boldsymbol{u}_{j,1} \right\|^k &\le \frac{(c_3 k)^{k/2}}{T^{k/2}} \mathbb{E} \left(\sum_{t=1}^{T} \mathbb{E} \left[\boldsymbol{u}_{t,1}^2 | \mathscr{F}_{t-1} \right] \right)^{k/2} \\
&\overset{(a)}{\lesssim} \frac{(c_3 k)^{k/2}}{T^{k/2}} \mathbb{E} \left(\sum_{t=1}^{T} \| \boldsymbol{x}_t - \boldsymbol{x}^\star \|^2 \right)^{k/2} \\
&\overset{(b)}{\lesssim} \frac{(c_3 k)^{k/2}}{T} \mathbb{E} \sum_{t=1}^{T} \| \boldsymbol{x}_t^\star - \boldsymbol{x}^\star \|^k \\
&\overset{(c)}{\lesssim} \frac{(c_3 k)^{k/2} C_k^k}{T} \sum_{t=1}^{T} \eta_t^{k/2} \log^k T,
\end{aligned}$$

where (a) holds because $\mathbb{E}\left[\boldsymbol{u}_{t,1}^2 | \mathscr{F}_{t-1} \right] \lesssim \mathbb{E}\| \boldsymbol{x}_t - \boldsymbol{x}^\star \|^2$ as a result of Assumption 3.2.3, (b) follows from Jensen's inequality, (c) holds owing to the $(L^p, (1+\log t)\sqrt{\eta_t})$-consistency that implies $\mathbb{E}\| \boldsymbol{x}_t^\star - \boldsymbol{x}^\star \|^k \le C_k^k \eta_t^{k/2} \log^k T$. The last inequality together with Proposition 3.4.3 implies that

$$\tilde{d}(\boldsymbol{\psi}_{4,1}) \le \left(\mathbb{E} \sup_{0 \le t \le T} \left\| \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{G}^{-1} \boldsymbol{u}_{t,1} \right\|^k \right)^{\frac{1}{k+1}} = \tilde{\mathcal{O}} \left(\sqrt{k} C_k \cdot \left(\frac{1}{T} \sum_{t=0}^{T} \eta_t^{k/2} \right)^{\frac{1}{1+k}} \right).$$

For $\boldsymbol{\psi}_{4,2}$ Recall that $\boldsymbol{\psi}_{4,2}(r) := \frac{1}{\sqrt{T}} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \boldsymbol{u}_{t,2}$ with $\boldsymbol{u}_{t,2} = \boldsymbol{U}(\boldsymbol{x}^\star, \xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi_{t-1})$ and $\boldsymbol{\psi}(r) := \boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \boldsymbol{S}^{1/2} \boldsymbol{W}(r)$. We will apply Theorem 3.4.1 to bound the Lévy-Prokhorov distance between them. Since Theorem 3.4.1 holds only for $0 < \delta < 3/2$, we denote $\delta' = \delta \wedge (\frac{3}{2} - o(1))$ for very sufficiently small $o(1)$.[1]

First, the quadratic variation process of $\boldsymbol{\psi}_{4,2}$ is given by

$$\left\langle \boldsymbol{\psi}_{4,2}(\cdot) \right\rangle_r = \frac{1}{T} \sum_{t=0}^{\lfloor Tr \rfloor} \mathbb{E}[(\boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \boldsymbol{u}_{t,2})^2 | \mathscr{F}_{t-1}] = \frac{1}{T} \sum_{t=0}^{\lfloor Tr \rfloor} \boldsymbol{\theta}^\top \boldsymbol{G}^{-1} \mathbb{E}[\boldsymbol{u}_{t,2} \boldsymbol{u}_{t,2}^\top | \mathscr{F}_{t-1}] \boldsymbol{G}^{-\top} \boldsymbol{\theta}. \tag{D.8}$$

[1] We can always set the term $o(1)$ as small as expected, which is the reason we denote it by an infinitesimal $o(1)$.

Second, note that the partial-sum process $\boldsymbol{\psi}_{4,2}$ is càdlàg (that is right continuous with left limits) with all discontinuous points given by $\{t/T\}_{t\in[T]}$. Hence, its corresponding dual predictable projection is the point measure in $[0,1]\times\mathbb{R}$ (similar to the definition of the Poisson point process) and thus we can compute the following integral and obtain

$$\mathbb{E}\int_0^1\int_{\mathbb{R}}|x|^{2+2\delta'}\Pi^n(ds,dx)=\mathbb{E}\sum_{t=0}^{T}\frac{1}{T^{1+\delta'}}\left|\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2}\right|^{2+2\delta'}\le C_{4,2}^{2+2\delta'}T^{-\delta'},\tag{D.9}$$

where $C_{4,2}=\|\boldsymbol{G}^{-1}\|\cdot\sup_{t\ge0}(\mathbb{E}\|\boldsymbol{u}_{t,2}\|^p)^{\frac{1}{p}}<\infty$ due to Assumption 3.2.2.

By Theorem 3.4.1, it follows that

$$d_{\mathrm{P}}(\boldsymbol{\theta}^\top\boldsymbol{\psi}_{4,2},\boldsymbol{\theta}^\top\boldsymbol{\psi})=\tilde{\mathcal{O}}\left(T^{-\frac{\delta'}{3+2\delta'}}+\left[\mathbb{E}\sup_{r\in[0,1]}\left|\left\langle\boldsymbol{\psi}_{4,2}(\cdot)\right\rangle_r-\langle\boldsymbol{\psi}(\cdot)\rangle_r\right|\right]^{\frac{1}{3}}\right).\tag{D.10}$$

The second term in (D.10) is the expected supreme absolute difference between the quadratic variation processes of $\boldsymbol{\psi}_{4,2}$ and $\boldsymbol{\psi}$ over the fraction $r\in[0,1]$. We analyze that term in Lemma D.2.3 whose proof is deferred in Section D.5.

Lemma D.2.3. *Rewriting $p=2+2\delta$ with p given in Assumption 3.2.2. For simplicity, we denote by*

$$q(\xi_{t-1}):=\mathbb{E}\left[(\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2})^2|\mathcal{F}_{t-1}\right]\quad\textit{with}\quad\boldsymbol{u}_{t,2}=\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)-\mathcal{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1}).$$

Under Assumption 3.3.3, $\mathbb{E}q(\xi_t)=\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-T}\boldsymbol{\theta}$ for all $t\ge0$. Then

$$\mathbb{E}\sup_{t\le T}\left|\sum_{i=0}^{t}\{q(\xi_i)-\mathbb{E}_{\xi_i}q(\xi_i)\}\right|=\mathcal{O}\left(\sqrt{Tt_{\mathrm{mix}}}\right)$$

where $\mathcal{O}(\cdot)$ hides factors depending on $\|\boldsymbol{G}^{-1}\|, C_U, \kappa$ and $\sup_{\xi\in\Xi}\mathcal{P}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\|^2$.

Using the notation in Lemma D.2.3, we denote $q(\xi_{t-1}):=\mathbb{E}\left[(\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2})^2|\mathcal{F}_{t-1}\right]$. Then the quadratic variation of $\boldsymbol{\psi}_{4,2}$ can be expressed in terms of $q(\xi_{t-1})$'s as what follows

$$\left\langle\boldsymbol{\psi}_{4,2}(\cdot)\right\rangle_r=\left\langle\frac{1}{\sqrt{T}}\sum_{t=0}^{\lfloor T\cdot\rfloor}\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2}\right\rangle_r=\frac{1}{T}\sum_{t=1}^{\lfloor Tr\rfloor}q(\xi_{t-1}).$$

By Lemma D.2.3, we have $\mathbb{E}q(\xi_{t-1})=\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-T}\boldsymbol{\theta}$ for all $t\ge1$. Therefore,

$$\mathbb{E}\sup_{r\in[0,1]}\left|\left\langle\boldsymbol{\psi}_{4,2}(\cdot)\right\rangle_r-\langle\boldsymbol{\psi}(\cdot)\rangle_r\right|\le\mathbb{E}\sup_{n\in[T]}\frac{1}{T}\left|\sum_{t=1}^{n}\left\{q(\xi_{t-1})-\mathbb{E}q(\xi_{t-1})\right\}\right|$$

$$+ \sup_{r\in[0,1]} \left| \left(\frac{\lfloor Tr \rfloor}{T} - r \right) \boldsymbol{\theta}^{\top} \boldsymbol{G}^{-1} \boldsymbol{S} \boldsymbol{G}^{-T} \boldsymbol{\theta} \right|$$
$$= \mathcal{O}\left(\sqrt{\frac{t_{\mathrm{mix}}}{T}} \right).$$

□

D.3 Proof of Lemma D.2.1

Proof of Lemma D.2.1. For the first part, it follows that

$$\begin{aligned}
\boldsymbol{G}\boldsymbol{D}_t^n + \prod_{i=t}^{n+1}(\boldsymbol{I} - \eta_i \boldsymbol{G}) &= \sum_{j=t}^{n} \eta_{j+1} \prod_{i=t}^{j}(\boldsymbol{I} - \eta_i \boldsymbol{G}) + \prod_{i=t}^{n+1}(\boldsymbol{I} - \eta_j \boldsymbol{G}) \\
&= \sum_{j=t}^{n-1} \eta_{j+1} \prod_{i=t}^{j}(\boldsymbol{I} - \eta_i \boldsymbol{G}) + (\boldsymbol{I} - \eta_{n+1}\boldsymbol{G} + \eta_{n+1}\boldsymbol{G}) \prod_{i=t}^{n}(\boldsymbol{I} - \eta_j \boldsymbol{G}) \\
&= \sum_{j=t}^{n-1} \eta_{j+1} \prod_{i=t}^{j}(\boldsymbol{I} - \eta_i \boldsymbol{G}) + \prod_{i=t}^{n}(\boldsymbol{I} - \eta_j \boldsymbol{G}) \\
&= \boldsymbol{G}\boldsymbol{D}_t^{n-1} + \prod_{i=t}^{n}(\boldsymbol{I} - \eta_j \boldsymbol{G}) \\
&= \eta_{t+1}\boldsymbol{G}(\boldsymbol{I} - \eta_{t+1}\boldsymbol{G}) + (\boldsymbol{I} - \eta_t \boldsymbol{G})(\boldsymbol{I} - \eta_{t+1}\boldsymbol{G}) \\
&= \boldsymbol{I} - \eta_t \boldsymbol{G}.
\end{aligned}$$

Rearranging the last equation gives

$$\boldsymbol{D}_t^n - \boldsymbol{G}^{-1} = -\eta_t \boldsymbol{I} - \boldsymbol{G}^{-1} \prod_{i=t}^{n+1}(\boldsymbol{I} - \eta_i \boldsymbol{G}).$$

It follows from 1 in Lemma B.7.1 that there exist two constant $c_0,\ c > 0$ so that $\left\| \prod_{i=t}^{n+1} (\boldsymbol{I} - \eta_i \boldsymbol{G}) \right\|_2 \leq c_0 \exp\left\{ -c \sum_{i=t}^{n+1} \eta_i \right\}$ for all $n \geq t > 0$. We then complete the proof by triangular inequality.

For the second part, we bound the difference between $\boldsymbol{A}_{t-1}^n$ and $\boldsymbol{D}_t^n$ as following

$$\begin{aligned}
\left\| \boldsymbol{A}_{t-1}^n - \boldsymbol{D}_t^n \right\|_2 &= \left\| \eta_{t-1}\boldsymbol{I} + \sum_{j=t}^{n} (\eta_{t-1} - \eta_{j+1}) \prod_{i=t}^{j}(\boldsymbol{I} - \eta_i \boldsymbol{G}) \right\|_2 \\
&\leq \eta_{t-1} + \sum_{j=t}^{n} \sum_{i=t}^{j+1} (\eta_{i-1} - \eta_i) \left\| \prod_{i=t}^{j}(\boldsymbol{I} - \eta_i \boldsymbol{G}) \right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\le} \eta_{t-1} + \frac{2\alpha}{t} \sum_{j=t}^{n} \sum_{i=t}^{j} \eta_i \left\| \prod_{i=t}^{j} (\boldsymbol{I} - \eta_i \boldsymbol{G}) \right\|_2 \\
&\overset{(b)}{\le} \eta_{t-1} + \frac{2c_0\alpha}{t^{1-\alpha}} \sum_{j=t}^{n} \eta_{j+1} \sum_{i=t}^{j} \eta_j \exp\left\{ -c \sum_{i=t}^{j} \eta_i \right\} \\
&\overset{(c)}{=} \mathscr{O}\left(\eta_{t-1} + t^{\alpha-1} \right) = \mathscr{O}(t^{\alpha-1})
\end{aligned}
$$

where (a) uses $\eta_{i-1} - \eta_i \le \frac{\alpha}{i-1} \cdot \eta_{i-1} \le 2\alpha\eta_{i-1}\frac{1}{t}$ for $\eta_i = i^{-\alpha}$ and $i \ge t$, (b) uses $\left\| \prod_{i=t}^{j} (\boldsymbol{I} - \eta_i \boldsymbol{G}) \right\|_2 \le c_0 \exp\left\{ -c \sum_{i=t}^{j} \eta_i \right\}$ and $\eta_{j+1} \ge \eta_t$ for $j+1 \ge t$, and (c) uses $\sum_{j=t}^{n} \eta_{j+1} \sum_{i=t}^{j} \eta_j \exp\left\{ -c \sum_{i=t}^{j} \eta_i \right\} \precsim \int_0^\infty m \exp(-cm) dm < \infty$.

Finally, we comment that here we use polynomial step sizes that $t^{-\alpha}$ with $\frac{1}{2} < \alpha < 1$ for simplicity. It is possible to extend to general step sizes using a similar but more complicate argument. □

D.4 Proof of Lemma D.2.2

Proof of Lemma D.2.2. The proof can be viewed as a quantitative version of that of Lemma B.3.1. We suggest readers should be familiar with the notation and proof idea therein before diving into the details of this proof. At the beginning, we choose any $p_0 \in [2, p]$.

We first assume $\boldsymbol{G}$ is further diagonalizable. Recall the definition of $\mathscr{A}^c$ in (B.8). Similar to (B.9), one can show that $\mathbb{E}\|\boldsymbol{y}_{h_k}\|^{p_0} \le {p_0}^{p_0} c_4^{p_0} \cdot \eta_{h,k}^{\frac{p_0}{2}}$. By Markov's inequality, it follows that

$$
\begin{aligned}
\mathbb{P}(\mathscr{A}^c) &\le \sum_{k=0}^{n} \mathbb{P}\left(\frac{c_0'}{\sqrt{T}} \left\| \frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}} \right\| \ge \varepsilon \right) \le \frac{(c_0')^{p_0}}{T^{p_0/2}\varepsilon^{p_0}} \sum_{k=0}^{n} \mathbb{E} \left\| \frac{\boldsymbol{y}_{h_k}}{\eta_{h_k}} \right\|^{p_0} \\
&\le \frac{1}{T^{p_0/2}\varepsilon^{p_0}} \sum_{k=0}^{n} \frac{{p_0}^{p_0} C_{3,1}^{p_0}}{\eta_{h_k}^{p_0/2}} \le \frac{{p_0}^{p_0} C_{3,1}^{p_0} n}{(T\eta_T)^{p_0/2}\varepsilon^{p_0}}.
\end{aligned}
$$

where $C_{3,1} := c_0' \cdot c_4$ for short. On the other hand, by Lemma B.4.1, we know that for any $k \in [n]$ or $k = 0$,

$$
\mathscr{P}_k \le \frac{{p_0}^{p_0} C_3^{p_0}}{n^{p_0/2}\varepsilon^{p_0}} \le \frac{{p_0}^{p_0} C_{3,2}^{p_0}}{n^{p_0/2}\varepsilon^{p_0}}, \tag{D.11}
$$

where $C_{3,2} := \max\{C_{3,1}, C_3\}$ and C_3 is defined in Lemma B.4.1. We comment that though the bound in Lemma B.4.1 depends on p rather than p_0, one can repeat the proof therein to derive the inequality (D.11). A shortcut argument can be used is to assume the noise defined therein has p_0-th order moments rather than p-th ones. Then (D.11) directly follows by replacing p with p_0 in Lemma B.4.1.

Putting these bounds together, we have that for any $\varepsilon > 0$,

$$\begin{aligned}
\tilde{d}\left(\bar{\boldsymbol{y}}_T\right) &\le \mathbb{P}\left(\sup_{0\le t\le T} \frac{\|\boldsymbol{y}_t\|}{\sqrt{T}\eta_t} > 2\varepsilon\right) \vee 2\varepsilon \\
&\le 2\varepsilon \vee \left\{\mathbb{P}\left(\sup_{0\le t\le T} \frac{\|\boldsymbol{y}_t\|}{\sqrt{T}\eta_t} > 2\varepsilon;\ \mathscr{A}\right) + \mathbb{P}(\mathscr{A}^c)\right\} \\
&\le 2\varepsilon \vee \left\{\sum_{k=0}^{n-1} \mathscr{P}_k + \mathbb{P}(\mathscr{A}^c)\right\} \\
&\lesssim \varepsilon \vee \left[\frac{n{p_0}^{p_0} C_{3,2}^{p_0}}{\varepsilon^{p_0}} \cdot \left(\frac{1}{n^{p_0/2}} + \frac{1}{(T\eta_T)^{p_0/2}}\right)\right].
\end{aligned}$$

Since the last inequality holds for any n and ε, we will carefully set n and ε to make the bound as small as possible. First, we set $n = T^{1-\alpha}$ so that $\frac{1}{n^{p_0/2}} = \frac{1}{(T\eta_T)^{p_0/2}}$ as a result of $\eta_T = T^{-\alpha}$. Therefore,

$$\tilde{d}\left(\bar{\boldsymbol{y}}_T\right) \lesssim \varepsilon \vee \frac{{p_0}^{p_0} C_{3,2}^{p_0}}{T^{(1-\alpha)(p_0/2-1)} \varepsilon^{p_0}}.$$

Then, we let $\varepsilon = p_0 C_{3,2} \cdot T^{-\frac{(p_0/2-1)(1-\alpha)}{1+p_0}}$ which ensures that $\varepsilon = \frac{C_{3,2}^{p_0+1} {p_0}^{p_0+1}}{T^{(1-\alpha)(p_0/2-1)} \varepsilon^{p_0}}$. As a result,

$$\tilde{d}\left(\bar{\boldsymbol{y}}_T\right) \lesssim p_0 C_{3,2} \cdot T^{-\frac{(p_0/2-1)(1-\alpha)}{1+p_0}} \lesssim (1+l) \cdot T^{-\frac{l(1-\alpha)}{3+2l}},$$

where the last inequality is because we rewrite $p_0 = 2(1+l)$ with $0 \le l \le \delta$ and $p = 2 + 2\delta$.

We then consider the case where $\boldsymbol{G}$ is not diagonalizable. The idea is similar to what we did in Section B.3. Let its Jordan decomposition be $\boldsymbol{G} = \boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^{-1} = \boldsymbol{V}\mathrm{diag}\{\boldsymbol{J}_1, \cdots, \boldsymbol{J}_r\}\boldsymbol{V}^{-1}$, where $\boldsymbol{V}$ is the non-singular matrix and $\{\boldsymbol{J}_i\}_{1\le i\le r}$ collects all Jordan blocks. Recall that $\{\boldsymbol{y}_t\}_{t\ge 0}$ is defined in (3.24). Let $\tilde{\boldsymbol{y}}_t = \boldsymbol{V}^{-1}\boldsymbol{y}_t$, $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{V}^{-1}\boldsymbol{\varepsilon}_t$ be transformed vectors. Then the recursion formula (3.24) becomes

$$\tilde{\boldsymbol{y}}_{t+1} = (\boldsymbol{I} - \eta_t \boldsymbol{J})\tilde{\boldsymbol{y}}_t + \eta_t \tilde{\boldsymbol{\varepsilon}}_t.$$

Let $\left(\tilde{\boldsymbol{y}}_t\right)_k$ denote the k-th coordinate of the vector $\tilde{\boldsymbol{y}}_t$ and so does $\left(\tilde{\boldsymbol{\varepsilon}}_t\right)_k$. The associated process is denoted by

$$(\bar{\tilde{\boldsymbol{y}}}_T)_k(r) = \frac{(\tilde{\boldsymbol{y}}_{\lfloor (T+1)r\rfloor})_k}{\sqrt{T}\eta_{\lfloor (T+1)r\rfloor}} \text{ for } r \in [0, 1].$$

Then it follows that

$$\left|\!\left|\!\left|\bar{\boldsymbol{y}}_T\right|\!\right|\!\right| = \sup_{0\le t\le T} \frac{\|\boldsymbol{y}_{t+1}\|}{\sqrt{T}\eta_{t+1}} \lesssim \sup_{0\le t\le T} \frac{\|\tilde{\boldsymbol{y}}_{t+1}\|}{\sqrt{T}\eta_{t+1}} \le \sum_{k=1}^{d} \sup_{0\le t\le T} \frac{|(\tilde{\boldsymbol{y}}_{t+1})_k|}{\sqrt{T}\eta_{t+1}} = \sum_{k=1}^{d} \left|\!\left|\!\left|(\bar{\tilde{\boldsymbol{y}}}_T)_k\right|\!\right|\!\right|,$$

which implies

$$\tilde{d}(\bar{\boldsymbol{y}}_T) = \varepsilon \vee \mathbb{P}(\left|\!\left|\!\left|\bar{\boldsymbol{y}}_T\right|\!\right|\!\right| \ge \varepsilon) \lesssim \varepsilon \vee \sum_{k=1}^{d} \mathbb{P}\left(\left|\!\left|\!\left|(\bar{\tilde{\boldsymbol{y}}}_T)_k\right|\!\right|\!\right| \ge \frac{\varepsilon}{d}\right) \le d \cdot \sum_{k=1}^{d} \tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_k).$$

In the following, we will focus on each coordinate supreme $\left|\!\left|\!\left|(\bar{\tilde{\boldsymbol{y}}}_T)_k\right|\!\right|\!\right|$. Without loss of generality, we assume $\boldsymbol{G}$ is a matrix of Jordan canonical form, that is, $\boldsymbol{J}$ consists of only one Jordan block (B.5).

Note that the last coordinate process evolves as $(\tilde{\boldsymbol{y}}_{t+1})_d = (1-\eta_t\lambda)(\tilde{\boldsymbol{y}}_t)_d + \eta_t(\tilde{\varepsilon}_t)_d$. By what has been established early in this subsection, we have $\tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_d) \lesssim (1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}}$. We are going to finish the proof by induction. Suppose that we already have $\tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_i) \lesssim (1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}}$ for the coordinates $i = k, k+1, \cdots, d$, we will show $\tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_{k-1})$ is also bounded by that quantity. Using the structure of $\boldsymbol{J}$ in (B.5), we have

$$(\tilde{\boldsymbol{y}}_{t+1})_{k-1} = (1-\lambda\eta_t)(\tilde{\boldsymbol{y}}_t)_{k-1} - \eta_t(\tilde{\boldsymbol{y}}_t)_k + \eta_t(\tilde{\varepsilon}_t)_{k-1}. \tag{B.6}$$

To facilitate analysis, we construct a surrogate sequence $\{(\hat{\boldsymbol{y}}_t)_{k-1}\}$ defined by $\hat{\boldsymbol{y}}_0 = \boldsymbol{0}$ and

$$(\hat{\boldsymbol{y}}_{t+1})_{k-1} = (1-\lambda\eta_t)(\hat{\boldsymbol{y}}_t)_{k-1} + \eta_t(\tilde{\varepsilon}_t)_{k-1}. \tag{B.7}$$

Again, we have

$$\tilde{d}((\bar{\hat{\boldsymbol{y}}}_T)_{k-1}) \lesssim (1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}} \text{ with } (\bar{\hat{\boldsymbol{y}}}_T)_k(r) = \frac{(\hat{\boldsymbol{y}}_{\lfloor (T+1)r\rfloor})_k}{\sqrt{T}\eta_{\lfloor (T+1)r\rfloor}} \text{ for } r \in [0,1].$$

Let $\tilde{\boldsymbol{\Delta}}_t := \frac{(\tilde{\boldsymbol{y}}_t)_{k-1} - (\hat{\boldsymbol{y}}_t)_{k-1}}{\eta_t}$ be their normalized difference. From (B.6) − (B.7), it follows that

$$\tilde{\boldsymbol{\Delta}}_{t+1} = \frac{(1-\lambda\eta_t)\eta_t}{\eta_{t+1}}\tilde{\boldsymbol{\Delta}}_t - \frac{\eta_t^2}{\eta_{t+1}} \cdot \frac{(\tilde{\boldsymbol{y}}_t)_k}{\eta_t}$$

There exists an t_0 such that $\left|\frac{(1-\lambda\eta_t)\eta_t}{\eta_{t+1}}\right| \le 1 - 0.5\lambda\eta_t$ for any $t \ge t_0$. In this case, for any $t \ge t_0$,

$$|\tilde{\boldsymbol{\Delta}}_{t+1}| \le (1-0.5\lambda\eta_t)\cdot|\tilde{\boldsymbol{\Delta}}_t| + 2\eta_t\left|\frac{(\tilde{\boldsymbol{y}}_t)_k}{\eta_t}\right|,$$

by which one can show the following inequality by induction

$$\sup_{t_0\le t\le T} |\tilde{\boldsymbol{\Delta}}_{t+1}| \le \frac{4}{\lambda}\cdot\max\left\{|\tilde{\boldsymbol{\Delta}}_{t_0}|, \sup_{t_0\le t\le T}\left|\frac{(\tilde{\boldsymbol{y}}_t)_k}{\eta_t}\right|\right\}.$$

One can also show that there exists a constant $C_{3,3} > 0$ depending on t_0, λ and $\{\eta_t\}_{0\le t\le t_0}$ such that

$$\sup_{0\le t\le t_0} |\tilde{\mathbf{\Delta}}_{t+1}| \le C_{3,3} \cdot \sup_{0\le t\le t_0} \left|\frac{(\tilde{\boldsymbol{y}}_t)_k}{\eta_t}\right|.$$

As a result, we know that

$$\frac{1}{\sqrt{T}} \sup_{0\le t\le T} |\tilde{\mathbf{\Delta}}_{t+1}| \lesssim \frac{1}{\sqrt{T}} \sup_{0\le t\le T} \left|\frac{(\tilde{\boldsymbol{y}}_t)_k}{\eta_t}\right|.$$

which implies

$$\tilde{d}(\bar{\tilde{\mathbf{\Delta}}}_T) \lesssim \tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_k) \lesssim (1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}} \text{ with } (\bar{\tilde{\mathbf{\Delta}}}_T)(r) = \frac{\tilde{\mathbf{\Delta}}_{\lfloor (T+1)r\rfloor}}{\sqrt{T}} \text{ for } r \in [0,1].$$

Finally, we complete the induction by noting

$$\tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_{k-1}) \le 2(\tilde{d}((\bar{\tilde{\boldsymbol{y}}}_T)_{k-1}) + \tilde{d}(\bar{\tilde{\mathbf{\Delta}}}_T)) \lesssim (1+l)\cdot T^{-\frac{l(1-\alpha)}{3+2l}}.$$

□

D.5 Proof of Lemma D.2.3

Proof of Lemma D.2.3. Let $\boldsymbol{M} = \sup_{\xi\in\Xi}\sqrt{\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\|^2}$. From Assumption 3.3.3, we have $\boldsymbol{M} < \infty$. By Lemma 3.2.2, it follows that

$$\boldsymbol{U}(\boldsymbol{x}^\star,\xi) = \boldsymbol{H}(\boldsymbol{x}^\star,\xi) + \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi) \text{ and } \sup_{\xi\in\Xi}\|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\| \le C_U.$$

Therefore, we have

$$\sup_{\xi\in\Xi}\mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\|^2 \le \sup_{\xi\in\Xi} 2\left[\mathscr{P}\|\boldsymbol{H}(\boldsymbol{x}^\star,\xi)\|^2 + \|\mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi)\|^2\right] \le 2(M^2 + C_U^2).$$

The last equation implies

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{u}_{t,2}\|^2|\mathscr{F}_{t-1}] &= \mathbb{E}[\|\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t) - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2|\mathscr{F}_{t-1}]\\ &\le \mathbb{E}[\|\boldsymbol{U}(\boldsymbol{x}^\star,\xi_t)\|^2|\mathscr{F}_{t-1}] = \mathscr{P}\|\boldsymbol{U}(\boldsymbol{x}^\star,\xi_{t-1})\|^2 \le 2(M^2 + C_U^2)\end{aligned}$$

is uniformly bounded. As a result,

$$q(\xi_{t-1}) := \mathbb{E}\left[(\boldsymbol{\theta}^\top\boldsymbol{G}^{-1}\boldsymbol{u}_{t,2})^2|\mathscr{F}_{t-1}\right] \le 2\|\boldsymbol{\theta}\|_*^2\|\boldsymbol{G}^{-1}\|^2(M^2 + C_U^2) \quad \text{(D.12)}$$

is uniformly bounded and thus has any l-th order moment where $l > 0$. For simplicity, we set $X_t := q(\xi_t)$. From (D.12), we know that the sequence $\{X_t - \mathbb{E}X_t\}_{t\ge0}$ has uniform bounded

$1 + l$-th order moments for any $l \geq 0$. We denote its centralized L_{1+l}-norm by

$$M_{1+l} := \sup_{t\geq 0}(\mathbb{E}|X_t - \mathbb{E}X_t|^{1+l})^{\frac{1}{1+l}} = \sup_{t\geq 0}(\mathbb{E}|q(\xi_t) - \mathbb{E}q(\xi_t)|^{1+l})^{\frac{1}{1+l}}.$$

On the other hand, since we assume $\xi_0 \sim \pi$, then $\xi_t \sim \pi$ and thus

$$\mathbb{E}q(\xi_t) = \mathbb{E}_{\xi\sim\pi,\xi'\sim P(\xi,\cdot)}[(\boldsymbol{\theta}^\top \boldsymbol{G}^{-1}(\boldsymbol{U}(\boldsymbol{x}^\star, \xi') - \mathscr{P}\boldsymbol{U}(\boldsymbol{x}^\star, \xi))^2] = \boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\boldsymbol{\theta}.$$

For simplicity, we denote $q^\star = \boldsymbol{\theta}^\top \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\boldsymbol{\theta}$. Our target quanitity is

$$\mathbb{E}\sup_{t\leq T}\left|\sum_{i=0}^{t}\{q(\xi_i) - \mathbb{E}q(\xi_i)\}\right| = \mathbb{E}\sup_{t\leq T}\left|\sum_{i=0}^{t}(X_i - \mathbb{E}X_i)\right|,$$

where the expectation $\mathbb{E}(\cdot)$ is taken with respect to all randomness. To that end, we will make use of moment inequalities for fast mixing random variables in Lemma D.5.1. Before starting the analysis, we first introduce additional notations and preliminaries. We denote a given one-dimensional random variable $X \in \mathbb{R}$ by $Q_X(\cdot)$ as the quantile function of $|X|$. It is the inverse of the function $x \to \mathbb{P}(|X| > x)$, defined by $Q_X(u) = \inf\{x : \mathbb{P}(|X| > x) \leq u\}$. We present a useful tail bound for $Q(u) := \sup_{t\geq 0} Q_{X_t - \mathbb{E}X_t}(u)$ which mainly follows from the Markov inequality.

$$Q(u) \leq \left(\frac{1}{u}\sup_{t\geq 0}\mathbb{E}|q(\xi_t) - \mathbb{E}q(\xi_t)|^{1+l}\right)^{\frac{1}{1+l}} = M_{1+l}u^{-\frac{1}{1+l}} \quad \text{for any } l \geq 0. \tag{D.13}$$

For a sequence of real numbers $\{\alpha_t\}_{t\geq 0}$, we define by the function $\alpha^{-1}(u)$ the counting function on the indexes t's on which α_t is larger then a given input u, that is, $\alpha^{-1}(u) := \sum_{t=0}^{\infty} 1_{u<\alpha_t}$.

Definition D.5.1 (α-mixing coefficients). *Given two σ-field $\mathscr{A}$ and $\mathscr{B}$, the strong mixing coefficient between them is defined by*

$$\alpha(\mathscr{A}, \mathscr{B}) := 2\sup\left\{\mathrm{Cov}(1_A, 1_B) : A \in \mathscr{A}, B \in \mathscr{B}\right\}$$

where 1_A is the indicator function of the event A and similar is 1_B.

Definition D.5.2 (Strong mixing coefficients[174]). *Let $\{X_t\}_{t>0}$ be a sequence of real-valued random variables. Set $\mathscr{F}_k^{\mathrm{l}} = \sigma(\{X_t\}_{t\leq k})$ and $\mathscr{F}_l^{\mathrm{u}} = \sigma(\{X_t\}_{t\geq l})$. The strong mixing coefficients of $\{X_t\}_{t>0}$ are denoted by $\{\alpha_t\}_{t\geq 0}$ with definition as what follows*

$$\alpha_0 = 1/2 \quad \textit{and} \quad \alpha_t = \sup_{k\in\mathbb{N}}\alpha(\mathscr{F}_k^{\mathrm{l}}, \mathscr{F}_{k+t}^{\mathrm{u}}) \quad \textit{for any} \quad t \geq 1.$$

Lemma D.5.1 (Theorem 6.3 in Rio [175]). *Let $\{X_t\}_{t>0}$ be a sequence of real-valued and cen-*

tered random variables and $\{\alpha_t\}_{t\geq 0}$ be the corresponding strong mixing coefficients. Suppose that, for some $p \geq 2$, $\sup_{t>0} \mathbb{E}|X_t|^p < \infty$. Then with $S_k = \sum_{t=1}^k X_t$, we have

$$\mathbb{E}\left(\sup_{1\leq k\leq n} |S_k|^p\right) \leq a_p \left(\sum_{i=1}^n \sum_{j=1}^n \left|\mathrm{Cov}\left(X_i, X_j\right)\right|\right)^{\frac{p}{2}} + nb_p \int_0^1 \left[\alpha^{-1}(u) \wedge n\right]^{p-1} Q^p(u)du,$$

where

$$Q := \sup_{t>0} Q_{X_t}, \quad a_p = p4^{p+1}(p+1)^{p/2} \quad \text{and} \quad b_p = \frac{p}{p-1}4^{p+1}(p+1)^{p-1}.$$

Lemma D.5.1 replies on the concept of strong mixing coefficients which we introduce in Definition D.5.2. By Lemma D.5.2, we know that the the strong mixing coefficients of $\{\xi_t\}_{t\geq 0}$ is mixing exponentially fast. With this result, we can compute the bounds in Lemma D.5.1 as follows.

Lemma D.5.2 (Fast mixing)**.** *Under Assumption 3.2.4, the strong mixing coefficients of $\{\xi_t\}_{t\geq 0}$ vanishes exponentially fast, that is, $\alpha_t \leq \kappa\rho^t$ with $\kappa > 0$ and $\rho \in [0, 1)$ given in Assumption 3.2.4.*

We first compute that for any $i < j$,

$$\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) \\
&\overset{(a)}{=} \mathbb{E}\left[(X_i - \mathbb{E}X_i)\mathbb{E}[X_j - \mathbb{E}X_j|\mathscr{F}_i]\right] \\
&\overset{(b)}{=} \mathbb{E}(X_i - \mathbb{E}X_i)(\mathscr{P}^{j-i}X_i - \mathbb{E}X_j) \\
&= \mathbb{E}[(X_i - \mathbb{E}X_i)(\mathscr{P}^{j-i}X_i - q^\star)] \\
&\overset{(c)}{\leq} \mathbb{E}|X_i - \mathbb{E}X_i| \cdot \kappa\rho^{j-i} \sup_{\xi} |q(\xi) - q^\star| \\
&\leq \kappa M_1 M_\infty \rho^{j-i},
\end{aligned}$$

where (a) uses the law of total expectation and the notation $\mathscr{F}_i = \sigma(\{\xi_t\}_{t\leq i})$, (b) uses the equality $\mathbb{E}[X_j|\mathscr{F}_i] = \mathscr{P}^{j-i}X_i$ due to the Markov property, and (c) follows from Lemma 3.2.1. Therefore,

$$\begin{aligned}
\sum_{i=1}^T \sum_{j=1}^T |\mathrm{Cov}(X_i, X_j)| &= \sum_{i=1}^T \mathbb{E}|X_i - \mathbb{E}X_i|^2 + 2\sum_{1\leq i<j\leq n} |\mathrm{Cov}(X_i, X_j)| \\
&\lesssim TM_2^2 + 2\rho\kappa M_1 M_\infty \sum_{1\leq i<j\leq T} \rho^{j-i}
\end{aligned}$$

$$
\begin{aligned}
&= TM_2^2 + 2\rho\kappa M_1 M_\infty \sum_{i=1}^{T-1}\sum_{k=0}^{T-i-1} \rho^k \\
&\le T\left(M_2^2 + \frac{2\rho\kappa M_1 M_\infty}{1-\rho}\right) \le T\left(M_2^2 + 2\kappa M_1 M_\infty t_{\text{mix}}\right). \qquad \text{(D.14)}
\end{aligned}
$$

Now, we apply Lemma D.5.1 with $p = 2$ and obtain

$$
\begin{aligned}
&\mathbb{E}\sup_{t\le T}\left|\sum_{i=0}^{t}\{X_i - \mathbb{E}X_i\}\right| \le \left\{\mathbb{E}\sup_{t\le T}\left|\sum_{i=0}^{t}\{X_i - \mathbb{E}X_i\}\right|^2\right\}^{\frac{1}{2}} \\
&\lesssim \left\{\sum_{i=1}^{T}\sum_{j=1}^{T}|\mathrm{Cov}(X_i, X_j)| + T\int_0^1 \alpha^{-1}(u)Q(u)^2 du\right\}^{\frac{1}{2}} \\
&\overset{(a)}{\le} \left\{T\left(M_2^2 + 2\kappa M_1 M_\infty t_{\text{mix}}\right) + T\int_0^1\left(\sum_{j=0}^{\infty} 1_{u<\alpha_j}\right)Q(u)^2 du\right\}^{\frac{1}{2}} \\
&= \left\{T\left(M_2^2 + 2\kappa M_1 t_{\text{mix}}\right) + T\sum_{j=0}^{\infty}\int_0^{\alpha_j} Q(u)^2 du\right\}^{\frac{1}{2}} \\
&\overset{(b)}{\le} \left\{T\left(M_2^2 + 2\kappa M_1 M_\infty t_{\text{mix}}\right) + \frac{1+l}{l-1}M_{1+l}^2 T\sum_{j=0}^{\infty}\alpha_j^{\frac{l-1}{1+l}}\right\}^{\frac{1}{2}} \\
&\overset{(c)}{\le} \left\{T\left(M_2^2 + 2\kappa M_1 M_\infty t_{\text{mix}}\right) + \frac{\kappa^{\frac{l-1}{1+l}}}{1-\rho}M_{1+l}^2\left(\frac{1+l}{l-1}\right)^2 T\right\}^{\frac{1}{2}} \\
&\lesssim T^{\frac{1}{2}}\left\{M_2^2 + t_{\text{mix}}\left(\kappa M_1 M_\infty + \kappa^{\frac{l-1}{1+l}}M_{1+l}^2\left(\frac{1+l}{l-1}\right)^2\right)\right\}^{\frac{1}{2}} \\
&\overset{(d)}{\lesssim} \sqrt{T(M_2^2 + \kappa t_{\text{mix}}\left(M_1 M_\infty + M_\infty^2\right))} \lesssim \sqrt{T(1+t_{\text{mix}})} \overset{(e)}{\le} \sqrt{Tt_{\text{mix}}},
\end{aligned}
$$

where (a) holds due to the bound (D.14) for the sum of covariance, and (b) holds due to the inequality (D.13), and (c) uses Lemma D.5.2 and the inequality that $\left(1-\rho^{\frac{\delta-1}{\delta+1}}\right)^{-1} \le \frac{1}{1-\rho}\frac{\delta+1}{\delta-1}$, (d) follows by setting $l \to \infty$, and (e) uses the fact that $t_{\text{mix}} \ge 1$ when $t_{\text{mix}} > 0$. □

In the end, we provide the proof for Lemma D.5.2.

Proof of Lemma D.5.2. From Section 3 in[176], we know that if $\{\xi_t\}_{t>0}$ is a (not necessarily stationary) Markov chain, then by the Markov property and an elementary argument,

$$\alpha_t = \sup_{k\in\mathbb{N}} \alpha(\sigma(\xi_k), \sigma(\xi_{k+t})).$$

By Definition D.5.1, it follows that

$$\alpha(\sigma(\xi_k), \sigma(\xi_{k+t})) = 2\sup\left\{\mathrm{Cov}(1_{\xi_k\in A}, 1_{\xi_{k+t}\in B}) \,:\, \text{both } A, B \text{ measurable}\right\}.$$

In the following, we fix $k \in \mathbb{N}$ and two measurable sets A, B. For simplicity, we denote $h_1(\xi_k) = 1_{\xi_k\in A} - \mathbb{P}(\xi_k \in A)$ and $h_2(\xi_{k+t}) = 1_{\xi_{k+t}\in B} - \mathbb{P}(\xi_{k+t} \in B)$. It then follows that

$$\begin{aligned}
\mathrm{Cov}(1_{\xi_k\in A}, 1_{\xi_{k+t}\in B}) &= \mathbb{E}h_1(\xi_k)h_2(\xi_{k+t}) = \mathbb{E}[h_1(\xi_k)\mathbb{E}[h_2(\xi_{k+t})|\mathscr{F}_k]] \\
&= \mathbb{E}[h_1(\xi_k)\mathscr{P}^t h_2(\xi_k)] \\
&= \mathbb{E}[h_1(\xi_k)(\mathscr{P}^t h_2(\xi_k) - \mathbb{E}_{\xi\sim\pi}h_2(\xi))] + \mathbb{E}h_1(\xi_k)\cdot\mathbb{E}_{\xi\sim\pi}h_2(\xi) \\
&\overset{(a)}{=} \mathbb{E}[h_1(\xi_k)(\mathscr{P}^t h_2(\xi_k) - \mathbb{E}_{\xi\sim\pi}h_2(\xi))] \\
&\le \mathbb{E}|h_1(\xi_k)|\cdot|\mathscr{P}^t h_2(\xi_k) - \mathbb{E}_{\xi\sim\pi}h_2(\xi)| \overset{(b)}{\le} \mathbb{E}|h_1(\xi_k)|\cdot\kappa\rho^t \le \kappa\rho^t,
\end{aligned}$$

where (a) uses $\mathbb{E}h_1(\xi_k) = 0$ and (b) uses Lemma 3.2.1 and the fact that both $h_1(\cdot)$ and $h_2(\cdot)$ are uniformly bounded by 1. Taking maximum over all $k \in \mathbb{N}$ and measurable sets A, B, we conclude that $\alpha_t \le \kappa\rho^t$. □

D.6 Additional Experimental Details

Further details for Table 3.2 Each time we initialize $\boldsymbol{x}_0$ as a zero vector, set $\eta_t = d^{-0.5}t^{-0.505}$, and always abandon the first 5% iterates for a warm up. We set $d = 10$ and $\rho_\varepsilon = 0.9$ in all experiments. The bootstrap method discards the first 400 samples as a warm up. According to Ramprasad, Li, Yang, Wang, Sun, Cheng [43], we set the step size as $\eta_t = 0.75\cdot t^{-0.75}$ and use $B \in \{10, 100, 200\}$.

Details bout Figure 3.3 The random MDP is generated in a similar way as[21]. In particular, for each (s, a) pair, the random reward $R(s, a) \sim \mathcal{N}(r(s, a), 1)$ is normally distributed with the mean $r(s, a)$ sampled from $[0, 1]$ uniformly initially and the transition probability $P(s'|s, a) = u(s')/\sum_s u(s)$, where $u(s) \overset{i.i.d.}{\sim} \mathcal{U}(0, 1)$. We choose the MDP size to be $|\mathcal{S}| = |\mathcal{A}| = 5$ and $\gamma = 0.6$. More iterations are required to conduct statistical inference in larger MDPs with larger γ. We choose a zero initial Q-value function. We abandon the first 4000 iterates as a warm up and use the following 50000 iterates to conduct statistical inference. We repeat the process for 200 times and use the polynomial step size $\eta_t = (t + 1)^{-0.51}$ and zero initial point each time.

Details bout Figure 3.4 We abandon the first 3000 iterates as a warm-up and use the following 40000 iterates to conduct statistical inference. Here we set $d = 5$. Both $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ are initialized as zero vectors. Larger d requires more iterations and more carefully parameter tuning to produce comparable performance. We set the step size as $\eta_t = t^{-0.501}$ for all the experiments in this figure. The target parameter is $\boldsymbol{\theta}^\top \boldsymbol{x}^\star$ where $\boldsymbol{\theta} = (1, \cdots, 1)^\top / \sqrt{d} \in \mathbb{R}^d$ and $\boldsymbol{x}^\star$'s coordinates evenly spread in the interval $[0, 1]$.

Details bout Figure 3.5 The problem dimension is $d = 5$ where both $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ are initialized as zero vectors. The target parameter is $\boldsymbol{\theta}^\top \boldsymbol{x}^\star$ where $\boldsymbol{\theta} = (1, \cdots, 1)^\top / \sqrt{d} \in \mathbb{R}^d$ and $\boldsymbol{x}^\star$'s coordinates evenly spread in the interval $[0, 1]$. For experiments in the first row, we abandon the first 3000 iterates as a warm-up and use the following 50000 iterates to conduct statistical inference. The step size $\eta_t = t^{-\alpha}$. For experiments in the second row, we again abandon the first 3000 iterates as a warm-up and use the following 50000 iterates to conduct statistical inference. The step size $\eta_t = \eta t^{-0.501}$. For experiments in the last row, we again abandon the first N iterates as a warm-up and use the following 50000 iterates to conduct statistical inference. The step size $\eta_t = t^{-0.501}$.

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Prof. Zhihua Zhang for his guidance, support and encouragement throughout my doctoral journey. The freedom to choose research questions and the respect that my mentor provided were the most valuable treasures during my academic journey. With this autonomy and encouragement, I was able to discover the field that resonated with me the most, and that I felt was worth my unwavering dedication.

I would like to express my gratitude to the senior fellows and collaborators who have been instrumental in shaping my research and developing my skills. I want to thank Haishan Ye, Michael I. Jordan, Qiang Sun, Shusen Wang, Weijie Su, and Xiangyu Chang (in alphabetical order) for their expert knowledge, constructive feedback, and patience. Their valuable insights and mentorship have been invaluable in guiding my research. I am also grateful to Naiyan Wang and Shuchang Zhou for providing me with the opportunity to work as an intern in high-tech companies. These internships not only allowed me to develop my coding skills but also enriched my project experience, which will undoubtedly facilitate my future academic development.

I would like to extend my sincere thanks to my friends and colleagues who have supported me throughout my doctoral journey. Their camaraderie, encouragement, and collaboration have been invaluable to me, and I am grateful for the friendships and professional connections that have grown from the shared experiences. Those people include Chengzhuo Ni, Dachao Lin, Guangzeng Xie, Hengchao Chen, Jiadong Liang, Kun Chen, Lei Ding, Naixin Guo, Shuyun Lin, Wenhao Yang,Yitan Wang, Yuze Han (in an alphabet order). Among them, I would like to express my special gratitude to Jiadong Liang for his insightful feedback, stimulating discussions, willingness to offer help, and collaborative daily workout, which have made my academic and personal life more fulfilling. I feel honored to have had the opportunity to work alongside him.

I would like to acknowledge my family for their love, encouragement, and understanding, which gave me the strength and motivation to pursue this degree.

I also want to thank those harsh reviewers. It is their meticulousness that makes me pay more attention to details and logical reasoning and improve writing and presentation clarity.

Finally, I would like to thank the funding agencies and institutions that provided financial support for my research, without which my works would not have been possible.

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：李翔　日期：2023 年 5 月 7 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/☑三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：李翔　导师签名：张志华

日期：2023 年 5 月 9 日