Junchi Li

Probabilistic Machine Learning Vol. II: Learning Theory

Topics on Online Learning Algorithms: Quantile Estimation, Regression, ROOT-SGD, and Beyond

Stanford University

Current Draft: December 9, 2024

Springer

Contents

Chapter 0
Overview

This overview chapter introduces from a high level the recent progresses on learning theory, including topics on online learning algorithms: quantile estimation, regression, ROOT-SGD, and beyond.

0.3 Deviation Theorems in Quantile Estimation and Online Quantile Regression

Quantile estimation and regression are robust statistical methods widely used in analyzing the relationship between a response variable and a set of covariates. Unlike traditional regression methods that focus on the mean, quantile regression provides insights into different points of the conditional distribution of the response variable. This approach is particularly useful in applications such as risk management and econometrics, where understanding the distribution of outcomes, rather than just the average, is crucial.

The estimation of quantiles from empirical data has been a focal point in statistical analysis, with extensive research dedicated to understanding the asymptotic properties of sample quantiles. These properties are often studied through the lens of moderate and large deviation principles, which provide essential insights into the likelihood of significant deviations in quantile estimates as the sample size increases. Moderate deviations bridge the gap between the central limit theorem, which addresses small fluctuations, and large deviations, which consider the probability of rare events. Understanding these deviations is critical for deriving statistical guarantees and assessing the performance of quantile-based methods under varying conditions.

In parallel, *quantile regression* (QR) is a powerful tool that does not assume a specific distribution for the error term, making it highly effective for handling skewed or heterogeneous data. However, traditional QR methods face significant challenges in dynamic environments where data arrive in streams, requiring real-time updates. Conventional QR techniques, which rely on static datasets, are often impractical for real-time applications due to their computational complexity and storage demands.

To address these limitations, we propose an *Online Quantile Regression* (`Online-QR`) algorithm designed for efficient real-time processing of streaming data. Our approach reformulates the check loss optimization problem inherent in QR as a least squares problem, significantly enhancing computational efficiency. By leveraging a Bayesian framework, the `Online-QR` algorithm performs online updates using the posterior distribution from previous data as the prior for new data. This method allows for rapid adaptation to new information, maintaining high accuracy while supporting the demands of real-time data streams.

We extend the `Online-QR` algorithm to *Multiple Quantile Regression* (`Multiple-QR`), providing a flexible tool for simultaneous estimation across multiple quantile levels. This extension is particularly valuable in applications where understanding the full distribution of a response variable is necessary.

This chapter makes several key contributions. First, we offer new theoretical insights into the moderate and large deviations in quantile estimation, providing refined statistical guarantees for quantile-based methods in various contexts. Second, we introduce a novel `Online-QR` algorithm that significantly improves the efficiency of quantile regression in streaming data environments. Finally, we extend this algorithm to `Multiple-QR`, broadening its applicability.

0.4 ROOT-SGD: Sharp Nonasymptotics and Near-Optimal Asymptotics in a Single Algorithm

Stochastic optimization has become a cornerstone in machine learning and statistical learning, particularly for large-scale and high-dimensional data. Among various stochastic optimization techniques, *stochastic gradient descent* (SGD) stands out due to its simplicity and effectiveness [153]. However, the performance of SGD is heavily influenced by the stepsize schedule, which determines the balance between convergence speed and stability.

Let $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ be differentiable as a function of its first argument, and consider the following unconstrained minimization problem:

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{where } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big], \tag{0.1}$$

and where the expectation is taken over a random vector $\xi \in \Xi$ with distribution $\mathbb{P}$. Our goal is to approximately solve this minimization problem based on samples $(\xi_i)_{i=1,2,\cdots} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and moreover to do so in a way that is computationally efficient and statistically optimal. When the samples arrive as an online stream, it is desirable to compute the approximate solution in a single pass, without storing the data, and this chapter focuses on this online setting.

Stochastic optimization problems of this type underpin a variety of methods in large-scale machine learning and statistical inference. One of the simplest methods is *stochastic gradient descent* (SGD), which recursively updates a parameter vector θ_t by taking a step in the direction of a single stochastic gradient, with a (possibly) time-varying step-size η_t [153]. This simple strategy has been surprisingly successful in modern large-scale statistical machine learning problems [139, 28, 145]; however, it can be substantially improved, both in theory and in practice, by algorithms that make use of more than a single stochastic gradient. Such algorithms belong to the general family of *stochastic first-order methods*. Various procedures have been studied, involving different weightings of past stochastic gradients, and also a range of analysis techniques. The diversity of approaches is reflected by the wide range of terminology, including *momentum*, *averaging*, *acceleration*, and *variance reduction*. All of these ideas center around two main underlying goals—that of proceeding quickly to a minimum, and that of arriving at a final state that achieves the optimal statistical efficiency and also provides a calibrated assessment of the uncertainty associated with the solution.

More concretely, the former goal requires the algorithm to achieve a fast rate of convergence and low sample complexity, ideally matching that of the noiseless case and the information-theoretic limit. For example, gradient descent takes $O\left(\frac{L}{\mu}\right)$ number of iterations to optimize a L-smooth and μ-strongly convex function. It is therefore desirable that the sample-size requirement for a stochastic optimization algorithm scales linearly with $O\left(\frac{L}{\mu}\right)$, with additional terms characterizing the effect of random noise on optimality. On the other hand, the latter goal imposes a more fine-grained requirement on the estimator produced by the algorithm. Roughly speaking, we need the estimator to share the same *optimal statistical properties* typically possessed by the empirical risk minimizer (were it be computed exactly in the batch setting). The notion of *statistical efficiency*, in both its asymptotic and nonasymptotic forms, allows for a fine-grained study of these issues.

0.5 ROOT-SGD with Adaptive, Diminishing Stepsize for Statistically Efficient Stochastic Optimization

For the ROOT-SGD algorithm we introduced, the *diminishing stepsize* strategy has been proposed to overcome the limitations of fixed stepsize schemes, offering a way to improve both the efficiency and robustness of SGD. This strategy provides adaptive learning rates that decrease over time, leading to better convergence properties in various settings. Despite its potential, integrating diminishing stepsize strategies with SGD in a way that optimally balances stochastic optimization and statistical efficiency remains a challenge.

In this chapter, we revisit ROOT-SGD, recently studied by [114], a novel optimization framework that enhances both the convergence and stability of stochastic gradient methods. ROOT-SGD is designed to achieve theoretical optimality and practical effectiveness, producing estimators that exhibit the same *optimal statistical* *properties* as empirical risk minimizers. The estimator produced by the ROOT-SGD algorithm retains these optimal properties in both asymptotic and non-asymptotic settings.

The notion of statistical efficiency allows for a rigorous assessment of optimality. The (Bayesian) Cram'er-Rao lower bounds provide the fundamental limit of the *mean-squared error* (MSE) of an estimator in relation to the Fisher information.[1] Furthermore, local asymptotic minimax theorems demonstrate that, under any bowl-shaped loss function, the optimal asymptotic distribution is Gaussian [173, 57]. The asymptotic covariance reflects the local complexity, and it is desirable to achieve this optimal bound with a *unity* pre-factor. Under relatively mild conditions, the empirical risk minimizer achieves this.

In contrast, our understanding of which first-order stochastic algorithms are optimal (or non-optimal) in this fine-grained way remains complete. Most existing performance guarantees are too coarse for this purpose, as the convergence rates are measured with worst-case problem-specific parameters, and bounds are given up to universal constants instead of unity in the asymptotic limit. This motivates us to establish performance guarantees for an efficient algorithm that match the optimal statistical efficiency with *unity pre-factor*, both asymptotically and non-asymptotically.

[1] The standard Cram'er-Rao lower bounds are valid only for unbiased estimators, whereas the Bayesian Cram'er-Rao lower bound applies to the Bayes risk of *any* estimator [78].

Chapter 3
Deviation Theorems in Quantile Estimation and Online Quantile Regression

This chapter presents new theoretical insights and algorithms for quantile estimation and regression in streaming data environments. We explore moderate and large deviation principles within quantile estimation, offering refined approximations that bridge the gap between central limit theorem results and large deviation principles. Additionally, we introduce an efficient Online Quantile Regression (`Online-QR`) algorithm designed for real-time data processing. By leveraging a Bayesian framework, our method reformulates the traditional quantile regression problem into a least squares problem, enhancing computational efficiency while maintaining high accuracy. We also extend the `Online-QR` to support Multiple Quantile Regression (`Multiple-QR`), enabling simultaneous estimation across multiple quantile levels. Theoretical analysis provides asymptotic properties, including unbiasedness, asymptotic normality, and linear convergence of the proposed estimators. Our findings are significant for applications requiring robust, real-time analysis of skewed or heterogeneous data streams.

Keywords:

Online Quantile Regression; Moderate and Large Deviations; Bayesian Inference; Streaming Data Processing; Multiple Quantile Regression

3.1 Introduction

As briefly introduced in the overview section, we continue with the discussions in two stages:

Deviation theorems in quantile estimation

Quantile estimation is fundamental in statistical analysis, with applications ranging from risk management to econometrics. The estimation of quantiles from empirical data has been extensively studied, particularly concerning the asymptotic properties of sample quantiles. Theoretical developments often focus on understanding how sample quantiles deviate from their true values as the sample size increases. These deviations can be analyzed through moderate and large deviation principles, which provide critical insights into the likelihood of significant deviations in quantile estimates.

Moderate deviation principles bridge the gap between the central limit theorem, which addresses small fluctuations, and large deviation principles, which consider the probability of rare events. While moderate deviations offer refined approximations for probabilities within an intermediate range, large deviations focus on tail events with exponentially small probabilities. In quantile estimation, understanding these deviations is crucial for deriving statistical guarantees and assessing the performance of quantile-based methods under various conditions.

Streaming linear quantile regression

Unlike traditional regression techniques, *quantile regression* (QR) does not assume a specific distribution for the error term, making it particularly effective for handling skewed or heterogeneous data. This versatility has led to widespread applications in fields such as economics and finance, where understanding the distribution of outcomes, rather than just the average, provides deeper insights.

However, traditional QR methods face significant challenges in streaming data environments. In these scenarios, data arrive online in blocks, requiring real-time updates and rapid processing. Conventional QR techniques, which rely on static datasets, are not well-suited to these dynamic environments. The computational complexity and storage requirements of traditional methods hinder their practicality for real-time applications, where quick and efficient updates are essential.

To address these limitations, we propose an *Online Quantile Regression* (`Online-QR`) algorithm designed for efficient real-time processing of streaming data. Our approach reformulates the check loss optimization problem inherent in QR as a least squares problem, significantly enhancing computational efficiency. By leveraging a Bayesian framework, the `Online-QR` algorithm performs online updates using the posterior distribution from previous data as the prior for new data. This method allows for rapid adaptation to new information, maintaining high accuracy while supporting the demands of real-time data streams.

We extend the `Online-QR` algorithm to *Multiple Quantile Regression* (`Multiple-QR`) and provide a comprehensive theoretical analysis, demonstrating the unbiasedness, asymptotic normality, and linear convergence of the `Online-QR` estimator.

Organization

The remainder of this chapter is organized as follows. Section 3.2 delves into new deviation results in quantile estimation. Section 3.3 briefly reviews Gibbs sampling for Bayesian QR and provides a detailed description of the proposed `Online-QR` algorithm. In Section 3.4, we extend `Online-QR` to the `Multiple-QR` model. Section 3.5 presents the theoretical analysis for `Online-QR`, including the unbiasedness, asymptotic normality, linear convergence of the `Online-QR` estimator, and the regret growth rate of the `Online-QR` learning procedure. All technical proofs are deferred to Section 3.6. Section 3.7 offer concluding remarks.

Notations

We write $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$. Let $\Phi(\cdot)$ denote the cumulative distribution function of a standard normal random variable. Specifically, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\{-t^2/2\}\,dt$ represents the standard normal distribution function. Additional notations will be introduced as they appear.

3.2 Moderate and large deviations in quantile estimation

This section aims to advance the theoretical understanding of sample quantiles by deriving new results within the framework of moderate and large deviations. We build on existing work in this area, particularly the Bahadur-Rao large deviations theory, to establish tighter bounds and uncover more nuanced behavior of quantile estimators. Our approach also investigates Cramér-type deviations, further enriching the analysis of empirical quantile functions.

Mathematically, recall the definitions of the quantile of population and the quantile of sample. Assume that $\{X_i, 1 \le i \le n\}$ is a sample of a population X with a common distribution function $F(x)$. For $0 < p < 1$, denote

$$x_p := F^{-1}(p) = \inf\{x : F(x) \ge p\}$$

the p-th quantile of $F(x)$. An estimator of x_p is given by the sample p-th quantile defined as follows:

$$x_{n,p} := F_n^{-1}(p) = \inf\{x : F_n(x) \ge p\}$$

where

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le x), \quad x \in \mathbf{R}$$

is the empirical distribution function of the sample $\{X_i, 1 \le i \le n\}$. Here, $\mathbf{1}(A)$ denotes the indicator function of set A.

There are a number of literatures to study the asymptotic properties for sample quantiles. If $x_{n,p}$ is the unique solution of the equation $F(x-) \leq p \leq F(x)$, then $x_{n,p} \to x_p$ a.e.; Assume that $F(x)$ have a continuous density function $f(x)$ in a neighborhood of x_p such that $f(x_p) > 0$. Then $\frac{\sqrt{n} f(x_p)(x_{n,p}-x_p)}{\sqrt{p(1-p)}}$ converges to the standard normal random variable in distribution, see Serfling [158]. Suppose that $F(x)$ is twice differentiable at x_p, then Bahadur [18] proved that

$$x_{n,p} = x_p + \frac{p - F_n(x_p)}{f(x_p)} + R_n, \quad a.e.$$

where $R_n = O(n^{-3/4}(\log n)^{3/4})$ a.e., $n \to \infty$. Xu and Miao [188] (see also Miao, Chen and Xu [131] for order statistics) obtained the following moderate deviation principles for $x_{n,p} - x_p$. For any sequence of real numbers $\{a_n\}_{n\geq 1}$ satisfying $a_n \to \infty$ and $a_n/\sqrt{n} \to 0$ as $n \to \infty$, it holds for any $r > 0$,

$$\lim_{n\to\infty} \frac{1}{a_n^2} \log \mathbb{P}\left(\frac{\sqrt{n}}{a_n} |x_{n,p} - x_p| \geq t \right) = -\frac{f(x_p)^2 r^2}{2p(1-p)} \tag{3.1}$$

They also obtained the following large deviation principles for $x_{n,p} - x_p$: the following two equality hold for any $t > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(x_{n,p} - x_p \geq t \right) = -\inf_{y \geq 1-p} \Lambda_+^*(y) \tag{3.2}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(x_{n,p} - x_p \leq -t \right) = -\inf_{y \geq p} \Lambda_-^*(y) \tag{3.3}$$

where

$$\Lambda_+^*(y) = y \log \frac{y}{F(x_p + t)} + (1-y) \log \frac{1-y}{1 - F(x_p + t)}$$

and

$$\Lambda_-^*(y) = y \log \frac{y}{F(x_p - t)} + (1-y) \log \frac{1-y}{1 - F(x_p - t)}$$

In this section, we are interested in sharp moderate and large deviations between the quantiles of population and the quantiles of samples. More precisely, we establish Cramér type moderate deviations and Bahadur-Rao type large deviations for $x_{n,p} - x_p$. Our results refine the moderate and large deviation principle results (3.1)–(3.3).

Main results

For brevity, denote

$$R_n(x,p)=\frac{\sqrt{n}f(x_p)(x_{n,p}-x_p)}{\sqrt{p(1-p)}},\quad p\in(0,1)$$

Denote $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\{-t^2/2\}dt$ the standard normal distribution function. The following theorem gives a Cramér type moderate deviation for sample quantiles.

Theorem 3.1. *Let $f(x)$ be the density function of X and let $p\in(0,1)$. If $f'(x)$ is bounded in a neighborhood of $x=x_p$ and $f(x_p)>0$, then it holds*

$$\log\frac{\mathbb{P}\big(\pm R_n(x,p)\ge t\big)}{1-\Phi(t)}=O\Big(\frac{1+t^3}{\sqrt{n}}\Big)\tag{3.4}$$

uniformly for $0\le t=o(\sqrt{n})$ as $n\to\infty$.

Remark 3.1. Let us comment on the result of Theorem 3.1.

(i) Assume that $f'(x)$ is uniformly bounded on $\mathbf{R}$ and that $f(x)$ is positive for all $x\in\mathbf{R}$. When p is replaced by p_n which may depend on n, by inspecting the proof of Theorem 3.1, the following equality

$$\log\frac{\mathbb{P}\big(\pm R_n(x,p_n)\ge t\big)}{1-\Phi(t)}=O\Big(\frac{1+t^3}{\sqrt{np_n(1-p_n)}}\Big)\tag{3.5}$$

holds uniformly for $0\le t=o\big(\sqrt{np_n(1-p_n)}\big)$ as $n\to\infty$. Clearly, if $p_n(1-p_n)\to 0$ as $n\to\infty$, then the last range tends to smaller than $0\le t=o(\sqrt{n})$.

(ii) By an argument similar to the proof of Corollary 3 in [64], the following moderate deviation principle (MDP) result is a consequence of Theorem 3.1. Let $\{a_n\}_{n\ge1}$ be a sequence of real numbers satisfying $a_n\to\infty$ and $a_n/\sqrt{n}\to0$ as $n\to\infty$. Then for each Borel set B,

$$\begin{aligned}-\inf_{x\in B^o}\frac{x^2}{2}&\le\liminf_{n\to\infty}\frac{1}{a_n^2}\log\mathbb{P}\Big(\frac{R_n(x,p)}{a_n}\in B\Big)\\&\le\limsup_{n\to\infty}\frac{1}{a_n^2}\log\mathbb{P}\Big(\frac{R_n(x,p)}{a_n}\in B\Big)\le-\inf_{x\in\overline{B}}\frac{x^2}{2}\end{aligned}\tag{3.6}$$

where B^o and $\overline{B}$ denote the interior and the closure of B, respectively. When B is $[t,\infty)$ or $(-\infty,t]$ for some $t>0$, the MDP result (3.6) has been established by Xu and Miao [188] (cf. equality (3.1)). Notice that, in Xu and Miao [188], MDP result holds without the assumption that $f'(x)$ is bounded in a neighborhood of $x=x_p$.

Using the inequality $|e^x-1|\le e^c|x|$ valid for $|x|\le c$, from Theorem 3.1, we obtain the following result about the relative errors of normal approximations.

Corollary 3.1. *Assume that the conditions of Theorem 3.1 are satisfied. Then it holds*

$$\frac{\mathbb{P}\big(\pm R_n(x,p) \geq t\big)}{1-\Phi(t)} = 1 + O\Big(\frac{1+t^3}{\sqrt{n}}\Big) \tag{3.7}$$

uniformly for $0 \leq t = O(n^{1/6})$ *as* $n \to \infty$, *which implies that*

$$\frac{\mathbb{P}\big(\pm R_n(x,p) \geq t\big)}{1-\Phi(t)} = 1 + o(1) \tag{3.8}$$

uniformly for $0 \leq x = o(n^{1/6})$.

From (3.7) in Corollary 3.1, by an argument similar to the proof of Corollary 2.2 in Fan *et al.* [63], we can obtain the following Berry-Esseen bound:

$$\sup_{t \in \mathbf{R}} \Big| \mathbb{P}\big(\pm R_n(x,p) \leq t\big) - \Phi(t) \Big| = O\Big(\frac{1}{\sqrt{n}}\Big), \quad n \to \infty \tag{3.9}$$

The last convergence rate coincides with the classical result established by Reiss [152].

Theorem 3.1 is devoted to the moderate deviations. For sharp large deviations, we have the following Bahadur-Rao type large deviation expansions.

Theorem 3.2. *For any* $t \geq 0$, *it holds*

$$\mathbb{P}\big(x_{n,p} - x_p \geq t\big) = \frac{1}{\tau_t^+ \, \sigma_p \sqrt{2\pi n}} e^{-n\Lambda^+(t)} \Big[1 + o(1)\Big], \quad n \to \infty$$

where

$$\tau_t^+ = \log \frac{F(x_p+t)(1-p)}{p(1-F(x_p+t))}, \quad \sigma_p = \sqrt{p(1-p)} \quad \textit{and}$$

$$\Lambda^+(t) = p \log \frac{p}{F(x_p+t)} + (1-p) \log \frac{1-p}{1-F(x_p+t)}$$

Similarly, it also holds for any $t \geq 0$,

$$\mathbb{P}\big(x_{n,p} - x_p \leq -t\big) = \frac{1}{\tau_t^- \, \sigma_p \sqrt{2\pi n}} e^{-n\Lambda^-(t)} \Big[1 + o(1)\Big], \quad n \to \infty$$

where

$$\tau_t^- = \log \frac{p(1-F(x_p-t))}{F(x_p-t)(1-p)} \quad \textit{and}$$

$$\Lambda^-(t) = p \log \frac{p}{F(x_p-t)} + (1-p) \log \frac{1-p}{1-F(x_p-t)}$$

Denote c a finite and positive constant which does not depend on n. For two sequences of positive numbers $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, write $a_n \asymp b_n$ if there exists a c such that $a_n/c \leq b_n \leq c\,a_n$ for all sufficiently large n. By Theorem 3.2, we have for any given constants $t > 0$ and $p \in (0,1)$,

$$\mathbb{P}\big(x_{n,p} - x_p \geq t\big) \asymp \frac{1}{\sqrt{n}} e^{-n\Lambda^{+}(t)} \quad \text{and} \quad \mathbb{P}\big(x_{n,p} - x_p \leq -t\big) \asymp \frac{1}{\sqrt{n}} e^{-n\Lambda^{-}(t)}, \quad n \to \infty$$

In particular, from the last line, we recover the following large deviation principle (LDP) result of Xu and Miao [188]: for any $t > 0$, the following equalities hold

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(x_{n,p} - x_p \geq t\big) = -\Lambda^{+}(t) \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(x_{n,p} - x_p \leq -t\big) = -\Lambda^{-}(t)$$

see equalities (3.2)–(3.3). Notice that $\Lambda^{+}(t) = \inf_{y\geq 1-p} \Lambda_{+}^{*}(y)$ and $\Lambda^{-}(t) = \inf_{y\geq p} \Lambda_{-}^{*}(y)$, see Remark 1 in Xu and Miao [188].

3.3 The `Online-QR` algorithm

Recall that linear QR assumes that the τ th $(0 < \tau < 1)$ conditional quantile of y given x is

$$Q_{y|x}(\tau) = \boldsymbol{\beta}^{o}(\tau) + x^{\top} \boldsymbol{\beta}(\tau)$$

Given a data set $\{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, N\}$, the $\mathbb{R}^{p+1}$-valued unknown regression parameters $\boldsymbol{\beta}^{\mathrm{QR}}(\tau) = (\boldsymbol{\beta}^{o}(\tau), \boldsymbol{\beta}(\tau))^{\top}$ at a fixed quantile level is estimated by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{QR}}(\tau) = \arg\min_{\boldsymbol{\beta}^{o}, \boldsymbol{\beta}} \sum_{i=1}^{N} \rho_{\tau}(y_i - \boldsymbol{\beta}^{o} - x_i^{\top} \boldsymbol{\beta}) \tag{3.10}$$

where $\rho_{\tau}(u) = u[\tau - \mathbb{I}(u < 0)]$ $(u \in \mathbb{R})$ is the check loss function and $\mathbb{I}(\cdot)$ is the indicator function. QR [97] is a powerful tool for studying the dependency of a response variable $y \in \mathbb{R}$ on a set of covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{T} \in \mathbb{R}^{p}$. Through modeling the conditional quantiles of y given x, it gives a comprehensive portrayal for their relationship. Moreover, QR does not impose distributional assumptions on the error and is hence suitable for analyzing skewed or heterogeneous data. Its applications appear widely in economics [42, 68, 70, 69] and finance [128, 5]. For example, studying the herding behavior in stock markets of Finland [156] and India [25], analyzing the inflation-return [3] and volatility-return relationships [17], and discussing the impacts of economic development and schooling on wage structure [31, 9].

Traditional statistical analysis is often based on a static data set. That is, the analysis is performed after all data become available. However, stream data, as data blocks arriving in an online fashion in time, demand frequent update of the analysis result as more and more data are collected [200, 75]. Due to rapid development of information technology, stream data have become ubiquitous in modern real-time data processing infrastructures. For example, financial institutions track the changes of stock price in real-time and calculate the value-at-risk to re-balance portfolio [179]. Land and resource management departments record the real-time house prices to timely grasp market dynamics [60, 79]. E-commerce companies analyze users' browsing and purchasing history in real-time for more accurate personalized recommendations [8, 66]. Although collected over time, compared to traditional time series, the analysis for stream data usually does not focus on modeling the temporal correlation structure [123]. Rather, the two key issues are (3.10) the storage challenge that old data are already discarded when new data arrive, and (3.11) the speed challenge that the analysis must be updated in a very fast way to support real-time information processing. Before the old data are discarded, information therein is typically compressed efficiently as a fixed dimensional object whose size does not grow over time. In this chapter, we assume that at a given time point t, only the current data block $\left(X_t, \boldsymbol{y}_t\right)$ and the compressed information based on the previous $(t-1)$ data blocks are available (see Fig. 1), and term the data processing and information aggregation at a single data block as local processing.

Conventionally, because of the non-smoothness in the check loss function, the QR estimation problem (3.10) cannot be directly solved by gradient-based methods. Rather it is written as a linear programming (LP) problem and solved by the interior point (IP) method [151] whose complexity is $O\left(N^{1+\alpha} p^3 \log N\right)$, $\alpha \in (0, 1/2)$ [84]. Its computation under the big data context has undergone rapid development recently. Common strategies include subsampling and divide-and-conquer. The subsampling methods rely on analyzing a subset of the big data. Other than the naive simple random sampling, the subset can be chosen by projecting the big data to a lower dimensional space [190] or based on optimal design principles [2]. The divide-and-conquer technique processes partitions of the entire data in parallel and an aggregated estimator is then computed to approximate the QR estimator $\widehat{\boldsymbol{\beta}}_{QR}(\tau)$. According to the number of rounds for information aggregation required by the estimation process, the divide-and-conquer algorithms can be classified into one-shot algorithms [39, 177, 38, 35] and multi-shot algorithms [198, 199, 66, 182, 109]. However, most of these techniques are only applicable to static big data but not stream data except the one-shot divide-and-conquer algorithms when viewing stream data as data blocks partitioned over the time domain. The subsampling methods obviously require the availability of the entire data set to sample from it, and the multi-shot divide-and-conquer algorithms need to repeatedly access the partitioned data blocks, which is infeasible in stream data. On the other hand, while the one-shot divideand-conquer algorithms can be naturally extended to stream data by implementing the aggregation in a block-by-block manner, their asymptotic theory often requires the block size to grow to infinity, while the block size in stream data is generally determined by hardware capacity and hence fixed. In addition, updates based on direct applications of one-shot divide-and-conquer algorithms may not be efficient enough to support the frequent update in the stream data context.

Another set of techniques closely related to our context is the recently proposed renewable regression estimators, including linear regression [176], generalized linear model [125] and linear mixed effect model [126]. Direct applications of the renewable estimator to QR are straightforward extensions of the one-shot divide-and-conquer technique, such as the simple average estimation (SAE) method developed in [202] and the renewable QR estimation method proposed in [179]. However, local processing in both methods require solving a QR problem, resulting in updates that are too slow for real-time processing of stream data. For example, the electronic health record data collected by the Scientific Registry of Transplant Recipients are updated every 10 min [126]; the partial discharges signal data in power grids are collected by ultra-high frequency sensors every 0.4 ns [178].

Our proposed online QR (`Online-QR`) algorithm was partly motivated by Bayesian QR for that online updates occur naturally under the Bayesian setup, i.e., using the posterior distribution obtained from the past analysis as the prior information for later analysis. Bayesian QR uses a linear model with an asymmetric Laplace (AL) error term to achieve equivalence between the original QR estimator and the Bayesian maximum-a-posteriori (MAP) estimator [194]. However, posterior simulation in Bayesian inference does not meet the computational speed requirement for stream data. In each update of the `Online-QR` algorithm, we essentially impose a Gaussian prior for the regression parameter centered at the previous estimate, which is equivalent to putting a ridge penalty. Then using the scale-mixture representation of the AL distribution [99], we can derive the updated regression parameter estimate from a normal linear model based on a proper estimate of the scale parameter. Hence, each update in `Online-QR` only requires solving a least squares problem and significantly improves the computational efficiency. The scale parameter update is done similar to the stochastic approximation (SA) for generalized linear mixed models (GLMM) [33].

1. Store $(X_t, \boldsymbol{y}_t)$ and $\overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}$
2. Compute $\widetilde{\boldsymbol{\nu}}_t$ and $\widetilde{\boldsymbol{\beta}}_t^{\text{QR}}$
3. Update $\overline{\boldsymbol{\beta}}_t^{\text{QR}}$

The general idea of `Online-QR` can easily extend to other variants of QR. In addition to the standard QR, we also exemplify its application to multiple, or composite, quantile regression (`Multiple-QR`) [208].

To speed up local processing to accommodate the frequent update in stream data, we convert the check loss optimization based on each data block to a least squares problem, which is partly motivated by the Gibbs sampling in Bayesian QR. To ensure that this chapter is self-contained, in this section we first briefly review the Gibbs sampler for Bayesian QR and then detail the proposed `Online-QR`. Throughout this section, we consider QR at a fixed quantile level, so we omit " τ " in all notations for simplicity.

3.3.1 The Gibbs sampler for Bayesian QR

The linear QR model at quantile level τ can also be written as $y = x^{*\top}\boldsymbol{\beta}^{\mathrm{QR}} + \varepsilon$, where $x^* = \left(1, x^\top\right)^\top \in \mathbb{R}^{p+1}$ and $Q_\varepsilon(\tau) = 0$. The most popular way to formulate Bayesian QR is proposed by [194], which imposed an asymmetric Laplace (AL) distribution, $\mathrm{AL}(0,1,\tau)$, on ε with location parameter 0, scale parameter 1, and skew parameter τ [196]. This gives the error density $f(\varepsilon) = \tau(1-\tau)\exp\{-\rho_\tau(\varepsilon)\}$, and the model likelihood is then

$$f\left(\boldsymbol{y} \mid \boldsymbol{\beta}^{\mathrm{QR}}\right) = \tau^N (1-\tau)^N \exp\left\{-\sum_{i=1}^{N} \rho_\tau(y_i - x_i^{*\top}\boldsymbol{\beta}^{\mathrm{QR}})\right\} \tag{3.11}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^\top \in \mathbb{R}^N$. Under a flat prior, finding the Bayesian MAP estimator is equivalent to solving the original QR problem (3.10). Posterior simulation can then be implemented through a Gibbs sampler [99] based on a scale-mixture representation of the AL distribution, which is given next in Lemma 3.3.

Lemma 3.3 (Proposition 1 of [99]). *If a random variable ε follows* $\mathrm{AL}(0,1,\tau)$, *then it can be represented as*

$$\varepsilon = \xi_1 v + \xi_2 \sqrt{v} u \qquad \textit{where } \xi_1 = \frac{1-2\tau}{\tau(1-\tau)} \textit{ and } \xi_2 = \sqrt{\frac{2}{\tau(1-\tau)}}$$

and v is a standard exponential variable and u is an (independent) standard normal variable.

Using Lemma 3.3, we have

$$y_i = x_i^{*\top}\boldsymbol{\beta}^{\mathrm{QR}} + \xi_1 v_i + \xi_2\sqrt{v_i}u_i \qquad i = 1, 2, \ldots, N \tag{3.12}$$

Conditioning on the scale parameter v_i, (3.12) is then a normal linear model,

$$y_i \mid \boldsymbol{\beta}^{\mathrm{QR}}, v_i \sim N(x_i^{*\top}\boldsymbol{\beta}^{\mathrm{QR}} + \xi_1 v_i, \xi_2^2 v_i) \tag{3.13}$$

Therefore, it is easy to derive the full conditional distributions of $\boldsymbol{\beta}^{\mathrm{QR}}$ and v_i as

$$\boldsymbol{\beta}^{\mathrm{QR}} \mid \boldsymbol{y}, \boldsymbol{v} \sim N(\widehat{\boldsymbol{\beta}}, \widehat{B}) \tag{3.14}$$

and

$$v_i \mid y_i, \boldsymbol{\beta}^{\mathrm{QR}} \sim \mathscr{GIG}\left(\frac{1}{2}, \widehat{\delta}_i, \widehat{\gamma}\right) \tag{3.15}$$

under the normal prior $\boldsymbol{\beta}^{\mathrm{QR}} \sim N(\boldsymbol{\beta}^o, B_0)$, where

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_N)^\top \qquad \widehat{\boldsymbol{\beta}} = \widehat{B}\left[\sum_{i=1}^{N} \frac{x_i^*(y_i - \xi_1 v_i)}{\xi_2^2 v_i} + B_0^{-1}\boldsymbol{\beta}^o\right]$$

$$\widehat{B}^{-1} = \sum_{i=1}^{N} \frac{x_i^* x_i^{*\top}}{\xi_2^2 v_i} + B_0^{-1} \qquad \widehat{\delta}_i^2 = \frac{\left(y_i - x_i^{*\top} \boldsymbol{\beta}^{\mathrm{QR}}\right)^2}{\xi_2^2} \qquad \widehat{\gamma}^2 = 2 + \frac{\xi_1^2}{\xi_2^2}$$

Here $\mathscr{G}\mathscr{I}(\cdot)$ represents the generalized inverse Gaussian distribution [47]. Eqs. (3.14) and (3.15) then form the Gibbs sampler for Bayesian QR.

3.3.2 The Online-QR algorithm

In this subsection, we give the algorithmic details of Online-QR. Assume that the data block arriving at time point $t (t = 1, 2, \ldots)$ is $(X_t, \boldsymbol{y}_t)$, where $X_t = (x_{1t}, x_{2t}, \ldots, x_{n_t t})^\top \in \mathbb{R}^{n_t x p}$ and $\boldsymbol{y}_t = (y_{1t}, y_{2t}, \ldots, y_{n_t t})^\top \in \mathbb{R}^{n_t}$. In the local processing at time point t, existing methods for QR estimation in stream data, including SAE [202] and the renewable QR [179], compute the following local QR estimator,

$$\widehat{\boldsymbol{\beta}}_t^{\mathrm{QR}} = \arg\min_{\boldsymbol{\beta}^{\mathrm{QR}}} \sum_{i=1}^{n_t} \rho_\tau (y_{it} - x_{it}^{*\top} \boldsymbol{\beta}^{\mathrm{QR}}) \tag{3.16}$$

which is a LP problem with time complexity $O(n_t^{1+\alpha} p^3 \log n_t)$, $\alpha \in (0, 1/2)$ [84]. Our Online-QR algorithm instead only solves a least squares problem and hence is significantly faster than these existing solutions.

The key observation is that, when conditioning on v_i, (3.12) is a normal linear model. Thus, if we can find a proper estimate $\widetilde{v}_{it}$ for v_{it} (here v_{it} is the scale parameter corresponding to (x_{it}, y_{it})) based on the compressed information till time point $(t-1)$, we can construct a pseudo response $y_{it}^* = y_{it} - \xi_1 \widetilde{v}_{it}$, and a local estimate of $\boldsymbol{\beta}^{\mathrm{QR}}$ can be given by the least squares estimator for model (3.12), i.e.,

$$\widetilde{\boldsymbol{\beta}}_t^{\mathrm{QR}} = (X_t^{*\top} X_t^*)^{-1} X_t^* \boldsymbol{y}_t^* \tag{3.17}$$

where $X_t^* = [\mathbf{1}_{n_t}, X_t] \in \mathbb{R}^{n_t \times (p+1)}$ and $\boldsymbol{y}_t^* = (y_{1t}^*, y_{2t}^*, \ldots, y_{n_t t}^*)^\top \in \mathbb{R}^{n_t}$.

Let $\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}$ denote the Online-QR estimator at time point $(t-1)$ (whose estimation will be discussed later in this section). By the full conditional distribution (3.15) in the Gibbs sampler for Bayesian QR, we have

$$f(v_{it} \mid y_{it}, \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}) \propto v_{it}^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\widehat{\delta}_{it}^2 v_{it}^{-1} + \widehat{\gamma}^2 v_{it}) \right\}$$

where

$$\widehat{\delta}_{it}^2 = \frac{(y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}})^2}{\xi_2^2}$$

We next define an ”estimate” of v_{it} as

$$\widetilde{v}_{it} = \frac{\sqrt{\widehat{\delta}_{it}^2\widehat{\gamma}^2 - \frac{1-2\tau+2\tau^2}{\tau(1-\tau)}|\widehat{\gamma}|\xi_2^{-1} + \widehat{\gamma}^2}}{\widehat{\gamma}^2} \tag{3.18}$$

This choice of $\widetilde{v}_{it}$ is developed based on the form of the intuitive "estimate", the expectation of the conditional distribution of $v_{it} \mid y_{it}, \overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}$, to guarantee the unbiasedness of the `Online-QR` estimator (see (3.32) in the proof of Theorem 3.4 in Section 3.6). This construction is inspired by the SA method developed for GLMM [33], as v_i is the random intercept when viewing model (3.12) as a linear mixed effect model.

Next, we discuss how to update the `Online-QR` estimator at time point t. In renewable QR, the update is done by taking a weighted average over the local estimates $\widehat{\boldsymbol{\beta}}_l^{\text{QR}}$ $(l = 1,2,\ldots,t)$. We use a different idea motivated by the online update in Bayesian inference and obtain an even simpler one. Specifically, we conduct the update by solving the following nonzero centered ridge regression problem,

$$\overline{\boldsymbol{\beta}}_t^{\text{QR}} = \arg\min_{\boldsymbol{\beta}^{\text{QR}}} \left\| y_t^* - X_t^*\boldsymbol{\beta}^{\text{QR}} \right\|_2^2 + \lambda_t \left\| C_t(\boldsymbol{\beta}^{\text{QR}} - \overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}) \right\|_2^2 \tag{3.19}$$

where C_t is a given matrix. The penalty term can be viewed equivalently as setting a Gaussian prior $N(\overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}, \lambda_t^{-1}(C_t^\top C_t)^{-1})$ for the Bayesian linear regression $y_t^* \mid \boldsymbol{\beta}^{\text{QR}} \sim N(X_t^*\boldsymbol{\beta}^{\text{QR}}, I_{n_t})$, where I_{n_t} is the n_t-dimensional identity matrix. Hence the penalty parameter λ_t serves a similar role as scaling the prior variance, i.e., how strongly we believe the current estimate should be similar to the previous one. Recently, [176] used this ridge regression idea in online linear regression for big data, in which λ_t was selected by cross validation and C_t was set as I_{p+1}. Since cross validation is infeasible in stream data, we set $\lambda_t = \frac{\sum_{l=1}^{t-1} n_l}{n_t}$ from its intuitive connection to the prior variance and let $C_t = X_t^*$ to guarantee that the problem (3.19) has a closed-form solution given as follows.

$$\overline{\boldsymbol{\beta}}_t^{\text{QR}} = \frac{n_t}{\sum_{l=1}^t n_l}\widetilde{\boldsymbol{\beta}}_t^{\text{QR}} + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l}\overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}} \tag{3.20}$$

The update of our proposed `Online-QR` algorithm at time point t involves evaluating (3.17), (3.18) and (3.20). And we summarize `Online-QR` in Algorithm 1 and Fig. 2, where $\widetilde{\boldsymbol{v}}_t = (\widetilde{v}_{1t}, \widetilde{v}_{2t}, \ldots, \widetilde{v}_{n_t t})^\top \in \mathbb{R}^{n_t}$, and $\overline{\boldsymbol{\beta}}_0^{\text{QR}}$ is the initial value of the algorithm, which can be set as the local estimate based on the first data block. When the duration of the data stream is T, we denote $\overline{\boldsymbol{\beta}}_T^{\text{QR}}$ as the final estimate for $\boldsymbol{\beta}^{\text{QR}}$.

Algorithm 1 The `Online-QR` algorithm for QR estimation in stream data

```
Input: data blocks $(X_t, y_t)$, $t = 1, 2, \ldots, T$ quantile level $\tau$, and the initial estimate $\overline{\boldsymbol{\beta}}_0^{\text{QR}}$
Set $\xi_1 = \frac{1-2\tau}{\tau(1-\tau)}$, $\xi_2 = \sqrt{\frac{2}{\tau(1-\tau)}}$ and $\widehat{\gamma}^2 = 2 + \frac{\xi_1^2}{\xi_2^2}$
for $t = 1, 2, \ldots, T$ do
    Compute $\widehat{\delta}_{it}^2 = \frac{\left(y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}\right)^2}{\xi_2^2}$ and $\widetilde{v}_{it} = \frac{\sqrt{\widehat{\delta}_{it}^2 \widehat{\gamma}^2} - \frac{1-2\tau+2\tau^2}{\tau(1-\tau)} |\widehat{\gamma}| \xi_2^{-1} + \widehat{\gamma}^2}{\widehat{\gamma}^2}$, $i = 1, 2, \ldots, n_t$
    Set $\boldsymbol{y}_t^* = \boldsymbol{y}_t - \xi_1 \widetilde{\boldsymbol{v}}_t$ and compute $\widetilde{\boldsymbol{\beta}}_t^{\text{QR}} = \left(X_t^{*\top} X_t^*\right)^{-1} X_t^* \boldsymbol{y}_t^*$
    Update $\overline{\boldsymbol{\beta}}_t^{\text{QR}} = \frac{n_t}{\sum_{l=1}^{t} n_l} \widetilde{\boldsymbol{\beta}}_t^{\text{QR}} + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^{t} n_l} \overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}}$
    $t = t + 1$
end for
return the final estimate $\overline{\boldsymbol{\beta}}_T^{\text{QR}}$
```

3.3.3 The space complexity of `Online-QR`

In stream data analysis, space complexity, i.e., the volume of objects compressed from historical data that are required in the algorithm update, is an important consideration [44]. It is closely related to the efficient use of storage resources, as we need to store these objects with the new data even after the old data are discarded. See Fig. 1. In our `Online-QR` algorithm, it only needs to store the fixed dimensional vector $\overline{\boldsymbol{\beta}}_{t-1}^{\text{QR}} \in R^{p+1}$ in updating $\overline{\boldsymbol{\beta}}_t^{\text{QR}}$, and thus the complexity is $O(p)$. The SAE method has the same storage requirement, because it also only needs the estimate from the previous time point. In contrast, the complexity of renewable QR is higher at $O\left(p^2\right)$ since it also needs to store the matrix $\sum_{l=1}^{t-1} X_l^{*\top} X_l^*$ for constructing the weights used in averaging $\widehat{\boldsymbol{\beta}}_l^{\text{QR}}$ $(l = 1, 2, \ldots, t)$.

3.4 The extension of `Online-QR` to `Multiple-QR`

The general idea of `Online-QR` developed in Section 3.3 can also extend to other variants of QR. In this section, we give the implementation of `Online-QR` for `Multiple-QR`.

`Multiple-QR` considers simultaneous modeling at many quantile levels $0 < \tau_1 < \tau_2 < \ldots < \tau_S < 1$ to improve the estimation efficiency of QR. It assumes identical slope parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ at different quantile levels but allows the intercept $\boldsymbol{\beta}^{o\tau_s} \in \mathbb{R} (s = 1, 2, \ldots, S)$ to vary. The local `Multiple-QR` estimator at time point t based on $(X_t, \boldsymbol{y}_t)$ is given by

$$\left(\widehat{\boldsymbol{\beta}}_{0\tau_1 t}, \widehat{\boldsymbol{\beta}}_{0\tau_2 t}, \ldots, \widehat{\boldsymbol{\beta}}_{0\tau_S t}, \widehat{\boldsymbol{\beta}}_t\right) = \underset{\left\{\boldsymbol{\beta}^{o\tau_s}\right\}_S, \boldsymbol{\beta}}{\arg\min} \sum_{s=1}^{S} \sum_{i=1}^{n_t} \rho_{\tau_s}(y_{it} - \boldsymbol{\beta}^{o\tau_s} - x_{it}^\top \boldsymbol{\beta}) \tag{3.21}$$

For simplicity, we denote $(\boldsymbol{\beta}^{o\tau_1}, \boldsymbol{\beta}^{o\tau_2}, \ldots, \boldsymbol{\beta}^{o\tau_S}, \boldsymbol{\beta})$ and $(\widehat{\boldsymbol{\beta}}_{0\tau_1 t}, \widehat{\boldsymbol{\beta}}_{0\tau_2 t}, \cdots, \widehat{\boldsymbol{\beta}}_{0\tau_S t}, \widehat{\boldsymbol{\beta}}_t)$ as $\boldsymbol{\beta}_{\texttt{Multiple-QR}} \in \mathbb{R}^{p+S}$ and $\widehat{\boldsymbol{\beta}}_{\texttt{Multiple-QR}t} \in \mathbb{R}^{p+S}$, respectively. Note that solving (3.21) is even slower than solving (3.16) in QR as both its dimensionality (i.e., $(p+S)$) and size (i.e., $S \times n_t$) are larger. In the following, we show how to convert this problem to a least squares problem using a similar idea as in Section 3.3 .

The scale-mixture representation for Bayesian `Multiple-QR` [4] is

$$y_{it} = \boldsymbol{\beta}^{o\tau_s} + x_{it}^\top \boldsymbol{\beta} + \xi_{1\tau_s} v_{it\tau_s} + \xi_{2\tau_s} \sqrt{v_{it\tau_s}} u_{it\tau_s} \qquad s = 1, 2, \ldots, S$$

where $v_{it\tau_s}$'s are standard exponential variables, $u_{it\tau_s}$'s are standard normal variables, $\xi_{1\tau_s} = \frac{1-2\tau_s}{\tau_s(1-\tau_s)}$ and $\xi_{2\tau_s} = \sqrt{\frac{2}{\tau_s(1-\tau_s)}}$. Thus, given a reasonable estimate $\widetilde{v}_{\text{itt}}$ for $v_{it\tau_s}$, (3.21) can also be converted to the least squares problem,

$$\widetilde{\boldsymbol{\beta}}_{\texttt{Multiple-QR}t} = \underset{\{\boldsymbol{\beta}^{o\tau_s}\}_S, \boldsymbol{\beta}}{\arg\min} \frac{1}{2} \sum_{s=1}^{S} \left\| \boldsymbol{y}_{\tau_s}^* - \boldsymbol{\beta}^{o\tau_s} \mathbf{1}_{n_t} - X_t \boldsymbol{\beta} \right\|_2^2 \tag{3.22}$$

where $y_{t\tau_s}^* = \left(y_{1t\tau_s}^*, y_{2t\tau_s}^*, \ldots, y_{n_t t\tau_s}^*\right)^\top \in \mathbb{R}^{n_t}$ and $y_{it\tau_s}^* = y_{it} - \xi_{1\tau_s} \widetilde{v}_{it\tau_s}$. Following the method in Section 3.3, we obtain $\widetilde{v}_{it\tau_s}$ by

$$\widetilde{v}_{it\tau_s} = \frac{\sqrt{\widehat{\delta}_{it\tau_s}^2 \widehat{\gamma}_{\tau_s}^2} - \frac{1-2\tau_s+2\tau_s^2}{\tau_s(1-\tau_s)} \left|\widehat{\gamma}_{\tau_s}\right| \xi_{2\tau_s}^{-1} + \widehat{\gamma}_{\tau_s}^2}{\widehat{\gamma}_{\tau_s}^2} \tag{3.23}$$

where $\widehat{\delta}_{it\tau_s}^2 = \frac{\left(y_{it} - \overline{\boldsymbol{\beta}}_{0\tau_s(t-1)} - x_{it}^\top \overline{\boldsymbol{\beta}}_{t-1}\right)^2}{\xi_{2\tau_s}^2}$ and $\widehat{\gamma}_{\tau_s}^2 = 2 + \frac{\xi_{1\tau_s}^2}{\xi_{2\tau_s}^2}$. And the online update of the `Online-QR` estimator of $\boldsymbol{\beta}_{\texttt{Multiple-QR}}$ at time point t is then

$$\overline{\boldsymbol{\beta}}_{\texttt{Multiple-QR}t} = \frac{n_t}{\sum_{l=1}^{t} n_l} \widetilde{\boldsymbol{\beta}}_{\texttt{Multiple-QR}t} + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^{t} n_l} \overline{\boldsymbol{\beta}}_{\texttt{Multiple-QR}(t-1)} \tag{3.24}$$

where $\overline{\boldsymbol{\beta}}_{\texttt{Multiple-QR}t} = (\overline{\boldsymbol{\beta}}_{0\tau_1 t}, \overline{\boldsymbol{\beta}}_{0\tau_2 t}, \ldots, \overline{\boldsymbol{\beta}}_{0\tau_S t}, \overline{\boldsymbol{\beta}}_t^\top)^\top$. Eqs. (3.22), (3.23), (3.24) then form the `Online-QR` algorithm for `Multiple-QR` estimation.

3.5 Theoretical analysis

In this section, we analyze the theoretical properties of `Online-QR`. Under mild conditions, we show that $\overline{\boldsymbol{\beta}}_t^{\text{QR}}$ can achieve unbiasedness after finite time and further as $t \to \infty$, this estimator is asymptotically normal. The asymptotic covariance matrix can be updated through a recursive formula. Besides, we also analyze the convergence rate of $\overline{\boldsymbol{\beta}}_t^{\text{QR}}$ and the regret growth rate of the `Online-QR` learning procedure. We provide these properties in Theorems 3.4, 3.5, 3.6.

Theorem 3.4. *Assume that there exists an infinite sequence of data blocks* $\{(X_t, \boldsymbol{y}_t), t = 1,2,\ldots\}$ *and for any* t, $(X_t, \boldsymbol{y}_t)$ *satisfies:*

(i) $\operatorname{rank}(X_t) = p$;
(ii) $y_t = (y_{1t}, y_{2t}, \ldots, y_{n_t t})^\top$ *are independent.*

Then under model (3.12), *the* `Online-QR` *estimator can achieve unbiasedness after finite time, i.e., there exists* $t_0 > 0$, *such that for* $t > t_0$, *we have*

$$\mathbb{E}(\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}) = \boldsymbol{\beta}^{\mathrm{QR}}$$

Further, if we assume

(iii) $\lambda_{\min} \le \lambda_j \left(\frac{X_t^\top X_t}{n_t}\right) \le \lambda_{\max}$, *where* $\lambda_{\min}$ *and* $\lambda_{\max}$ *are two positive constants, and* $\lambda_j(\cdot)$ *represents the* j*th eigenvalue of a matrix, and*
(iv) all data block sizes are on the same magnitude, i.e., $\max_t n_t / \min_t n_t = O(1)$.

Then as $t \to \infty$, *the* `Online-QR` *estimator is asymptotically normal*

$$\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} \xrightarrow{d} N\left(\boldsymbol{\beta}^{\mathrm{QR}}, \Sigma_t\right)$$

with the asymptotic covariance matrix updated recursively by

$$\Sigma_t = \frac{\left(\sum_{l=1}^{t-1} n_l\right)^2 + 2\left(\sum_{l=1}^{t-1} n_l\right) n_t (1-2\tau)^4 + n_t^2 (1-2\tau)^2}{\left(\sum_{l=1}^{t} n_l\right)^2} \Sigma_{t-1} + \frac{n_t^2 g(\tau)}{\left(\sum_{l=1}^{t} n_l\right)^2} (X_t^{*\top} X_t^*)^{-1}$$

where

$$g(\tau) = \frac{1 - 2\tau + 2\tau^2 + (1-2\tau)^2\left(1 - 2\tau - 2\tau^2 + 8\tau^3 - 4\tau^4\right) - 2(1-2\tau)^4}{\tau^2(1-\tau)^2}$$

is a function of the quantile level τ.

In proving the unbiasedness of the `Online-QR` estimator, we utilize the fact that the sequence $\{\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}\}, t = 1,2,\ldots$ is a Markov process, and the unbiasedness then comes as a result of the stationarity of this Markov process. The asymptotic normality of `Online-QR`, on the other hand, follows from the martingale central limit theorem after treating $(\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} - \boldsymbol{\beta}^{\mathrm{QR}})$ as the weighted sum of a martingale difference sequence. The detailed proof is given in Section 3.6.

In the recursive update of asymptotic covariance matrix, the initialization of Σ_1 is determined by the choice of the initial estimate in `Online-QR`. For example, if we set the initial estimate as the standard QR estimate based on the first data block $(X_1, \boldsymbol{y}_1)$, then $\Sigma_1 = \tau(1-\tau) f_\varepsilon^{-2}(0) \left(X_1^{*\top} X_1^*\right)^{-1}$ [97].

Theorem 3.5. *Under the assumptions (i)-(iv) in Theorem 3.4, the* `Online-QR` *estimator enjoys a sublinear convergence rate*

$$\left\|\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} - \boldsymbol{\beta}^{\mathrm{QR}}\right\|_2 = O_p\left(\frac{1}{t}\right)$$

The estimation error of `Online-QR` can be decomposed as follows:

$$\begin{aligned}\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} - \boldsymbol{\beta}^{\mathrm{QR}} &= (1-r_t)\left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} - \boldsymbol{\beta}^{\mathrm{QR}}\right) - r_t\left(\boldsymbol{\beta}^{\mathrm{QR}} - \widetilde{\boldsymbol{\beta}}_t^{\mathrm{QR}}\right)\\ &= \left[\prod_{l=1}^{t}(1-r_l)\right]\left(\overline{\boldsymbol{\beta}}_0^{\mathrm{QR}} - \boldsymbol{\beta}^{\mathrm{QR}}\right) - \sum_{l=1}^{t} r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left(\boldsymbol{\beta}^{\mathrm{QR}} - \widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}\right)\end{aligned} \tag{3.25}$$

where $r_t := \frac{n_t}{\sum_{l=1}^{t} n_l}$. In our proof, the initial error due to $\overline{\boldsymbol{\beta}}_0^{\mathrm{QR}}$ is controlled by quantifying the multiplying factor in front of it, while the additional errors from $\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}$'s occurring in the following updates can be controlled by the Pinelis-Bernstein inequality for martingale difference sequences [168]. Then we are able to prove the linear convergence rate, typically true for optimization problems with smooth objectives [134], without imposing more strict assumptions. The detailed proof is given in Section 3.6. Besides, it is worth noting that in the proof we do not require $n_t \to \infty$, but only conditions on the homogeneity of data block sizes, i.e., $\max_t n_t / \min_t n_t = O(1)$, which is different from other existing methods based on divide-and-conquer techniques, e.g., [202].

In the analysis of online learning algorithms, regret is an important statistical risk measure that evaluates the predictive performance of the algorithm. In Theorem 3.6, we give the regret growth rate of our `Online-QR`.

Theorem 3.6. *Under the assumptions (i)-(iv) in Theorem 3.4, the regret*

$$R_t(\boldsymbol{\omega}) = \sum_{l=1}^{t}\frac{1}{n_l}\left[\sum_{i=1}^{n_l}\rho_\tau(y_{il} - {x_{il}^*}^\top \overline{\boldsymbol{\beta}}_l^{\mathrm{QR}}) - \sum_{i=1}^{n_l}\rho_\tau(y_{il} - {x_{il}^*}^\top \boldsymbol{\omega})\right] \qquad \boldsymbol{\omega} \in \mathbb{R}^{p+1}$$

of the `Online-QR` learning procedure grows at a rate of $O(\sqrt{t})$.

For the online learning procedure with Lipschitz loss, [32] showed that the lower bound of the regret growth rate is $O(\sqrt{t})$. By utilizing the additive inequality of the check loss function (see Lemma 1 in [66]), we can prove the regret of our `Online-QR` learning procedure also grows at the rate of $O(\sqrt{t})$.

3.5.1 Future directions

Our proposed methods may be further extended to several promising directions, including but not limited to:

(i) We currently estimate coefficients in the normal linear regression model (3.12) by ordinary least squares (OLS) for its simplicity and computational efficiency.

In general, any consistent linear regression estimation techniques can be considered. For example, it is tempting to use weighted least squares (WLS) for the heterogeneity in the model's error term. However, we found using the WLS resulted in lower estimation accuracy compared to the OLS. The choice of the regression estimator still deserves more careful study in our future work. Second, possible improvement might be done to the choice of the penalty parameter λ_t in the nonzero centered ridge regression (3.19). In this chapter, we set λ_t as a function of the data block size to guarantee the problem has a closed-form solution. Potentially, a better calibrated penalty parameter, e.g., letting λ_t reflects the uncertainty in the `Online-QR` estimate at time point $(t-1)$, may further improve the estimation accuracy of `Online-QR`. Third, `Online-QR` might give conservative confidence intervals at the extreme quantile $\tau = 0.9$. In the QR literature, this is a situation that often requires special treatment, such as using extreme value theory in [41] or the correction in [192].

(ii) As the idea in our proposed `Online-QR` is fairly general, it may possibly further extend to other variants of QR, such as censored QR [191, 83] and Tobit QR [195, 89]. Another intriguing topic is to consider `Online-QR` for high-dimensional QR analysis of stream data. However, as sparsity is usually assumed for high-dimensional regression, the prior choice would require more careful consideration. Then it is nontrivial how to adapt the Bayesian formulation into the `Online-QR` framework and meanwhile maintain the computational simplicity.

(iii) Although the popular stochastic gradient descent (SGD) is not applicable to QR problems due to the non-smooth objective, it is tempting to consider its variant based on sub-gradient, due to the fact that it does not require any matrix inversion. However, this algorithm can be numerically unstable and sensitive to the choice of error distributions, likely because the sub-gradient does not necessarily correspond to a real gradient ascent direction [193]. Alternatively, one may consider applying SGD to a smoothed approximation to the check loss, e.g., the conquer method [84]. Potential challenge with this approach is the selection of smoothing parameter when facing the storage and speed challenges in stream data. The idea in [166] for choosing local smoothing parameter in distributed QR might provide valuable insight on solving this issue. Further exploration along this direction may also prove useful for solving high-dimensional QR and censored QR in stream data based on the works of [83], [167], [129], and [132].

3.6 Deferred proofs

3.6.1 *Proof of Theorem 3.1*

Let $(Y_i)_{i\geq 1}$ be a sequence of i.i.d. and centered random variables. Denote $\sigma^2 = \mathbb{E}Y_1^2$ and $T_n = \sum_{i=1}^n Y_i$. Cramér [43] has established the following asymptotic expansion on the probabilities of moderate deviations for T_n.

Lemma 3.7. *Assume that* $\mathbf{E}e^{\lambda|Y_1|} < \infty$ *for a constant* $\lambda > 0$. *Then it holds*

$$\log \frac{\mathbb{P}(T_n > x\sigma\sqrt{n})}{1-\Phi(x)} = O\left(\frac{1+x^3}{\sqrt{n}}\right) \quad as \ \ n \to \infty, \tag{3.26}$$

uniformly for $0 \leq x = o(n^{1/2})$.

The Cramér moderate deviations have attracted a lot of interests. We refer to Petrov [147], Beknazaryan, Sang and Xiao [21], Fan, Hu and Ma [64], Fan, Hu and Xu [65] for more such type results.

We first give a proof for the case $R_n(x,p)$. For all $t \geq 0$, it is easy to see that

$$\begin{aligned}\mathbb{P}\big(R_n(x,p) \geq t\big) &= \mathbb{P}\left(\frac{\sqrt{n}f(x_p)(x_{n,p}-x_p)}{\sqrt{p(1-p)}} \geq t\right)\\ &= \mathbb{P}\left(x_{n,p}-x_p \geq \frac{t\sqrt{p(1-p)}}{\sqrt{n}f(x_p)}\right)\end{aligned}$$

Write $x_{n,p,t} = x_p + \frac{t\sqrt{p(1-p)}}{\sqrt{n}f(x_p)}$. Then, by the definition of $x_{n,p,t}$, we get for all $t \geq 0$,

$$\begin{aligned}\mathbb{P}\big(R_n(x,p) \geq t\big) &= \mathbb{P}\big(x_{n,p} \geq x_{n,p,t}\big) \ = \ \mathbb{P}\big(p \geq F_n(x_{n,p,t})\big)\\ &= \mathbb{P}\Big(np \geq \sum_{i=1}^n \mathbf{1}(X_i \leq x_{n,p,t})\Big)\\ &= \mathbb{P}\Big(np - nF(x_{n,p,t}) \geq \sum_{i=1}^n \big(\mathbf{1}(X_i \leq x_{n,p,t}) - F(x_{n,p,t})\big)\Big)\\ &= \mathbb{P}\left(\sum_{i=1}^n \frac{\mathbf{1}(X_i \leq x_{n,p,t}) - F(t_{n,p}(x))}{\sqrt{nF(x_{n,p,t})(1-F(x_{n,p,t}))}} \leq \frac{\sqrt{n}(p-F(x_{n,p,t}))}{\sqrt{F(x_{n,p,t})(1-F(x_{n,p,t}))}}\right)\end{aligned}$$

Recall that $F'(x_p) = f(x_p)$ and that $F''(x) = f'(x)$ is bounded in a neighborhood of $x = x_p$. Thus, it holds uniformly for $0 \leq t = o(\sqrt{n})$,

$$\begin{aligned}p - F(x_{n,p,t}) &= F(x_p) - F(x_{n,p,t})\\ &= -\frac{t}{\sqrt{n}}\sqrt{p(1-p)} + O\left(\frac{t^2}{n}\right)\end{aligned}$$

From the last line, we deduce that uniformly for $0 \leq t = o(\sqrt{n})$,

$$F(x_{n,p,t})\big(1-F(x_{n,p,t})\big) = p(1-p)+\frac{(1-2p)t}{\sqrt{np(1-p)}}p(1-p)+O\Big(\frac{t^2}{n}\Big) \quad (3.27)$$

Hence, we have uniformly for $0 \le t = o(\sqrt{n})$,

$$\frac{\sqrt{n}(p-F(x_{n,p,t}))}{\sqrt{F(x_{n,p,t})(1-F(x_{n,p,t}))}} = -t+O\Big(\frac{t^2}{\sqrt{np(1-p)}}\Big)$$

Therefore, we deduce that uniformly for $0 \le t = o(\sqrt{n})$,

$$\mathbb{P}\big(R_n(x,p) \ge t\big) = \mathbb{P}\Bigg(\sum_{i=1}^n \frac{\mathbf{1}(X_i \le x_{n,p,t}) - F(t_{n,p}(x))}{\sqrt{nF(x_{n,p,t})(1-F(x_{n,p,t}))}} \le -t+O\Big(\frac{t^2}{\sqrt{np(1-p)}}\Big)\Bigg)$$

Denote $Z_i = \mathbf{1}(X_i \le x_{n,p,t}) - F(t_{n,p}(x)), 1 \le i \le n$. Notice that $(Z_i)_{1 \le i \le n}$ are i.i.d. and centered random variables, and satisfy that for all $1 \le i \le n$,

$$|Z_i| \le 1 \quad \text{and} \quad \mathrm{Var}(Z_i) = F(x_{n,p,t})\big(1-F(x_{n,p,t})\big)$$

The last line and (3.27) implies that

$$\sum_{i=1}^n \mathrm{Var}(Z_i) = nF(x_{n,p,t})\big(1-F(x_{n,p,t})\big) = np(1-p)+O\big(\sqrt{n}\big) \quad (3.28)$$

By Lemma 3.7 and (3.28), we obtain that it holds

$$\log\frac{\mathbb{P}\big(R_n(x,p) \ge t\big)}{1-\Phi(t)} = O\Big(\frac{1+t^3}{\sqrt{n}}\Big) \text{ as } n \to \infty \quad (3.29)$$

uniformly for $0 \le t = o(n^{1/2})$, which gives the desired inequality for $R_n(x,p)$. For $-R_n(x,p)$, the desired inequality follows by a similar argument. □

3.6.2 Proof of Theorem 3.2

Recall that $(Y_i)_{i \ge 1}$ is a sequence of i.i.d. and centered random variables and $T_n = \sum_{i=1}^n Y_i$. Assume that $\mathbf{E}e^{\lambda|Y_1|} < \infty$ for a constant $\lambda > 0$. Denote

$$\Lambda^*(x) = \sup_{\lambda \ge 0}\{\lambda x - \log \mathbf{E}e^{\lambda Y_1}\}$$

the Fenchel-Legendre transform of the cumulant function of Y_1. The function $\Lambda^*(x)$ is known as the good rate function in LDP theory, see Dembo and Zeitouni [50]. Bahadur and Rao [19] have established the following sharp large deviations.

Lemma 3.8. *Assume that $\mathbf{E}e^{\lambda|Y_1|} < \infty$ for a constant $\lambda > 0$. For any $y > 0$, let τ_y and σ_y be the positive solutions of the following equations:*

$$h'(\tau_y)=0 \quad and \quad \sigma_y=\sqrt{-h''(\tau_y)}$$

where $h(\tau)=\tau y-\log \mathbf{E}e^{\tau Y_1}$. Then for a given positive constant y, it holds

$$\mathbb{P}\left(\frac{T_n}{n}>y\right)=\frac{e^{-n\Lambda^*(y)}}{\sigma_y\,\tau_y\sqrt{2\pi n}}\Big[1+o(1)\Big], \qquad n\to\infty \tag{3.30}$$

Such type large deviations have attracted a lot of attentions. We refer to Bercu and Rouault [23], Joutard [92], Fan, Grama and Liu [62], Li [113] for more such type results.

We are in position to prove Theorem 3.2. For all $t\geq 0$, it holds

$$\begin{aligned}\mathbb{P}\big(x_{n,p}-x_p\geq t\big)&=\mathbb{P}\big(x_{n,p}\geq x_p+t\big)\;=\;\mathbb{P}\big(p\geq F_n(x_p+t)\big)\\&=\mathbb{P}\Big(n(1-p)\leq\sum_{i=1}^n\mathbf{1}(X_i>x_p+t)\Big)\\&=\mathbb{P}\left(n(F(x_p+t)-p)\leq\sum_{i=1}^n U_i\right)\end{aligned}$$

where

$$U_i=\mathbf{1}(X_i>x_p+t)-1+F(x_p+t), \qquad i=1,\ldots,n$$

Notice that $(U_i)_{1\leq i\leq n}$ are i.i.d. and centered random variables with $|U_i|\leq 1$. By some simple calculations, it is easy to see that

$$\begin{aligned}\mathbf{E}e^{\lambda U_1}&=e^{\lambda F(x_p+t)}\big(1-F(x_p+t)\big)+e^{\lambda(F(x_p+t)-1)}F(x_p+t)\\&=\exp\left\{\lambda F(x_p+t)+\log\Big(1-F(x_p+t)+e^{-\lambda}F(x_p+t)\Big)\right\}\end{aligned}$$

Therefore, we have

$$\begin{aligned}\Lambda^*(F(x_p+t)-p)&=\sup_{\lambda\geq 0}\Big\{\lambda(F(x_p+t)-p)-\log\mathbf{E}e^{\lambda U_1}\Big\}\\&=\sup_{\lambda\geq 0}\Big\{-\lambda p-\log\Big(1-F(x_p+t)+e^{-\lambda}F(x_p+t)\Big)\Big\}\\&=\Lambda^+(t)\end{aligned}$$

where

$$\Lambda^+(t)=p\log\frac{p}{F(x_p+t)}+(1-p)\log\frac{1-p}{1-F(x_p+t)}$$

Denote

$$h_1(\tau)=\tau(F(x_p+t)-p)-\log\mathbf{E}e^{\tau U_1}$$

Let τ_t^+ and σ_t^+ be the positive solutions of the following equations:

$$h_1'(\tau_t^+)=0 \quad \text{and} \quad \sigma_t^+=\sqrt{-h_1''(\tau_t^+)}$$

Then we have

$$\tau_t^+ = \log \frac{F(x_p+t)(1-p)}{p(1-F(x_p+t))} \quad \text{and} \quad \sigma_t^+ = \sigma_p = \sqrt{p(1-p)}$$

Applying Lemma 3.8 to $(U_i)_{1\le i\le n}$, with

$$Y_i = U_i, \quad y = F(x_p+t)-p, \quad \tau_t = \tau_t^+ \quad \text{and} \quad \sigma_t = \sigma_p$$

we get for any $t>0$,

$$\mathbb{P}\big(x_{n,p}-x_p \ge t\big) = \frac{1}{\sigma_p\,\tau_t^+\sqrt{2\pi n}} \exp\Big\{-n\Lambda^+(t)\Big\}\Big[1+o(1)\Big], \quad n\to\infty$$

which gives the first desired equality. An argument in symmetry gives the second desired equality.

3.6.3 *Proof of Theorem 3.4*

We first show that the `Online-QR` estimator can achieve the unbiasedness after finite time, and then prove its asymptotic normality as $t\to\infty$. Finally, we give a recursive formula for the update of its asymptotic covariance matrix.

3.6.3.1 The unbiasedness of the `Online-QR` estimator

From the update rule of `Online-QR`, we see that $\Big\{\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} \mid t=1,2,\dots\Big\}$ actually forms a time-homogeneous Markov process. According to Theorem 8.2.14 in [163], under the following three conditions: (i) the irreducibility of the process; (ii) the geometric drift of the process to the center; (iii) the uniform integrability of the marginal densities of the process, a time-homogeneous Markov process is stationary. Condition (iii) follows from the general argument laid out in [163]. Next, we first verify conditions (i) and (ii), and then prove the unbiasedness of `Online-QR` under the stationarity of this Markov process.

From (3.19), $\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}$ is obtained by solving a penalized linear regression problem, and thus the state space S of the process is a compact subset of $\mathbb{R}^{p+1}$. To verify the irreducibility, we only need to show that with positive probability any $\boldsymbol{\beta}'\in S$ is reachable from any $\boldsymbol{\beta}''\in S$ after finite time. Note that $\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}$ can be rewritten as

$$\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} = \frac{n_t}{\sum_{l=1}^t n_l}(X_t^{*\top}X_t^*)^{-1}\left[X_t^{*\top}\boldsymbol{y}_t^* + \frac{\sum_{l=1}^{t-1} n_l}{n_t} X_t^{*\top}X_t^*\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right]$$

from which we see that its value is constrained to the subspace spanned by the rows of X_t^*. Therefore, if $\bigcap_{l=t_0}^{t_0+t_0'} \mathrm{null}(X_l^*) = \mathbf{0}_{p+1}$ for $t_0, t' \in \mathbb{N}$ and t' is large enough, any value in S is reachable after finite time. Here $\mathrm{null}(X_l^*)$ represents the null space of the linear map induced by X_l^*. While if $\bigcap_{l=t_0}^{t_0+t'} \mathrm{null}(X_l^*) \neq \mathbf{0}_{p+1}$, the state space of the process is reducible to $S' = S \backslash \bigcap_{l=t_0}^{t_0+t'} \mathrm{null}(X_l^*)$, and any value in S' is reachable after finite time. To assess the geometric drift of the process to the center, we utilize

$$\begin{aligned}
\left\| \overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} \right\|_2 &= \left\| \frac{n_t}{\sum_{l=1}^t n_l} \widetilde{\boldsymbol{\beta}}_t^{\mathrm{QR}} + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right\|_2 \\
&\leq \frac{n_t}{\sum_{l=1}^t n_l} \left\| \widetilde{\boldsymbol{\beta}}_t^{\mathrm{QR}} \right\|_2 + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l} \left\| \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right\|_2 \leq \frac{c_1}{t} + c_2 \left\| \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right\|_2
\end{aligned} \tag{3.31}$$

where $c_1, c_2 \in \mathbb{R}$ are positive constants and $c_2 < 1$. From (3.31), we can conclude the tightness of $\{\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}} : t = 1, 2, \ldots\}$. This then implies that, after finite time t_0, the process can reach the stationarity.

Once the stationarity is reached, we have $\mathbb{E}(\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}) = \mathbb{E}(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}})$. In what follows, we denote this expectation as $\boldsymbol{\beta}^*$ and will show that it is equal to the true value of the regression parameter $\boldsymbol{\beta}$. By (3.17) and (3.20), we have

$$\begin{aligned}
\boldsymbol{\beta}^* &= \mathbb{E}\left[\frac{n_t}{\sum_{l=1}^t n_l} (X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \boldsymbol{y}_t^* + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right] \\
&= \frac{n_t}{\sum_{l=1}^t n_l} \mathbb{E}\left[(X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \boldsymbol{y}_t^* \right] + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l} \boldsymbol{\beta}^* \\
&= \frac{\xi_1 n_t}{|\widehat{\gamma}| \xi_2 \sum_{l=1}^t n_l} (X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \left(\frac{1 - 2\tau + 2\tau^2}{\tau(1-\tau)} \mathbf{1}_{n_t} - \mathbb{E}|\boldsymbol{y}_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| \right) + \frac{n_t}{\sum_{l=1}^t n_l} \boldsymbol{\beta} + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l} \boldsymbol{\beta}^*
\end{aligned} \tag{3.32}$$

where $\mathbf{1}_{n_t}$ and $\left| y_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right|$ respectively are the n_t-dimensional vectors composed by 1 and $|y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|$. For the i th element of $\mathbb{E}\left| y_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \right|$, we have

$$\begin{aligned}
&\mathbb{E}|y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| = \mathbb{E}\left[\mathbb{E}\left(|y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| \mid y_{it} \right) \right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\left| y_{it} - x_{it}^{*\top} \boldsymbol{\beta}^* \right| + \nabla h\left(\boldsymbol{\beta}'\right)^\top \left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} - \boldsymbol{\beta}^* \right) \mid y_{it} \right) \right] = \mathbb{E}\left| y_{it} - \boldsymbol{x}_{it}^{*\top} \boldsymbol{\beta}^* \right|
\end{aligned}$$

where $\nabla h\left(\boldsymbol{\beta}'\right)$ is the sub-gradient of the absolute value function at $\boldsymbol{\beta}'$, and $\boldsymbol{\beta}'$ lies between $\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}$ and $\boldsymbol{\beta}^*$. Under the assumption $\varepsilon_{it} \sim \mathrm{AL}(0, 1, \tau)$, the variable $w_{it} = y_{it} - x_{it}^{*\top} \boldsymbol{\beta}^* = x_{it}^{*\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \varepsilon_{it}$ follows $\mathrm{AL}\left(x_{it}^{*\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), 1, \tau\right)$ with density $f(w_{it}) = \tau(1-\tau) \exp\left\{ -\rho_\tau(w_{it} - \boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)) \right\}$. Then we have

$$\mathbb{E}|y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| = \mathbb{E}|w_{it}| = \int_0^{+\infty} w_{it} f(w_{it}) d_{w_{it}} - \int_{-\infty}^0 w_{it} f(w_{it}) d_{w_{it}} \tag{3.33}$$

Next, we evaluate the integrals in the right-hand side by separating into two cases, (i) $x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \geq 0$ and (ii) $x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \leq 0$. As the evaluations are done in a similar way, we mainly focus on case (i) in the following. The first integral in (3.33) can be written as

$$\int_0^{+\infty} w_{it} f(w_{it}) d_{w_{it}} = \int_0^{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} w_{it} f(w_{it}) d_{w_{it}} + \int_{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^{+\infty} w_{it} f(w_{it}) d_{w_{it}}$$

where the first term in the right-hand side is

$$\begin{aligned}
&\int_0^{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} w_{it} f(w_{it}) d_{w_{it}} \\
&= \int_0^{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} w_{it}\tau(1-\tau)\exp\left\{(1-\tau)\left[w_{it}-x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}} \\
&= \tau\left\{ w_{it}\exp\left\{(1-\tau)\left[w_{it}-x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\}\Big|_0^{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} - \int_0^{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} \exp\left\{(1-\tau)\left[w_{it}-\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}} \right\} \\
&= \tau\left\{ x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \frac{1}{1-\tau} + \frac{1}{1-\tau}\exp\left\{(\tau-1)x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\}\right\}
\end{aligned}$$

and the second term is

$$\begin{aligned}
&\int_{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^{+\infty} w_{it} f(w_{it}) d_{w_{it}} = \int_{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^{+\infty} w_{it}\tau(1-\tau)\exp\left\{-\tau\left[w_{it}-x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}} \\
&= (\tau-1)\left\{ w_{it}\exp\left\{-\tau\left[w_{it}-x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\}\Big|_{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^{+\infty} \right. \\
&\qquad\qquad \left. - \int_{x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}^{+\infty} \exp\left\{-\tau\left[w_{it}-x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}} \right\} \\
&= (1-\tau)\left\{ x_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \frac{1}{\tau}\right\}
\end{aligned}$$

Thus

$$\int_0^{+\infty} w_{it} f(w_{it}) d_{w_{it}} = \boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \xi_1 + \frac{\tau}{1-\tau}\exp\left\{(\tau-1)\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\}$$

Meanwhile, the second integral in (3.33) is

$$\begin{aligned}
&\int_{-\infty}^{0} w_{it} f(w_{it}) d_{w_{it}} = \int_{-\infty}^{0} w_{it}\tau(1-\tau)\exp\left\{(1-\tau)\left[w_{it}-\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}} \\
&= \tau\left\{ w_{it}\exp\left\{(1-\tau)\left[w_{it}-\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\}\Big|_{-\infty}^{0} - \int_{-\infty}^{0}\exp\left\{(1-\tau)\left[w_{it}-\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right]\right\} d_{w_{it}}\right\} \\
&= -\frac{\tau}{1-\tau}\exp\left\{(\tau-1)\boldsymbol{x}_{it}^{*\top}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right\}
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left|y_{it} - \boldsymbol{x}_{it}^{*}{}^{\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right| = \boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}) + \xi_1 + \frac{2\tau}{1-\tau}\exp\left\{(\tau-1)\boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*})\right\} \quad (3.34)$$

On the other hand, if $x_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}) \leq 0$, it is easy to find using similar techniques that

$$\mathbb{E}\left|y_{it} - \boldsymbol{x}_{it}^{*}{}^{\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right| = -x_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}) - \xi_1 + \frac{2(1-\tau)}{\tau}\exp\left\{\tau x_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*})\right\} \quad (3.35)$$

Substituting (3.34) and (3.35) into (3.32), and utilizing the fact that $0 < \tau < 1$, $\exp\left\{(\tau-1)\boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*})\right\} \leq 1$ for $\boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}) \geq 0$, and $\exp\left\{\tau\boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*})\right\} \leq 1$ for $\boldsymbol{x}_{it}^{*}{}^{\top}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}) \leq 0$, we then finally obtain $\boldsymbol{\beta}^{*} = \boldsymbol{\beta}$.

3.6.3.2 The asymptotic normality of the `Online-QR` estimator

By the update rule of `Online-QR`, we can write $\left(\overline{\boldsymbol{\beta}}_{t}^{\mathrm{QR}} - \boldsymbol{\beta}\right)$ as

$$\overline{\boldsymbol{\beta}}_{t}^{\mathrm{QR}} - \boldsymbol{\beta} = \sum_{l=2}^{t}\frac{n_l}{\sum_{m=1}^{t} n_m}(\widetilde{\boldsymbol{\beta}}_{l}^{\mathrm{QR}} - \boldsymbol{\beta}) + \frac{n_1}{\sum_{m=1}^{t} n_m}(\overline{\boldsymbol{\beta}}_{1}^{\mathrm{QR}} - \boldsymbol{\beta})$$

By (3.34) and (3.35), we have

$$\begin{aligned}
&\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_{l}^{\mathrm{QR}} - \boldsymbol{\beta} \mid \overline{\boldsymbol{\beta}}_{1}^{\mathrm{QR}}, \widetilde{\boldsymbol{\beta}}_{2}^{\mathrm{QR}}, \ldots, \widetilde{\boldsymbol{\beta}}_{(l-1)}^{\mathrm{QR}}\right) \\
&= \mathbb{E}\left[(X_l^{*}{}^{\top}X_l^{*})^{-1}X_l^{*}{}^{\top}(X_l^{*}\boldsymbol{\beta} + \varepsilon_l - \xi_1\widetilde{\boldsymbol{v}}_l) - \boldsymbol{\beta} \mid \overline{\boldsymbol{\beta}}_{1}^{\mathrm{QR}}, \widetilde{\boldsymbol{\beta}}_{2}^{\mathrm{QR}}, \ldots, \widetilde{\boldsymbol{\beta}}_{(l-1)}^{\mathrm{QR}}\right] \\
&= (X_l^{*}{}^{\top}X_l^{*})^{-1}X_l^{*}{}^{\top}\left[\frac{1-2\tau+2\tau^2}{\tau(1-\tau)}\mathbf{1}_{n_t} - \mathbb{E}\left(\left|\boldsymbol{y}_l - X_l^{*}\overline{\boldsymbol{\beta}}_{(l-1)}^{\mathrm{QR}}\right| \overline{\boldsymbol{\beta}}_{1}^{\mathrm{QR}}, \widetilde{\boldsymbol{\beta}}_{2}^{\mathrm{QR}}, \ldots, \widetilde{\boldsymbol{\beta}}_{(l-1)}^{\mathrm{QR}}\right)\right] \\
&= \mathbf{0}_{p+1} \qquad l = 2, 3, \ldots
\end{aligned}$$

where $\varepsilon_l = \left(\varepsilon_{1l}, \varepsilon_{2l}, \ldots, \varepsilon_{n_l l}\right)^{\top} \in \mathbb{R}^{n_l}$. Then if the initialization of `Online-QR` is chosen as the standard QR estimator based on $(X_1, \boldsymbol{y}_1)$, $(\widetilde{\boldsymbol{\beta}}_{l}^{\mathrm{QR}} - \boldsymbol{\beta})(l = 1, 2, \ldots)$ is a martingale difference sequence. Here we write $\overline{\boldsymbol{\beta}}_{1}^{\mathrm{QR}}$ as $\widetilde{\boldsymbol{\beta}}_{1}^{\mathrm{QR}}$ for simplicity. Thus, we can use the martingale central limit theorem to prove the asymptotic normality. Next, we verify the Lindeberg-type condition.

For any $z \in \mathbb{R}^{p+1}$ satisfying $\|z\|_2 = 1$ and $\alpha > 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left|\boldsymbol{z}^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right|^{2+\alpha} \\
&= \mathbb{E}\left|\boldsymbol{z}^\top\left[(X_l^{*\top} X_l^*)^{-1} X_l^{*\top}\left(\xi_1 \boldsymbol{v}_l + \xi_2\sqrt{\boldsymbol{v}_l}\circ \boldsymbol{u}_l - \xi_1 \widetilde{\boldsymbol{v}}_l\right)\right]\right|^{2+\alpha} \\
&\leq \left(2\lambda_{\min}^{-1}\sqrt{\lambda_{\max}}/\sqrt{n_l}\right)^{2+\alpha}\left[\mathbb{E}\left\|\xi_1\left(\boldsymbol{v}_l - \widetilde{\boldsymbol{v}}_l\right)\right\|_2^{2+\alpha} + \mathbb{E}\left\|\xi_2\sqrt{\boldsymbol{v}_l}\circ\boldsymbol{u}_l\right\|_2^{2+\alpha}\right] \\
&\leq c_3\left(2\lambda_{\min}^{-1}\sqrt{\lambda_{\max}}\right)^{2+\alpha}
\end{aligned}
\tag{3.36}
$$

where $c_3 \in \mathbb{R}$ is a constant, "$\circ$" represents the Hadamard product, and $u_l = \left(u_{1l}, u_{2l}, \ldots, u_{n_l l}\right) \in \mathbb{R}^{n_l}$. In (3.36), we utilize

$$
\mathbb{E}\left(v_{il} - \widetilde{v}_{il}\right) = \frac{1-2\tau+2\tau^2}{\tau(1-\tau)} - \mathbb{E}\left|y_{il} - x_{il}^{*\top}\overline{\boldsymbol{\beta}}_{(l-1)}^{\text{QR}}\right| = 0 \tag{3.37a}
$$

$$
\mathbb{E}\left(\sqrt{v_{il}}u_{il}\right) = \mathbb{E}\left(\sqrt{v_{il}}\right)\mathbb{E}\left(u_{il}\right) = 0 \tag{3.37b}
$$

$$
\operatorname{Var}\left(v_{il} - \widetilde{v}_{il}\right) \leq 2\operatorname{Var}\left(v_{il}\right) + 2\operatorname{Var}\left(\widetilde{v}_{il}\right) = 2 + 2\frac{\xi_1^2}{\widehat{\gamma}^2\xi_2^2}\operatorname{Var}\left(\left|y_{il} - x_{il}^{*\top}\overline{\boldsymbol{\beta}}_{(l-1)}^{\text{QR}}\right|\right) = O(1) \tag{3.37c}
$$

and

$$
\operatorname{Var}\left(\sqrt{v_{il}}u_{il}\right) = \mathbb{E}\left(\operatorname{Var}\left(\sqrt{v_{il}}u_{il} \mid v_{il}\right)\right) + \operatorname{Var}\left(\mathbb{E}\left(\sqrt{v_{il}}u_{il} \mid v_{il}\right)\right) = 1 \tag{3.38}
$$

The specific form of $\operatorname{Var}\left(\left|y_{il} - x_{il}^{*\top}\overline{\boldsymbol{\beta}}_{(l-1)}^{\text{QR}}\right|\right)$ is given in (3.42) in the third step of this proof. By (3.36), we can further obtain

$$
\begin{aligned}
&\frac{1}{t}\sum_{l=1}^{t}\mathbb{E}\left[\left|z^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right|^2 \mathbb{I}\left(\left|\boldsymbol{z}^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right| > \varepsilon\sqrt{t}\right)\right] \\
&\leq \frac{1}{t}\sum_{l=1}^{t}\varepsilon^{-\alpha}t^{-\frac{\alpha}{2}}\mathbb{E}\left[\left|\boldsymbol{z}^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right|^{2+\alpha}\mathbb{I}\left(\left|z^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right| > \varepsilon\sqrt{t}\right)\right] \\
&\leq \varepsilon^{-\alpha}t^{-\frac{\alpha}{2}-1}\sum_{l=1}^{t}\mathbb{E}\left|\boldsymbol{z}^\top(\widetilde{\boldsymbol{\beta}}_l^{\text{QR}} - \boldsymbol{\beta})\right|^{2+\alpha} \to 0 \quad \text{as} \quad t \to \infty
\end{aligned}
$$

for any $\varepsilon > 0$, and thus $\overline{\boldsymbol{\beta}}_t^{\text{QR}}$ is asymptotically normal as we have $\frac{n_l}{\sum_{m=1}^{t} n_m} = O\left(\frac{1}{t}\right)$ under the assumption $\max_l n_l / \min_l n_l = O(1)$.

3.6.3.3 The asymptotic covariance matrix of the `Online-QR` estimator

Denote $\operatorname{Var}(\overline{\boldsymbol{\beta}}_t^{\text{QR}})$ as $\boldsymbol{\Sigma}_t$. By (3.17) and (3.20), we have

$$
\boldsymbol{\Sigma}_t = \operatorname{Var}\left(\mathbb{I}_1\right) + \operatorname{Var}\left(\mathbb{I}_2\right) + \operatorname{Var}\left(\mathbb{I}_3\right) - 2\operatorname{Cov}\left(\mathbb{I}_1 + \mathbb{I}_2, \mathbb{I}_3\right) \tag{3.39}
$$

where $\mathbb{I}_1 = \frac{n_t}{\sum_{l=1}^{t} n_l}(X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \boldsymbol{y}_t, \mathbb{I}_2 = \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^{t} n_l} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}$ and $\mathbb{I}_3 = \frac{n_t \xi_1}{\sum_{l=1}^{t} n_l}(X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \widetilde{\boldsymbol{v}}_t$. For the first two terms on the right-hand side of (3.39), we have

$$\begin{aligned}\operatorname{Var}(\mathbb{I}_1) &= \frac{n_t^2}{\left(\sum_{l=1}^{t} n_l\right)^2} \cdot \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2} (X_t^{*\top} X_t^*)^{-1} X_t^{*\top} X_t^* (X_t^{*\top} X_t^*)^{-1} \\ &= \frac{n_t^2}{\left(\sum_{l=1}^{t} n_l\right)^2} \cdot \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2} (X_t^{*\top} X_t^*)^{-1}\end{aligned} \tag{3.40}$$

and

$$\operatorname{Var}(\mathbb{I}_2) = \frac{\left(\sum_{l=1}^{t-1} n_l\right)^2}{\left(\sum_{l=1}^{t} n_l\right)^2} \Sigma_{t-1} \tag{3.41}$$

while the third term on the right-hand side of (3.39) is

$$\begin{aligned}\operatorname{Var}(\mathbb{I}_3) &= \frac{n_t^2 \xi_1^2}{\left(\sum_{l=1}^{t} n_l\right)^2} (X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \operatorname{Var}(\widetilde{\boldsymbol{v}}_t) X_t^* (X_t^{*\top} X_t^*)^{-1} \\ &= \frac{n_t^2}{\left(\sum_{l=1}^{t} n_l\right)^2} \cdot \frac{\xi_1^2}{\widetilde{\gamma}^2 \xi_2^2} (X_t^{*\top} X_t^*)^{-1} X_t^{*\top} \operatorname{Var}(|\boldsymbol{y}_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|) X_t^* (X_t^{*\top} X_t^*)^{-1}\end{aligned}$$

Next, we first analyze the diagonal elements of $\operatorname{Var}(|\boldsymbol{y}_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|)$, and then give its off-diagonal elements. For the i th diagonal element, we have

$$\operatorname{Var}\left(|y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right) = \mathbb{E}\left(y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right)^2 - \mathbb{E}^2\left(\left|y_{it} - x_t^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right|\right)$$

in which

$$\begin{aligned}&\mathbb{E}(y_{it} - x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}})^2 = \mathbb{E}(y_{it})^2 + \mathbb{E}(x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}})^2 - 2\mathbb{E}(y_{it})\mathbb{E}(x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}) \\ &= \operatorname{Var}(y_{it}) + \mathbb{E}^2(y_{it}) + \operatorname{Var}(x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}) + \mathbb{E}^2(x_{it}^{*\top} \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}) - 2\left(x_{it}^{*\top} \boldsymbol{\beta} + \xi_1\right) x_{it}^{*\top} \boldsymbol{\beta} \\ &= \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2} + \left(x_{it}^{*\top} \boldsymbol{\beta} + \xi_1\right)^2 + x_{it}^{*\top} \Sigma_{t-1} x_{it}^* + \left(x_{it}^{*\top} \boldsymbol{\beta}\right)^2 - 2\left(x_{it}^{*\top} \boldsymbol{\beta} + \xi_1\right) x_{it}^{*\top} \boldsymbol{\beta} \\ &= \frac{2-6\tau+6\tau^2}{\tau^2(1-\tau)^2} + x_{it}^{*\top} \boldsymbol{\Sigma}_{t-1} x_{it}^*\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}\left(\left|y_{it}-\boldsymbol{x}_t^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right|\right)=\mathbb{E}\left[\mathbb{E}\left(\left|y_{it}-x_t^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right|y_{it}\right)\right]\\
&=\mathbb{E}\left[\mathbb{E}\left(\left|y_{it}-x_t^{*\top}\boldsymbol{\beta}\right|+\nabla h\left(\boldsymbol{\beta}'\right)^{\top}\left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}-\boldsymbol{\beta}\right)\mid y_{it}\right)\right]=\mathbb{E}\left|\varepsilon_{it}\right|\\
&=\int_0^{+\infty}\varepsilon_{it}f\left(\varepsilon_{it}\right)d_{\varepsilon_{it}}-\int_{-\infty}^{0}\varepsilon_{it}f\left(\varepsilon_{it}\right)d_{\varepsilon_{it}}\\
&=\int_0^{+\infty}\varepsilon_{it}\tau(1-\tau)\exp\left\{-\tau\varepsilon_{it}\right\}d_{\varepsilon_{it}}-\int_{-\infty}^{0}\varepsilon_{it}\tau(1-\tau)\exp\left\{(1-\tau)\varepsilon_{it}\right\}d_{\varepsilon_{it}}\\
&=(\tau-1)\left[\varepsilon_{it}\exp\left\{-\tau\varepsilon_{it}\right\}\Big|_0^{+\infty}-\int_0^{+\infty}\exp\left\{-\tau\varepsilon_{it}\right\}d_{\varepsilon_{it}}\right]\\
&\qquad-\tau\left[\varepsilon_{it}\exp\left\{(1-\tau)\varepsilon_{it}\right\}\Big|_{-\infty}^{0}-\int_{-\infty}^{0}\exp\left\{(1-\tau)\varepsilon_{it}\right\}d_{\varepsilon_{it}}\right]=\frac{1-2\tau+2\tau^2}{\tau(1-\tau)}
\end{aligned}
$$

Thus

$$
\operatorname{Var}\left(\left|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right|\right)=x_{it}^{*\top}\boldsymbol{\Sigma}_{t-1}x_{it}^{*}+\frac{1-2\tau-2\tau^2+8\tau^3-4\tau^4}{\tau^2(1-\tau)^2} \tag{3.42}
$$

For the (i,j) th off-diagonal element of $\operatorname{Var}(|\boldsymbol{y}_t-X_t^{*}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|)$, we have

$$
\begin{aligned}
&\operatorname{Cov}\left(|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|,|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right)\\
&=\mathbb{E}\left[\operatorname{Cov}\left(\sqrt{\left(y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right)^2},\sqrt{\left(y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right)^2}\mid y_{it},y_{jt}\right)\right]\\
&\quad+\mathbb{E}\left\{\left[\mathbb{E}\left(|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\mid y_{it}\right)-\mathbb{E}|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right]\right.\\
&\qquad\left.\cdot\left[\mathbb{E}\left(|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\mid y_{jt}\right)-\mathbb{E}|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right]\right\}
\end{aligned}
$$

in which

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{Cov}\left(\sqrt{\left(y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right)^2},\sqrt{\left(y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right)^2}\mid y_{it},y_{jt}\right)\right]\\
&=\mathbb{E}\left[\operatorname{Cov}\left(x_{it}^{*\top}\left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}-\boldsymbol{\beta}\right),\boldsymbol{x}_{jt}^{*\top}\left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}-\boldsymbol{\beta}\right)\mid y_{it},y_{jt}\right)\right]=\boldsymbol{x}_{it}^{*\top}\boldsymbol{\Sigma}_{t-1}x_{jt}^{*}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left\{\left[\mathbb{E}\left(|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\mid y_{it}\right)-\mathbb{E}|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right]\right.\\
&\qquad\left.\left[\mathbb{E}\left(|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\mid y_{jt}\right)-\mathbb{E}|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right]\right\}\\
&=\mathbb{E}\left[\left(|\varepsilon_{it}|-\mathbb{E}|\varepsilon_{it}|\right)\left(\left|\varepsilon_{jt}\right|-\mathbb{E}\left|\varepsilon_{jt}\right|\right)\right]=0
\end{aligned}
$$

Thus

$$
\operatorname{Cov}\left(|y_{it}-x_{it}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|,|y_{jt}-x_{jt}^{*\top}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right)=x_{it}^{*\top}\Sigma_{t-1}x_{jt}^{*} \tag{3.43}
$$

Combining (3.42) and (3.43), we obtain

$$\operatorname{Var}(|\boldsymbol{y}_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|) = X_t^* \boldsymbol{\Sigma}_{t-1} X_t^{*\,\top} + \frac{1-2\tau-2\tau^2+8\tau^3-4\tau^4}{\tau^2(1-\tau)^2} I_{n_t}$$

and further

$$\operatorname{Var}(\mathbb{I}_3) = \frac{n_t^2(1-2\tau)^2}{\left(\sum_{l=1}^t n_l\right)^2}\left[\Sigma_{t-1} + \frac{1-2\tau-2\tau^2+8\tau^3-4\tau^4}{\tau^2(1-\tau)^2}(X_t^{*\,\top} X_t^*)^{-1}\right] \tag{3.44}$$

Finally, for the fourth term in the right-hand side of (3.39), we have

$$\operatorname{Cov}(\mathbb{I}_1+\mathbb{I}_2, \mathbb{I}_3) = \mathbb{E}\left[\operatorname{Cov}(\mathbb{I}_1+\mathbb{I}_2, \mathbb{I}_3 \mid \boldsymbol{y}_t)\right] - \mathbb{E}\left\{\left[\mathbb{E}(\mathbb{I}_1+\mathbb{I}_2 \mid \boldsymbol{y}_t) - \mathbb{E}(\mathbb{I}_1+\mathbb{I}_2)\right]\left[\mathbb{E}(\mathbb{I}_3 \mid \boldsymbol{y}_t) - \mathbb{E}(\mathbb{I}_3)\right]^\top\right\}$$

in which

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Cov}(\mathbb{I}_1+\mathbb{I}_2, \mathbb{I}_3 \mid \boldsymbol{y}_t)\right] \\
=&\mathbb{E}\left[\operatorname{Cov}\left(\frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}, \frac{n_t \xi_1}{\left(\sum_{l=1}^t n_l\right)|\widehat{\gamma}|\xi_2}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top} |\boldsymbol{y}_t - X_t^* \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| \boldsymbol{y}_t\right)\right] \\
=&\frac{\left(\sum_{l=1}^{t-1} n_l\right) n_t}{\left(\sum_{l=1}^t n_l\right)^2} \cdot \frac{\xi_1}{|\widehat{\gamma}|\xi_2}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top} \mathbb{E}\left[\operatorname{Cov}\left(-\frac{y_t - X_t^* \boldsymbol{\beta}}{|\boldsymbol{y}_t - X_t^* \boldsymbol{\beta}|} \circ X_t^*\left(\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} - \boldsymbol{\beta}\right), \overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \mid \boldsymbol{y}_t\right)\right] \\
=&-\frac{\left(\sum_{l=1}^{t-1} n_l\right) n_t}{\left(\sum_{l=1}^t n_l\right)^2} \cdot \frac{\xi_1}{|\widehat{\gamma}|\xi_2} \mathbb{E}\left(\frac{\varepsilon}{|\varepsilon|}\right) \Sigma_{t-1} = -\frac{\left(\sum_{l=1}^{t-1} n_l\right) n_t (1-2\tau)^2}{\left(\sum_{l=1}^t n_l\right)^2} \Sigma_{t-1}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}(\mathbb{I}_1+\mathbb{I}_2 \mid \boldsymbol{y}_t) - \mathbb{E}(\mathbb{I}_1+\mathbb{I}_2) \\
=&\mathbb{E}\left(\frac{n_t}{\sum_{l=1}^t n_l}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top} \boldsymbol{y}_t + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}} \,\middle|\, y_t\right) \\
&-\mathbb{E}\left(\frac{n_t}{\sum_{l=1}^t n_l}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top} \boldsymbol{y}_t + \frac{\sum_{l=1}^{t-1} n_l}{\sum_{l=1}^t n_l}\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}\right) \\
=&\frac{n_t}{\sum_{l=1}^t n_l}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top}\left[\boldsymbol{y}_t - \mathbb{E}(\boldsymbol{y}_t)\right] = \frac{n_t}{\sum_{l=1}^t n_l}(X_t^{*\,\top} X_t^*)^{-1} X_t^{*\,\top}(\boldsymbol{\varepsilon}_t - \xi_1 \mathbf{1}_{n_t})
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{I}_3 \mid \boldsymbol{y}_t\right)-\mathbb{E}\left(\mathbb{I}_3\right)\\
=&\mathbb{E}\left(\left.\frac{n_t\xi_1}{\sum_{l=1}^t n_l}(X_t^{*\top}X_t^*)^{-1}X_t^{*\top}\widetilde{\boldsymbol{v}}_t\right| \boldsymbol{y}_t\right)-\mathbb{E}\left(\frac{n_t\xi_1}{\sum_{l=1}^t n_l}(X_t^{*\top}X_t^*)^{-1}X_t^{*\top}\widetilde{\boldsymbol{v}}_t\right)\\
=&\frac{n_t\xi_1}{\left(\sum_{l=1}^t n_l\right)|\widehat{\gamma}|\xi_2}(X_t^{*\top}X_t^*)^{-1}X_t^{*\top}\left[\mathbb{E}\left(|\boldsymbol{y}_t-X_t^*\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}| \mid \boldsymbol{y}_t\right)-\mathbb{E}|\boldsymbol{y}_t-X_t^*\overline{\boldsymbol{\beta}}_{t-1}^{\mathrm{QR}}|\right]\\
=&\frac{n_t(1-2\tau)}{\left(\sum_{l=1}^t n_l\right)}(X_t^{*\top}X_t^*)^{-1}X_t^{*\top}\left(|\varepsilon_t|-\frac{1-2\tau+2\tau^2}{\tau(1-\tau)}\mathbf{1}_{n_t}\right)
\end{aligned}$$

Besides, $\mathbb{E}\left[(\varepsilon_t-\xi_1\mathbf{1}_{n_t})\left(|\varepsilon_t|-\frac{1-2\tau+2\tau^2}{\tau(1-\tau)}\mathbf{1}_{n_t}\right)^\top\right]=\frac{(1-2\tau)^3}{\tau^2(1-\tau)^2}I_{n_t}$. Thus, we have

$$\operatorname{Cov}\left(\mathbb{I}_1+\mathbb{I}_2,\mathbb{I}_3\right)=-\frac{\left(\sum_{l=1}^{t-1}n_l\right)n_t(1-2\tau)^2}{\left(\sum_{l=1}^t n_l\right)^2}\Sigma_{t-1}+\frac{n_t^2(1-2\tau)^4}{\left(\sum_{l=1}^t n_l\right)^2\tau^2(1-\tau)^2}(X_t^{*\top}X_t^*)^{-1} \tag{3.45}$$

Combining (3.39), (3.40), (3.41) and (3.44), (3.45), we finally obtain that

$$\Sigma_t=\frac{n_t^2 g(\tau)}{\left(\sum_{l=1}^t n_l\right)^2}(X_t^{*\top}X_t^*)^{-1}+\frac{\left(\sum_{l=1}^{t-1}n_l\right)^2+n_t^2(1-2\tau)^2+2\left(\sum_{l=1}^{t-1}n_l\right)n_t(1-2\tau)^4}{\left(\sum_{l=1}^t n_l\right)^2}\Sigma_{t-1}$$

where $g(\tau)=\frac{1-2\tau+2\tau^2+(1-2\tau)^2\left(1-2\tau-2\tau^2+8\tau^3-4\tau^4\right)-2(1-2\tau)^4}{\tau^2(1-\tau)^2}$.

3.6.4 *Proof of Theorem 3.5*

By (3.25), we have

$$\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}-\boldsymbol{\beta}=\left[\prod_{l=1}^t(1-r_l)\right](\overline{\boldsymbol{\beta}}_0^{\mathrm{QR}}-\boldsymbol{\beta})-\sum_{l=1}^t r_l\left[\prod_{m=l+1}^t(1-r_m)\right](\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}) \tag{3.46}$$

where $r_t=\frac{n_t}{\sum_{l=1}^t n_l}$. Next, we separately analyze the two terms in the right-hand side of (3.46). Firstly, by using $[\prod_{l=1}^t(1-r_l)]=O\left(\frac{1}{t}\right)$, we have

$$\left\|\left[\prod_{l=1}^t(1-r_l)\right](\overline{\boldsymbol{\beta}}_0^{\mathrm{QR}}-\boldsymbol{\beta})\right\|_2=\left[\prod_{l=1}^t(1-r_l)\right]\left\|\overline{\boldsymbol{\beta}}_0^{\mathrm{QR}}-\boldsymbol{\beta}\right\|_2=O_p\left(\frac{1}{t}\right)$$

Secondly, as analyzed before, $(\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}-\boldsymbol{\beta})$, $l=1,2,\ldots$ forms a martingale difference sequence, thus to control the norm of the second term, we need to control $\left\|r_l\left[\prod_{m=l+1}^t(1-r_m)\right](\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}})\right\|_2$ and $\sum_{l=1}^t\mathbb{E}\left\|r_l\left[\prod_{m=l+1}^t(1-r_m)\right](\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}})\right\|_2$.

By utilizing (3.37a), (3.37b), (3.37c), (3.38), we have

$$\begin{aligned}
&\left\| r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}\right)\right\|_2\\
&= r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left\|(X_l^{*\top}X_l^{*})^{-1}X_l^{*\top}\left(\xi_1\boldsymbol{v}_l+\xi_2\sqrt{\boldsymbol{v}_l}\circ\boldsymbol{u}_l-\xi_1\widetilde{\boldsymbol{v}}_l\right)\right\|_2\\
&\le \left(\lambda_{\min}^{-1}\sqrt{\lambda_{\max}}/\sqrt{n_l}\right) r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left[\|\xi_1\left(\boldsymbol{v}_l-\widetilde{\boldsymbol{v}}_l\right)\|_2+\|\xi_2\sqrt{\boldsymbol{v}_l}\circ\boldsymbol{u}_l\|_2\right]=O_p\left(\frac{1}{t}\right)
\end{aligned}$$

Further, we can obtain $\sum_{l=1}^{t}\mathbb{E}\left\| r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}\right)\right\|_2 = O\left(\frac{1}{t}\right)$. Then by applying the Pinelis-Bernstein inequality (see Proposition A.3 in [168]), we have

$$\left\|\sum_{l=1}^{t} r_l \left[\prod_{m=l+1}^{t}(1-r_m)\right]\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}_l^{\mathrm{QR}}\right)\right\| = O_p\left(\frac{1}{t}\right)$$

and thus $\left\|\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}-\boldsymbol{\beta}\right\|_2 = O_p\left(\frac{1}{t}\right)$. That is, the `Online-QR` estimator $\overline{\boldsymbol{\beta}}_t^{\mathrm{QR}}$ enjoys a linear convergence rate of $O\left(\frac{1}{t}\right)$.

3.6.5 *Proof of Theorem 3.6*

For the regret of `Online-QR`, we have

$$\begin{aligned}
R_t(\boldsymbol{\omega}) &= \sum_{l=1}^{t}\frac{1}{n_l}\left[\sum_{i=1}^{n_l}\rho_\tau(y_{il}-x_{il}^{*\top}\overline{\boldsymbol{\beta}}_l^{\mathrm{QR}})-\sum_{i=1}^{n_l}\rho_\tau(y_{il}-x_{il}^{*\top}\boldsymbol{\omega})\right]\\
&\le \sum_{l=1}^{t}\frac{1}{n_l}\sum_{i=1}^{n_l}\rho_\tau(x_{il}^{*\top}\boldsymbol{\omega}-x_{il}^{*\top}\overline{\boldsymbol{\beta}}_l^{\mathrm{QR}}) \qquad (3.47)\\
&\le \max(\tau,1-\tau)\sum_{l=1}^{t}\frac{1}{n_l}\sum_{i=1}^{n_l}\left|x_{il}^{*\top}\boldsymbol{\omega}-x_{il}^{*\top}\overline{\boldsymbol{\beta}}_l^{\mathrm{QR}}\right|\\
&\le \max(\tau,1-\tau)\frac{\sqrt{2}\max_l(n_t/n_l)}{n_t}\sqrt{\sum_{l=1}^{t}\left\|X_l^{*}(\boldsymbol{\omega}-\overline{\boldsymbol{\beta}}_l^{\mathrm{QR}})\right\|_2^2} \qquad (3.48)\\
&\le \max(\tau,1-\tau)\frac{\sqrt{2}\max_l(n_t/n_l)}{n_t}\sqrt{\sum_{l=1}^{t}\lambda_{\max}n_l\|\boldsymbol{\omega}-\overline{\boldsymbol{\beta}}_l^{\mathrm{QR}}\|_2^2}=O_p(\sqrt{t})
\end{aligned}$$

In (3.47), we utilize the additive inequality of the check loss function (see Lemma 1 in [66]), and in (3.48), we use the inequality between the l_1-norm and l_2-norm of a vector and the assumption $\max_l n_l/\min n_l = O(1)$.

3.7 Conclusion

In this chapter, we proposed an efficient online algorithm, `Online-QR`, for QR estimation in stream data, and further extended it to the `Multiple-QR` model. Our proposal is motivated by the Gibbs sampler for Bayesian QR and the iteration in the SA method for GLMM.

We converted the non-smooth check loss optimization to a normal linear regression estimation problem, and greatly improved the local processing speed because solving the converted problem only requires computing a least squares estimator. It is worth noting that the proposed `Online-QR` algorithm can be applied beyond the context of stream data. For example, when analyzing static big data stored distributedly over a network, standard divide-and-conquer techniques require the existence of a central node to perform the aggregation. However, in modern networks, this type of centralized computing is often a weakness to cyber attacks and decentralized computing is more preferred. Under such a situation, our `Online-QR` algorithm may serve as an efficient way to perform QR analysis when coupled with a smart way to traverse the network.

Chapter 4
ROOT-SGD: Sharp Nonasymptotics and Near-Optimal Asymptotics in a Single Algorithm

We study the problem of solving strongly convex and smooth unconstrained optimization problems using stochastic first-order algorithms. We devise a novel algorithm, referred to as *Recursive One-Over-T SGD* (ROOT-SGD), based on an easily implementable, recursive averaging of past stochastic gradients. We prove that it simultaneously achieves state-of-the-art performance in both a finite-sample, nonasymptotic sense and an asymptotic sense. On the nonasymptotic side, we prove risk bounds on the last iterate of ROOT-SGD with leading-order terms that match the optimal statistical risk with a unity pre-factor, along with a higher-order term that scales at the sharp rate of $O(n^{-3/2})$ under the Lipschitz condition on the Hessian matrix. On the asymptotic side, we show that when a mild, one-point Hessian continuity condition is imposed, the rescaled last iterate of (multi-epoch) ROOT-SGD converges asymptotically to a Gaussian limit with the Cramér-Rao optimal asymptotic covariance, for a broad range of step-size choices.

Keywords:

Stochastic first-order optimization, nonasymptotic finite-sample convergence rate, asymptotic efficiency, local asymptotic minimax, Cramér-Rao lower bound, variance-reduced gradient method, Polyak-Ruppert-Juditsky (PRJ) procedure.

4.1 Introduction

Let θ^* denote the minimizer of F, and define the matrices

$$H^* := \nabla^2 F(\theta^*), \quad \text{and} \quad \Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top\right].$$

Under certain regularity assumptions, given a collection of n samples $(\xi_i)_{i\in[n]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, classical statistical theory guarantees that the minimizer $\widehat{\theta}_n^{\mathrm{ERM}} := \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^n f(\theta;\xi_i)$ of the associated empirical risk has the following asymptotic behavior:

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{ERM}} - \theta^*\right) \xrightarrow{d} \mathscr{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right). \tag{4.1}$$

Furthermore, the asymptotic distribution (4.1) is known to be locally asymptotic minimax, i.e. given a bowl-shaped loss function, the asymptotic risk of *any* estimator is lower bounded by the expectation under such a Gaussian distribution, in a suitably defined sequence of local neighborhoods. See, for example, [58] for a precise statement.

Unfortunately, the goals of rapid finite-sample convergence and optimal asymptotic behavior are in tension, and the literature has not yet arrived at a single algorithmic framework that achieves both goals simultaneously. Consider in particular two seminal lines of research:

(1) The Polyak-Ruppert-Juditsky (PRJ) procedure [150, 149, 155] incorporates slowly diminishing step-sizes into SGD, thereby achieving asymptotic normality with an optimal covariance matrix (and unity pre-factor). This meets the goal of calibrated uncertainty. However, the PRJ procedure is *not* optimal from a nonasymptotic point of view: rather, it suffers from large high-order nonasymptotic terms and fails to achieve the optimal sample complexity in general [15].

(2) On the other hand, variance-reduced stochastic optimization methods have been designed to achieve reduced sample complexity that is the sum of a *statistical error* and an *optimization error* [107, 160, 91, 111, 48]. These methods yield control on the optimization error, with sharp nonasymptotic rates of convergence, but the guarantees for the statistical error term yield an asymptotic behavior involving constant pre-factors that are strictly greater than unity, and is hence sub-optimal.

An open question:

Given this state of affairs, we are naturally led to the following question: can a single stochastic optimization algorithm simultaneously achieve optimal asymptotic and nonasymptotic guarantees? In particular, we would like such guarantees to enjoy the fine-grained statistical properties satisfied by the empirical risk minimizer, for a commensurate set of assumptions on the function F and the observations $f(\cdot;\xi)$ and including the same rate of decay of high-order terms.

In this chapter, we resolve this open question, in particular by proposing and analyzing a novel algorithm called *Recursive One-Over-T Stochastic Gradient Descent* (ROOT-SGD). It is very easy to describe and implement, and we prove that it is optimal in both asymptotic and nonasymptotic senses:

(1) On the nonasymptotic side, under suitable smoothness assumptions, we show that the estimator $\widehat{\theta}_n^{\mathrm{ROOT}}$ produced by the last iterate of the (multi-epoch) ROOT-SGD satisfies a bound of the following form:

$$\mathbb{E}\|\widehat{\theta}_n^{\mathrm{ROOT}} - \theta^*\|_2^2 \le \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{n} + O\left(\frac{1}{n^{3/2}}\right). \tag{4.2}$$

Note that the leading-order term of the bound (4.2) is exactly the squared norm of the Gaussian random vector in the local asymptotic minimax limit or Cramér-Rao lower bound, with unity pre-factor. Moreover, our bound is entirely nonasymptotic, valid for all finite n. We also prove that high-order term $O\big(n^{-3/2}\big)$ is unavoidable under a natural setup, and it improves upon existing $O(n^{-7/6})$ and $O(n^{-5/4})$ rates for the PRJ procedure [15, 189, 74]. We also derive similar bounds for the objective gap $F\big(\widehat{\theta}_n^{\mathrm{ROOT}}\big) - F(\theta^*)$ and the gradient norm $\|\nabla F\big(\widehat{\theta}_n^{\mathrm{ROOT}}\big)\|^2$.

(2) Furthermore, the nonasymptotic bound (4.2) holds true under a mild sample-size requirement. Indeed, given a L-smooth and μ-strongly convex population-level function F, and assuming that the noise $\varepsilon(\cdot;\xi) := \nabla f(\cdot;\xi) - \nabla F(\cdot)$ satisfies a stochastic Lipschitz condition with parameter ℓ_Ξ, the finite-sample bounds are viable as long as $n \gtrsim \frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}$. The first term $O\big(\frac{L}{\mu}\big)$ matches the iteration complexity of gradient descent, and the $O\big(\frac{\ell_\Xi^2}{\mu^2}\big)$ term is the sample complexity needed for distinguishing a μ-strongly convex function from a constant function. The high-order terms in Eq. (4.2) also depend on the parameters (μ, L, ℓ_Ξ) in a similar way. This exhibits the fast nonasymptotic convergence of our algorithm, matching state-of-the-art variance reduction algorithms.

(3) We also establish asymptotic guarantees for the ROOT-SGD algorithm. Assuming insetad a mild one-point Hessian continuity condition at the minimizer, for a broad range of step-size choices the last iterate $\widehat{\theta}_n^{\mathrm{ROOT}}$ converges in distribution to the optimal Gaussian law (4.1) whenever the Hessian matrix $\nabla^2 F$ is continuous at θ^*, a much weaker condition manifesting the difference between ROOT-SGD and the Polyak-Ruppert averaging procedure [150].

Notably, both the MSE bound of the form (4.2) and the asymptotic normality are fine-grained guarantees that are satisfied by the empirical risk minimizer, under comparable assumption posed on the continuity of Hessian matrix. To the best of our knowledge, such guarantees have not been available heretofore in the literature on stochastic optimization. The ROOT-SGD algorithm proposed in this chapter achieves these guarantees not only simultaneously, but also with sharp nonasymptotic sample complexity.

4.1.1 Additional related work

SGD and Polyak-Ruppert-Juditsky averaging procedure

The theory of the stochastic approximation method has a long history since its birth in the 1950s [153, 29, 201, 139, 27, 30] and recently regains its attention due to its superb performance in real-world application practices featured by deep learning [80], primarily due to its exceptional handling of the online samples. Classics on this topic include [24, 22, 124, 26] and many more. Specially on the study of asymptotic normality which can trace back to [61], the general idea of iteration averaging is based on the analysis of two-time-scale iteration techniques and it achieves asymptotic normality with an optimal covariance [155, 149, 150]. Recent work along this line includes [15, 16, 14, 49, 71, 52, 58, 54, 7, 53, 12], presenting attractive asymptotic and nonasymptotic properties under a variety of settings and assumptions. [1, 183] provide minimax lower bounds for stochastic first-order algorithms. [88, 86, 87] analyze SGD and its acceleration with *tail-averaging* that simultaneously achieves exponential forgetting and optimal statistical risk up to a constant, nonunity pre-factor. It is also worth mentioning that iteration averaging provides robustness and adaptivity [112]. Instead of averaging the iterates, our ROOT-SGD algorithm averages the past stochastic gradients with proper de-biasing corrections and achieves competitive asymptotic performance.[1] For statistical inferential purposes, recent work [37, 164] presents confidence interval assertions via online stochastic gradient with Polyak-Ruppert-Juditsky averaging procedure; analogous results for the ROOT-SGD algorithm are hence worth exploring, building upon the asymptotic normality that our work has established.

Variance-reduced gradient methods

In the field of smooth and convex stochastic optimization, variance-reduced gradient methods represented by, but not limited to, SAG [107], SDCA [160], SVRG [91, 98, 13, 111], SAGA [48], SARAH [144, 146] have been proposed to improve the theoretical convergence rate of (stochastic) gradient descent. Accelerated variants of SGD provide further improvements in convergence rate [122, 159, 6, 105, 100, 104]. More recently, a line of work on recursive variance-reduced stochastic first-order algorithms have been studied in the nonconvex stochastic optimization literature [67, 203, 180, 146, 148, 120]. These algorithms, as well as their hybrid siblings [46, 171], achieve optimal iteration complexities for an appropriate class of nonconvex functions and in particular are faster than SGD under mild additional smoothness assumption on the stochastic gradients and Hessians [10]. Limited by space, we refer interested readers to a recent survey article by [81], and while our ROOT-SGD algorithm can be viewed as a variant of variance-reduced algorithms, our goal is substantially different: we aim to establish for strongly convex objectives both a sharp, unity pre-factor nonasymptotic bound and asymptotic normality with Cramér-Rao optimal asymptotic covariance that matches the local asymptotic minimax optimality [58].

[1] A related but fundamentally different idea was proposed in [141, 186, 110] called *dual averaging* for optimizing the regularized objectives. In contrast to their method, we focus in this chapter on the smooth objective setting and augment our estimator with de-biasing corrections. See also [58, 172] for more on first-order optimization methods on Riemannian manifolds.

Sharp nonasymptotics and asymptotic efficiency

When the objective admits additional smoothness, nonasymptotic rate analyses for either SGD with iteration averaging or variance-reduced stochastic first-order algorithms have been studied in various settings. [15] presents a nonasymptotic analysis of SGD with PRJ averaging procedure showing that, after processing n samples, the algorithm achieves a nonasymptotic rate that matches the Cramér-Rao lower bound with a pre-factor equal to one with the additional term being $O(n^{-7/6})$ (see the discussions in §4.5). [189, 74] improves the additional term to $O(n^{-5/4})$ under comparable assumptions. [49, 52, 58, 12] achieves either sharp nonasymptotic bounds (in the quadratic case) or asymptotic efficiency that matches the local asymptotic minimax lower bound. The asymptotic efficiency of variance-reduced stochastic approximation methods, however, has been less studied. More related to work in this chapter, [73] establishes the nonasymptotic upper bounds on the objective gap for an online variant of the SVRG algorithm [91], where the leading-order nonasymptotic bound on the excess risk matches the optimal asymptotic behavior of the empirical risk minimizer under certain *self-concordant condition* posed on the objective function; the additional higher-order term reported is at least $\Omega(n^{-8/7})$.

Other related work

[103, 134] studies fixed-constant-step-size linear stochastic approximation with PRJ averaging procedure beyond an optimization algorithm [150], which includes many interesting applications in minimax game and reinforcement learning. To be specific, [103] provides general nonasymptotic bounds which suffer from a nonunity pre-factor on the optimal statistical risk, and [134] studies the PRJ averaging procedure for general linear stochastic approximation and precisely characterizes the asymptotic limiting Gaussian distribution, delineating the additional term that adds onto the Cramér-Rao asymptotic covariance and which vanishes as $\eta \to 0$ [53], and further establishes sharp concentration inequalities under stronger moment conditions on the noise. [11] proposes an extrapolation-smoothing scheme of *Implicit Gradient Transportation* to reduce the variance of the algorithm and provides convergence rates for quadratic objectives, which is further generalized to nonconvex optimization to improve the convergence rate of normalized SGD [45]. For the policy evaluation problem in reinforcement learning, [95] establishes an instance-dependent non-asymptotic upper bound on the ℓ_∞ estimation error, for a variance-reduced stochastic approximation algorithm. Their bound matches the risk of optimal Gaussian limit up to constant or logarithmic factors. Recently, [133] extends the algorithmic idea in this chapter and proposes the recursive variance-reduced stochastic approximation in span seminorm, which is applicable for generative models in reinforcement learning.

Organization

The rest of the chapter is organized as follows. We present the ROOT-SGD algorithm in §4.2, and delineate the asymptotic normality and nonasymptotic upper bounds in §4.3. Proof of Theorem 4.1 and extended analysis is provided in §4.4. We provides additional discussion on comparison of our results with concurrent work in §4.5. We present our conclusions in §4.6. Full proofs and discussions are provided in the appendix.

Notations

Given a pair of vectors $u, v \in \mathbb{R}^d$, we write $\langle u, v \rangle$ for the inner product, and $\|v\|_2$ for the Euclidean norm. For a matrix M, the ℓ_2-operator norm is defined as $|||M|||_{\text{op}} :$ $= \sup_{\|v\|_2=1} \|Mv\|_2$. For scalars $a, b \in \mathbb{R}$, we adopt the shorthand notation $a \wedge b :=$ $\min(a,b)$ and $a \vee b := \max(a,b)$. Throughout the chapter, we use the σ-fields $\mathscr{F}_t :=$ $\sigma(\xi_1, \xi_2, \cdots, \xi_t)$ for any $t \geq 0$. Unless indicated otherwise, C denotes some positive, universal constant whose value may change at each appearance. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive scalars, we denote $a_n \gtrsim b_n$ (resp. $a_n \lesssim b_n$) if $a_n \geq Cb_n$ (resp. $a_n \leq Cb_n$) for all n, and $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold simultaneously. We also write $a_n = O(b_n), a_n = \Theta(b_n), a_n = \Omega(b_n)$ as $a_n \lesssim b_n, a_n \asymp b_n, a_n \gtrsim b_n$, respectively.

4.2 Constructing the ROOT-SGD algorithm

In this section, we introduce the ROOT-SGD algorithm that is the focus of our study. We first motivate the algorithm from an averaging and variance reduction perspective. We then describe the burn-in and restarting mechanism, which contributes to the superior theoretical guarantees in the overall algorithm.

4.2.1 *Motivation and gradient estimator*

Our choice of step-size emerges from an overarching statistical perspective—rather than viewing the problem as one of correcting SGD via particular mechanisms such as averaging, variance reduction or momentum, we instead view the problem as one of utilizing all previous online data samples, $\xi_1, \ldots, \xi_t \sim P$, to form an *estimate* $\mathsf{Estimator}_t$ of $\nabla F(\theta_{t-1})$ at each round t. We then perform a gradient step based on this estimator—that is, we compute $\theta_t = \theta_{t-1} - \eta_t \cdot \mathsf{Estimator}_t$.

Concretely, our point of departure is the following *idealized* estimate of the error in the current gradient:

$$\mathsf{Estimator}_t - \nabla F(\theta_{t-1}) = \frac{1}{t}\sum_{s=1}^{t}(\nabla f(\theta_{s-1};\xi_s) - \nabla F(\theta_{s-1})). \tag{4.3}$$

Treating the terms $\nabla f(\theta_{s-1};\xi_s) - \nabla F(\theta_{s-1}), s = 1,\ldots,t$ as martingale differences, and assuming that the conditional variances of these terms are identical almost surely, it is straightforward to verify that the choice of equal weights $\frac{1}{t}$ minimizes the variance of the estimator over all such convex combinations. This simple but very specific choice of weights is central to our algorithm, which we refer to as *Recursive One-Over-T SGD* (ROOT-SGD).

The recursive aspect of the algorithm arises as follows. We set $\mathsf{Estimator}_1 = \nabla f(\theta_0;\xi_1)$ and express (4.3) as follows:

$$\mathsf{Estimator}_t - \nabla F(\theta_{t-1}) = \frac{1}{t}(\nabla f(\theta_{t-1};\xi_t) - \nabla F(\theta_{t-1})) + \frac{t-1}{t}(\mathsf{Estimator}_{t-1} - \nabla F(\theta_{t-2})).$$

Rearranging gives

$$\mathsf{Estimator}_t = \frac{1}{t}\nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})) + \frac{t-1}{t}\mathsf{Estimator}_{t-1}.$$

We now note that we do *not* generally have access to the bracketed term $\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})$, and replace the term by an unbiased estimator, $\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)$, based on the current sample ξ_t. Intuitively, the replacement should not affect much as long as the stochastic function admits some smoothness condition. Letting v_t denote $\mathsf{Estimator}_t$ with this replacement, we obtain the following recursive update:

$$\begin{aligned} v_t &= \frac{1}{t}\nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \frac{t-1}{t}v_{t-1} \\ &= \underbrace{\nabla f(\theta_{t-1};\xi_t)}_{\text{stochastic gradient}} + \underbrace{\frac{t-1}{t}(v_{t-1} - \nabla f(\theta_{t-2};\xi_t))}_{\text{correction term}}, \end{aligned} \tag{4.4}$$

consisting of both a stochastic gradient and a correction term.

Finally, performing a gradient step based on our estimator yields the ROOT-SGD algorithm:

$$v_t = \nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}(v_{t-1} - \nabla f(\theta_{t-2};\xi_t)) \tag{4.5a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t, \tag{4.5b}$$

where $\{\eta_t\}_{t\geq 1}$ is a suitably chosen sequence of positive step-sizes. Note that v_t defined in Eq. (4.4) is a recursive estimate of $\nabla F(\theta_{t-1})$ that is *unconditionally* unbiased in the sense that $\mathbb{E}[v_t] = \mathbb{E}[\nabla F(\theta_{t-1})]$. So the θ-update is an approximate gradient-descent step that moves along the negative direction $-v_t$.[2]

[2] Unlike many classical treatments of stochastic approximation, we structure the subscripts so they match up with those of the filtration corresponding to the stochastic processes.

We initialize $\theta_0 \in \mathbb{R}^d$, and, to avoid ambiguity, we define the update (4.5) at $t = 1$ to use only $v_1 = \nabla f(\theta_0; \xi_1)$. Overall, given the initialization $(\theta_0, v_0, \theta_{-1}) = (\theta_0, 0, \text{arbitrary})$, at each step $t \geq 1$ we take as input $\xi_t \sim P$, and perform an update of $(\theta_t, v_t, \theta_{t-1})$. This update depends only on $(\theta_{t-1}, v_{t-1}, \theta_{t-2})$ and ξ_t, and is first-order and Markovian.

4.2.2 Two-time-scale structure and burn-in period

For the purposes of both intuition and the proof itself, it is useful to observe that the iterates (4.5) evolve in a two-time-scale manner. Define the process $z_t :$ $= v_t - \nabla F(\theta_{t-1})$ for $t = 1, 2, \cdots$, which characterizes the *tracking error* of v_t as an estimator for the gradient. For each $\theta \in \mathbb{R}^d$ and $\xi \sim P$ we define the noise term $\varepsilon(\theta; \xi) = \nabla_\theta f(\theta; \xi) - \nabla F(\theta)$, and use the shorthand notation $\varepsilon_s(\cdot) = \varepsilon(\cdot; \xi_s)$. Some algebra yields the decomposition

$$t \cdot z_t = \sum_{s=1}^{t} \varepsilon_s(\theta_{s-1}) + \sum_{s=1}^{t} (s-1)\big(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\big), \quad \text{valid for } t = 1, 2, \ldots. \tag{4.6a}$$

In this way, we see that the process $(t \cdot z_t)_{t \geq 1}$ is a martingale difference sequence adapted to the natural filtration $(\mathscr{F}_t)_{t \geq 0}$. Indeed, the quantity z_t plays the role of averaging the noise as well as performing a weighted averaging of consecutive differences collected along the path. On the other hand, the process $(t \cdot v_t)_{t \geq 1}$ moves rapidly driven by the strong convexity of the function F:

$$tv_t = (t-1)\big\{v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\big\} + \nabla F(\theta_{t-1}) + \varepsilon_t(\theta_{t-1}) + (t-1)\big(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\big). \tag{4.6b}$$

Given an appropriate step-size η_t, the first term on the RHS of Eq. (4.6b) exhibits a contractive behavior. Consequently, the process $(tv_t)_{t \geq 1}$ plays the role of a *fast process*, driving the motion of iterates $(\theta_t)_{t \geq 0}$, and the noise-collecting process z_t is a *slow process*, collecting the noise along the path and contributing to the asymptotic efficiency of θ_t. Note that the fast process moves with a step-size η_t, making $\eta_t \mu$ progress when F is μ-strongly convex, while the slow process works with a step-size $\frac{1}{t}$. In order to make the iterates stable, we need the fast process to be fast in a relative sense, requiring that $\eta_t \mu \geq \frac{1}{t}$. This motivates a burn-in period in the algorithm, namely, in the first T_0 iterations, we run the recursion (4.5) with step-size zero and simply average the noise at θ_0; we then start the algorithm with an appropriate choice of step-size. Concretely, given some initial vector $\theta_0 \in \mathbb{R}^d$, we set $\theta_t = \theta_0$ for all $t = 1, \ldots, T_0 - 1$, and compute

$$v_t = \frac{1}{t} \sum_{s=1}^{t} \nabla f(\theta_0; \xi_s), \qquad \text{for all } t = 1, \ldots, T_0. \tag{4.7}$$

Algorithm 2 ROOT-SGD

```
Input: initialization $\theta_0$; step-size sequence $(\eta_t)_{t\geq 1}$
for $t = 1, 2, \ldots, T$ do
    $v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right)$
    $\theta_t = \theta_{t-1} - \eta_t v_t$
end for
Output: $\theta_T$
```

As suggested by our discussion, an algorithm with step-size $\eta_t = \eta$ will need a burn-in period of length $T_0 \asymp \frac{1}{\eta\mu}$ for a μ-strongly convex function F. Equivalently, we can view the step-sizes in the update for θ_t as being scheduled as follows:

$$\eta_t = \begin{cases} \eta, & \text{for } t \geq T_0, \\ 0, & \text{for } t = 1, \ldots, T_0 - 1, \end{cases} \tag{4.8}$$

briefed as $\eta_t = \eta \cdot 1[t \geq T_0]$, and, accordingly, the update rule from Eqs. (4.5a) and (4.5b) splits into two phases:

$$v_t = \begin{cases} \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right) & \text{for } t \geq T_0 + 1, \\ \frac{1}{t}\sum_{s=1}^{t} \nabla f(\theta_0; \xi_s) & \text{for } t = 1, \ldots, T_0, \end{cases}$$

$$\theta_t = \begin{cases} \theta_{t-1} - \eta v_t & \text{for } t \geq T_0, \\ \theta_0 & \text{for } t = 1, \ldots, T_0 - 1. \end{cases}$$

Such an algorithmic design has the length of the burn-in period for our algorithm is identical to the number of processed samples, so it features that the iteration number is identical to the sample complexity. The ROOT-SGD scheme is presented formally as Algorithm 2; for the remainder of this chapter, when referring to ROOT-SGD, we mean Algorithm 2 unless specified otherwise.

4.3 Main results

In this section, we present our main nonasymptotic and asymptotic results. We first establish a preliminary nonasymptotic result in §4.3.1. With augmented smoothness and moment assumptions, we then introduce in §4.3.2 sharp nonasymptotic upper bounds with unit pre-factor on the term characterizing the optimal statistical risk. Finally, in §4.3.3, we establish the asymptotic efficiency of ROOT-SGD.

4.3.1 Preliminary nonasymptotic results

We begin by presenting preliminary nonasymptotic results for ROOT-SGD. Before formally presenting the result, we detail our assumptions for the stochastic function $f(\cdot;\xi)$ and the expectation F.

First, we impose strong convexity and smoothness assumptions on the objective function:

Assumption 1 (Strong convexity and smoothness) *The population objective objective function F is twice continuously differentiable, μ-strongly-convex and L-smooth for some $0 < \mu \leq L < \infty$:*

$$\left\|\nabla F(\theta) - \nabla F(\theta')\right\|_2 \leq L\left\|\theta - \theta'\right\|_2, \qquad \textit{and} \quad \left\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \right\rangle \geq \mu \left\|\theta - \theta'\right\|_2^2,$$

for all pairs $\theta, \theta' \in \mathbb{R}^d$.

Second, we assume sufficient regularity for the covariance matrix at the global minimizer θ^*:

Assumption 2 (Finite variance at optimality) *At any minimizer θ^* of F, the stochastic gradient $\nabla f(\theta^*;\xi)$ has a positive definite covariance matrix, $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)(\nabla f(\theta^*;\xi))^\top\right]$, with its trace $\sigma_*^2 := \mathbb{E}\left\|\nabla f(\theta^*;\xi)\right\|_2^2$ assumed to be finite.*

Note that we only assume a finite variance on the stochastic gradient at the global minimizer θ^*. This is significantly weaker than the standard assumption of a globally bounded noise variance. See [145] and [112] for a detailed discussion of this assumption on the noise.

Third, we impose a mean-squared Lipschitz condition on the stochastic noise:

Assumption 3 (Lipschitz stochastic noise) *The noise function $\theta \mapsto \varepsilon(\theta;\xi)$ in the associated stochastic gradients satisfies the bound*

$$\mathbb{E}\left\|\varepsilon(\theta;\xi) - \varepsilon(\theta';\xi)\right\|_2^2 \leq \ell_\Xi^2 \left\|\theta - \theta'\right\|_2^2, \qquad \textit{for all pairs } \theta;\theta' \in \mathbb{R}^d. \tag{4.10}$$

We note that in making Assumption 3, we separate the stochastic smoothness (in the L^2 sense) of the noise, $\varepsilon(\theta;\xi) = \nabla f(\theta;\xi) - \nabla F(\theta)$, from the smoothness of the population-level objective. The magnitude of ℓ_Ξ and L are not comparable in general. This flexibility permits, for example, mini-batch algorithms where the population-level Lipschitz constant L remains fixed but the parameter ℓ_Ξ decreases with batch size. Such a separation has been adopted in nonconvex stochastic optimization literature [10].[3]

[3] Observe that Assumptions 1 and 3 imply a mean-squared Lipschitz condition on the stochastic gradient function:

$$\mathbb{E}\left\|\nabla f(\theta;\xi) - \nabla f(\theta';\xi)\right\|_2^2 = \left\|\nabla F(\theta) - \nabla F(\theta')\right\|_2^2 + \mathbb{E}\left\|\varepsilon(\theta;\xi) - \varepsilon(\theta';\xi)\right\|_2^2 \leq \left(L^2 + \ell_\Xi^2\right)\left\|\theta - \theta'\right\|_2^2,$$

where the final step uses the L-Lipschitz condition on the population function F.

Finally, we remark that all of these assumptions are standard in the stochastic optimization and statistical literature; and specific instantations of these assumptions are satisfied by a broad class of statistical models and estimators. We should note, however, that the strong convexity and smoothness (Assumption 1) is a global condition stronger than those typically used in the asymptotic analysis of M-estimators in the statistical literature. These conditions are needed for the fast convergence of the algorithm as an optimization algorithm, making it possible to establish nonasymptotic bounds. Assumptions 2 and 3 are standard for proving asymptotic normality of M-estimators and Z-estimators, see, e.g., [174, Theorem 5.21]. In contrast to some prior work, e.g., [76, 77], we *do not* assume uniform upper bounds on the variance of the stochastic gradient noise; this assumption fails to hold for various statistical models of interest, and theoretical results that dispense with it are of practical interest.

With the aforementioned assumptions in place, we provide our first preliminary nonasymptotic result for single-epoch ROOT-SGD, as follows:

Theorem 4.1 (Preliminary nonasymptotic results, single-epoch ROOT-SGD). *Under Assumptions 1, 2, 3, suppose that we run Algorithm 2 with burn-in period T_0 and step-size η such that*

$$T_0 := \frac{24}{\eta\mu} \qquad \textit{and} \quad \eta \in (0, \eta_{max}], \qquad \textit{where} \quad \eta_{max} := \frac{1}{4L} \wedge \frac{\mu}{8\ell_{\Xi}^2}. \tag{4.11}$$

Then, for any iteration $T \geq 1$, the iterate θ_T satisfies the bound

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \frac{28\,\sigma_*^2}{T} + \frac{2700\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 T^2}. \tag{4.12}$$

We provide a complete analysis of Theorem 4.1 in §4.4. In order to interpret the result, we make few remarks in order.

(i) Note when stating the upper bound (4.12) we adopt the convergence metrics in expected squared gradient norm. The guarantee in (4.12) consists of the sum of two terms which differs in magnitude as $T \to \infty$. The leading-order first term is contributed by the *optimal statistical risk*, and is determined by the noise variance σ_*^2 at the minimizer. The higher-order second term, on the other hand, exhibits an $O(\frac{1}{T^2})$-dependency on the initial condition, which is suboptimal and can be improved to an exponential dependency by properly restarting the algorithm. When the step-size η is fixed, a comparison of the two summand terms (4.12) yields that the optimal asymptotic risk $\frac{\sigma_*^2}{T}$ for the squared gradient holds up to an absolute constant whenever $T \gtrsim \frac{1}{\eta\mu} \vee \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2\sigma_*^2}$.

(ii) Suppose that we use the maximal step-size η_{max} permitted by condition (4.11). By converting the convergence rate bound (4.12) into a sample complexity bound, we then find that it suffices to take

$$C_{4.1}(\varepsilon) = \max\left\{ \frac{74}{\eta_{\max}\mu} \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{56\sigma_*^2}{\varepsilon^2} \right\} \asymp \max\left\{ \left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right) \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2} \right\} \tag{4.13}$$

samples in order to obtain an estimate of θ^* with gradient norm bounded as $O(\varepsilon)$. When the asymptotics holds as ε tends to zero with other problem-dependent constants being bounded away from zero, the leading-order term $\asymp \frac{\sigma_*^2}{\varepsilon^2}$ in $C_{4.1}(\varepsilon)$ matches the optimal statistical risk up to a constant pre-factor. To the best of our knowledge, such a bound on sample complexity is achieved for the first time by a stochastic first-order algorithm in the setting where only first-order smoothness condition holds, i.e. no continuity condition on the Hessians are posed. The only prior results which reported a near-optimal statistical risk under comparable settings in the leading-order stochastic optimization are due to [146] and [7], achieving the optimal risk up to a polylogarithmic factor. In Appendix §4.5, we compare our non-asymptotic results with these existing works in detail.

4.3.2 *Improved nonasymptotic upper bounds*

The convergence rate bound of Theorem 4.1 matches the optimal risk by a constant pre-factor c—to be precise, $c = 28$ in the provided analysis. In addition to this non-optimal pre-factor, this result does not match the efficiency of M-estimators in its higher-order dependency. So as to overcome these limitations, we now show how to apply Theorem 4.1 as the building block to seek to obtain a sharp fine-grained convergence rate via two-time-scale characterization, under additional smoothness and moment assumptions.

First, we need the following *Lipschitz continuity condition* for the Hessian at the optimum. We denote $H^* := \nabla^2 F(\theta^*)$ throughout.

Assumption 4 (Lipschitz continuous Hessians) *There exists a Lipschitz constant $L_\gamma > 0$ such that*

$$|||\nabla^2 F(\theta) - \nabla^2 F(\theta^*)|||_{op} \le L_\gamma \|\theta - \theta^*\|_2 . \tag{4.14}$$

We also need fourth-moment analogue of Assumptions 2 and 3, stated as follows. Note that these conditions are also exploited in prior work on nonasymptotic analyses of PRJ averaging procedure [15, 189, 74] and Streaming SVRG [73].

Assumption 5 (Finite fourth moment at minimizers) *Let Assumption 2 hold, and let $\widetilde{\sigma_*}^2 := \sqrt{\mathbb{E}\|\nabla f(\theta^*;\xi)\|_2^4}$ be finite.*

Observe that $\sigma_* \le \widetilde{\sigma_*}$ by Hölder's inequality. This distinction is important, as σ_*^2 corresponds to the optimal statistical risk (measured in gradient norm), while $\widetilde{\sigma_*}^2$ does not.

Algorithm 3 ROOT-SGD, multi-epoch version

```
**Input:** initialization $\theta_0$; fixed step-size $\eta$; burn-in time $T_0$; short epochs length $T^\flat \geq T_0$; short epochs number $B$
Set initialization for first epoch $\theta_0^{(1)} = \theta_0$
**for** $b = 1, 2, \cdots, B$ **do**
    Run ROOT-SGD (Algorithm 2) for $T^\flat$ iterates with burn-in time $T_0$ (i.e. step-size sequence $(\eta_t)_{t\geq 1}$ defined as in Eq. (4.8))
    Set the initialization $\theta_0^{(b+1)} := \theta_{T^\flat}^{(b)}$ for the next epoch
**end for**
Run ROOT-SGD (Algorithm 2) for $T := n - T^\flat B$ iterates with burn-in time $T_0$
**Output:** The final iterate estimator $\theta_n^{\mathrm{final}} := \theta_T^{(B+1)}$
```

For the higher-order moments of stochastic gradients, we introduce the following

Assumption 6 (Lipschitz stochastic noise in fourth moment) *The noise function $\theta \mapsto \varepsilon(\theta;\xi)$ in the associated stochastic gradients satisfies the bound*

$$\sqrt{\mathbb{E}\left\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\right\|_2^4} \leq \ell_\Xi^2 \left\|\theta_1 - \theta_2\right\|_2^2, \qquad \textit{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d. \quad (4.15)$$

Note that we slightly abuse the notation and denote ℓ_Ξ by both moment Lipschitz constants in Assumptions 3 and 6. In the presentations for the rest of this subsection, the notation ℓ_Ξ should be understood as the parameter in Assumption 6, which is strictly stronger than Assumption 3.

Formally, we present a multi-epoch version of the ROOT-SGD algorithm in Algorithm 3. The algorithm runs B short epochs and one long epoch. The goal of each short epoch is to "halve" the dependency on the initial condition $\|\nabla F(\theta_0)\|_2$, and it suffices to take $T^\flat = cT_0$ for some universal constant $c > 1$. We further impose the mild condition that the quantity $\|\nabla F(\theta_0)\|_2 / \sigma_*$ scales as a polynomial function of n.[4] In the following Theorem 4.2, we present the gradient norm bounds satisfied by the multi-epoch ROOT-SGD algorithm:

Theorem 4.2 (Improved nonasymptotic upper bound, multi-epoch ROOT-SGD). *Under Assumptions 1, 4, 5, 6, suppose that we run Algorithm 3 with the number of short epochs $B = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$, the burn-in time $T_0 = \frac{24}{\eta\mu}$, and the small epoch length $T^\flat = \frac{7340}{\eta\mu}$. Then for any step-size $\eta \in \left(0, \frac{1}{56L} \wedge \frac{\mu}{64\ell_\Xi^2}\right]$ and $n \geq T^\flat B + 1$, it returns an estimate $\theta_n^{\mathrm{final}}$ such that*

$$\mathbb{E}\left\|\nabla F(\theta_n^{\mathrm{final}})\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\log T}{\eta\mu T} + \frac{\ell_\Xi^2\log T}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \quad (4.16)$$

where $T := n - T^\flat B$, and C is a universal constant.

[4] This assumption is used only to simplify the presentation. If it does not hold true, the $\log n$ terms in the bounds will be replaced by $\log n + \log\left(1 + \|\nabla F(\theta_0)\|_2 / \sigma_*\right)$.

See §D.1.2 for the proof of this theorem.

In order to interpret this result, let us take $n \geq 2T^{\flat}B$ so that we have $\frac{1}{T} \leq \frac{1}{n} + \frac{2T^{\flat}B}{n^2}$. When the number of online samples n is given a priori and the (constant) step-size is optimized as $\eta = \frac{c}{\ell_{\Xi}\sqrt{n}} \wedge \frac{1}{4L}$ where $c = 0.49$, some algebra reduces the bound to

$$\mathbb{E}\left\|\nabla F(\theta_n^{\mathrm{final}})\right\|_2^2 - \frac{\sigma_*^2}{n} \lesssim \left\{\frac{\ell_{\Xi}}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}\frac{\sigma_*^2 \log n}{n} + \underbrace{\left\{\frac{\ell_{\Xi}}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}^{1/2} \frac{L_{\gamma}}{\mu^2}\left(\frac{\widetilde{\sigma_*}^2}{n}\right)^{3/2}}_{=:\widetilde{\mathscr{H}_n}}, \tag{4.17}$$

where $\widetilde{\mathscr{H}_n}$ is the linearization-induced term. Given the sufficiently large sample size n satisfying the requirement $n \gtrsim \frac{L}{\mu} + \frac{\ell_{\Xi}^2}{\mu^2}$, the pre-factors in the second term of (4.17), as well as the linearization error term $\widetilde{\mathscr{H}_n}$, start to diminish when the sample size n grows. The gradient norm bound (4.17) consists of three terms. We discuss each of them as follows:

(i) The leading-order term $\frac{\sigma_*^2}{n}$ is exactly the asymptotic risk of the optimal limiting Gaussian random vector in the local asymptotic minimax theorem, measured with squared gradient norm. Note that this term depends only on the noise at the optimum θ^*, instead of some uniform upper bounds.

(ii) The first term on the right-hand side consists of two parts that decay at different rates. If the sample size satisfies $\frac{n}{\log n} \gtrsim \frac{L}{\mu} + \frac{\ell_{\Xi}^2}{\mu^2}$ (which is a slightly stronger condition than the sample size needed for the theorem to hold true), this term is always smaller than the $\frac{\sigma_*^2}{n}$ term in the left hand side. For a large sample size n, the dominating high-order term decays at the rate $O(n^{-3/2}\log n)$.

(iii) The remaining high-order term $\widetilde{\mathscr{H}_n}$ in the bound (4.17) scales as $O(n^{-7/4})$, a faster rate of decay than the previous term. This term is induced by a linearization argument in our proof, and therefore depends on the Lipschitz constant L_1 of the Hessian matrix.

Additionally, we remark that the sample size requirement $n \gtrsim \frac{L}{\mu} + \frac{\ell_{\Xi}^2}{\mu^2}$ in Theorem 4.2 is natural: on the one hand, under the noise assumption 3, it requires $\Theta\left(\frac{\ell_{\Xi}^2}{\mu^2}\right)$ samples to distinguish the function $x \mapsto \frac{\mu}{2}\|x\|_2^2$ from a constant function 0; on the other hand, the $\frac{L}{\mu}$ term is consistent with the optimization essence of the problem — as the ROOT-SGD algorithm reduces to gradient descent in the noiseless case, we need to pay for the complexity of gradient descent to achieve any meaningful guarantees.

Besides the expected gradient norm squared metric, we also establish guarantees in alternative metrics including the estimation error $\|\theta_n - \theta^*\|_2$ and the objective gap $F(\theta_n) - F(\theta^*)$ with expectation taken. In order to state the theorem, we define the following linearization-induced error terms that appears in the bound

$$\widetilde{\mathscr{H}}_n^{(\mathrm{MSE})} := \frac{c}{\lambda_{\min}(H^*)^2} \cdot \left\{ \frac{L_1}{\mu^2} \cdot \left(\frac{\widetilde{\sigma}_*^2}{n} \right)^{3/2} + \frac{L_1^2}{\mu^4} \cdot \left(\frac{\widetilde{\sigma}_*^2}{n} \right)^2 \right\}, \quad \text{and} \tag{4.18a}$$

$$\widetilde{\mathscr{H}}_n^{(\mathrm{OBJ})} := \frac{c}{\mu} \cdot \frac{L_1}{\mu^2} \cdot \left(\frac{\widetilde{\sigma}_*^2}{n} \right)^{3/2} + \frac{c}{\lambda_{\min}(H^*)} \cdot \frac{L_1^2}{\mu^4} \cdot \left(\frac{\widetilde{\sigma}_*^2}{n} \right)^2. \tag{4.18b}$$

For simplicity we only consider the multi-epoch ROOT-SGD as specified in Theorem 4.2, where we conclude the following Corollary:

Corollary 4.3 (Nonasymptotic bounds in alternative metrics, multi-epoch ROOT-SGD) *Under the setup of Theorem 4.2, the multi-epoch ROOT-SGD algorithm with the optimal step-size choice of $\eta \asymp \frac{1}{L} \wedge \frac{1}{\ell_\Xi \sqrt{n}}$ produces an estimator that satisfies the following bound for $n \geq T^\flat B + 1$:*

$$\mathbb{E} \left\| \theta_n^{\mathrm{final}} - \theta^* \right\|_2^2 - \frac{1}{n} \mathrm{Tr}\left((H^*)^{-1} \Sigma^* (H^*)^{-1} \right) \leq c \left\{ \frac{\ell_\Xi}{\mu \sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{\lambda_{\min}(H^*)^2 n} + \widetilde{\mathscr{H}}_n^{(\mathrm{MSE})}, \tag{4.19a}$$

$$\mathbb{E} \left[F\left(\theta_n^{\mathrm{final}} \right) - F(\theta^*) \right] - \frac{1}{2n} \mathrm{Tr}\left((H^*)^{-1} \Sigma^* \right) \leq c \left\{ \frac{\ell_\Xi}{\mu \sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{\lambda_{\min}(H^*) n} + \widetilde{\mathscr{H}}_n^{(\mathrm{OBJ})}. \tag{4.19b}$$

See §D.1.3 for the proof of this result, where the key technical addition lies on the adoption of a generic matrix-induced bound on the stochastic processes. Obviously, the optimal step-size choice in alternative metrics is same in magnitude as the one in squared gradient norm metric in Theorem 4.2. We can compare the bounds in Corollary 4.3 with the gradient norm bound (4.17) induced by Theorem 4.2, discussed term-by-term as follows:

(i) The leading-order terms in the bound (4.19a) and (4.19b), specified in the subtracted second terms on the left hands, are both optimal in a local asymptotic minimax sense with unity pre-factor. In particular, they are exactly the asymptotic risk of the limiting Gaussian random variable $\mathscr{N}\left(\theta^*, \frac{1}{n} (H^*)^{-1} \Sigma^* (H^*)^{-1} \right)$ in corresponding metrics. We also note that in the special case of well-specified maximal-likelihood estimation, Fisher's identity $H^* = \Sigma^*$ holds true, and the leading-order terms in Eq. (4.19a) and (4.19b) become $\frac{1}{n} \mathrm{Tr}\left((H^*)^{-1} \right)$ and $\frac{d}{2n}$, respectively. This is in accordance with the classical asymptotic theory for M-estimators (c.f. [174, §5.3]);

(ii) The dominating term $\left\{ \frac{\ell_\Xi}{\mu \sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{n}$ in the gradient norm bound is correspondingly multiplied by a factor of $\lambda_{\min}(H^*)^{-2}$ (resp. $\lambda_{\min}(H^*)^{-1}$) in the MSE (resp. objective gap) bound on the right hand (4.19a) (resp. (4.19b)), which is intuitively consistent with conversions in metric;

(iii) While the dominating high-order term on the right hand matches the corresponding optimal statistical risk and the higher-order terms altogether scale as $O(n^{-3/2})$, the linearization-induced error terms $\widetilde{\mathscr{H}}_n^{\mathrm{MSE}}$ and $\widetilde{\mathscr{H}}_n^{\mathrm{OBJ}}$ both decay at a rate of $O\left(n^{-3/2}\right)$ as long as L_1 is bounded away from zero i.e. the objective is essentially nonquadratic. This is vastly different from the bound in gradient norm (4.16) of Theorem 4.2 where the linearization-induced terms are all incorporated in $O(n^{-7/4})$, primarily due to that the pre-factor $\left\{\frac{\ell_\Xi}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}^{1/2}$ in the linearization-induced term $\widetilde{\mathscr{H}_n}$ in (4.17) is replaced by unity.[5] In consistency with the metric conversion, the linearization-induced terms $\widetilde{\mathscr{H}_n}^{\mathrm{MSE}}$ and $\widetilde{\mathscr{H}_n}^{\mathrm{OBJ}}$ also incur additional factors related to the smallest eigenvalue of H^* on top of the linearization-induced error term $\widetilde{\mathscr{H}_n}$ in (4.17).

4.3.3 Asymptotic results

In this subsection, we study the asymptotic behavior of our ROOT-SGD algorithm. We aim to prove the asymptotic efficiency of the multi-epoch estimator of Algorithm 3 under minimal assumptions. In this case, Assumptions 1, 2 and 3 are the standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see e.g. [174, Theorem 5.21]). We first introduce our one-point Hessian continuity condition as follows as the qualitative counterpart of the continuity Assumption 4:

Assumption 7 (One-point Hessian continuity) *The Hessian mapping $\nabla^2 F(\theta)$ is continuous at the minimizer θ^*, i.e.,*

$$\lim_{\theta\to\theta^*} \left|\!\left|\!\left|\nabla^2 F(\theta) - H^*\right|\!\right|\!\right|_{op} = 0.$$

Note in Assumption 7 we assume only the continuity of Hessian matrix at θ^* without posing any bounds on its modulus of continuity. This is a much weaker condition than a Lipschitz or Hölder condition posted on the Hessian matrix as required in the analysis of the Polyak-Ruppert averaging procedure [150].

With this setup, we are ready to state our weak convergence asymptotic efficiency result for θ^{final} in the following theorem,[6] whose proof is provided in §D.1.4:

Theorem 4.4 (Asymptotic efficiency, multi-epoch ROOT-SGD). *Under Assumptions 1, 2, 3 and 7, suppose that we run the multi-epoch Algorithm 3 with burn-in time $T_0 = \frac{24}{\eta\mu}$, short-epoch length $T^\flat = \frac{7340}{\eta\mu}$ and number of short epochs $B = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$. Then as $n\to\infty$, $\eta\to 0$ such that $\eta(n - T^\flat B)\to\infty$ and $T^\flat B/n \to 0$, the estimate $\theta_n^{\mathrm{final},(\eta)}$ satisfies the weak convergence*

[5] This is because the Hessian-Lipschitz assumption plays a key role in relating MSE and objective gap to the underlying noise structure in the stochastic optimization problem, paying for larger linearization error; whereas in the gradient norm bound, the Hessian-Lipschitz assumption is employed only to mitigate the effect correlation that appears at even higher-order terms in the bound.

[6] We emphasize our estimator's dependency on the step-size η by explicitly bracketing it in the superscript.

$$\sqrt{n}\left(\theta_n^{\mathrm{final},(\eta)} - \theta^*\right) \xrightarrow{d} \mathscr{N}\left(0, [\nabla^2 F(\theta^*)]^{-1}\Sigma^*[\nabla^2 F(\theta^*)]^{-1}\right), \tag{4.20}$$

where $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top\right]$ is the covariance of the stochastic gradient at the minimizer.

We remark that Theorem 4.4 holds under the mere additional assumption of one-point continuity on the Hessian matrix, which is usually the minimal assumption needed for an asymptotic efficiency result to hold. Here we are adopting the multi-epoch ROOT-SGD with the same algorithmic specifications as in Theorem 4.2, and we achieve the asymptotic convergence to the Gaussian limit that matches the Cramér-Rao lower bound. The asymptotic covariance matrix in Eq. (4.20), however, carries significantly more information than the (scalar) optimal asymptotic risk. Our asymptotic result is in a triangular-array format: we let the fixed constant step-size scale down with n where the scaling condition is essentially $\eta \to 0$, $n \to \infty$ with $\frac{\eta n}{\log(\eta^{-1})} \to \infty$, which is satisfied when $\eta \asymp \frac{1}{n^{c_1}}$ for any fixed $c_1 \in (0,1)$. Although not directly comparable, the range of step-size asymptotics is broader than [150] and accordingly hints at potential advantages over PRJ, primarily due to our de-biasing corrections in our algorithm design and is consistent with our improved higher-order term in nonasymptotic result (Theorem 4.2 and Corollary 4.3). In Appendix §D.2, we establish an additional asymptotic normality result for ROOT-SGD with fixed *constant* step-size, which exhibits exactly the same limiting behavior as constant-step-size *linear* stochastic approximation with PRJ averaging procedure under comparable asssumptions [134].

We end this subsection by remarking that Theorem 4.4 only requires strong convexity, smoothness, and a set of noise moment assumptions standard in asymptotic statistics, but not any higher-order smoothness other than the continuity of Hessian matrices at θ^*. This matches the assumptions for asymptotic efficiency results in classical statistics literature [174, 175].

4.4 Proof of Theorem 4.1 and extended analysis

This section is devoted to an (extended) analysis and proof of Theorem 4.1. In our analysis we utilize the central object the *tracking error process* z_t defined as in (4.28), and we heavily use the fact that the process $(tz_t)_{t\geq T_0}$ is a martingale adapted to the natural filtration. Also in part of our analysis, as an alternative to our Lipschitz stochastic noise Assumption 3, we can impose the following *individual convexity and smoothness* condition [107, 91, 48, 144]:

Assumption 8 (Individual convexity/smoothness) *Almost surely, the (random) function $\theta \mapsto f(\theta;\xi)$ is convex, twice continuously differentiable and satisfies the Lipschitz condition*

$$\left\|\nabla f(\theta;\xi) - \nabla f(\theta';\xi)\right\|_2 \leq L_{max}\left\|\theta - \theta'\right\|_2 \ \textit{a.s.}, \quad \textit{for all pairs } \theta;\theta' \in \mathbb{R}^d. \tag{4.21}$$

All Assumptions 1 and 2 along with either Assumption 3 or 8, are standard in the stochastic optimization literature (cf. [145, 12, 112]). Note that Assumption 8 implies Assumption 3 with constant L_{max}; in many statistical applications, the quantity L_{max} can be significantly larger than $\sqrt{L^2 + \ell_\Xi^2}$ in magnitude.

With these assumptions in place, let us formalize the two cases in which we analyze the ROOT-SGD algorithm. We refer to these cases as the *Lipschitz Stochastic Noise* case (or **LSN** for short), and the *Individually Smooth and Convex* case (or **ISC** for short).

LSN Case: Suppose that Assumptions 1, 2 and 3 hold, and define

$$\eta_{max} := \frac{1}{4L} \wedge \frac{\mu}{8\ell_\Xi^2}. \tag{4.22}$$

ISC Case: Suppose that Assumptions 1, 2 and 8 hold, and define

$$\eta_{max} := \frac{1}{4L_{max}}. \tag{4.23}$$

As the readers shall see immediately, ω_{max} is a key quantity that plays a pivotal role in our analysis for both cases.

Theorem 4.5 (Unified nonasymptotic results, single-epoch ROOT-SGD). *Suppose that the conditions in either the **LSN** or **ISC** Case are in force, and let the step sizes be chosen according to the protocol* (4.8) *for some* $\eta \in (0, \eta_{max}]$*, and assume that we use the following burn-in time:*

$$T_0 := \left\lceil \frac{24}{\eta\mu} \right\rceil. \tag{4.24}$$

Then, for any iteration $T \geq 1$*, the iterate* θ_T *from Algorithm 2 satisfies the bound*

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \frac{2700\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2(T+1)^2} + \frac{28\,\sigma_*^2}{T+1}. \tag{4.25}$$

We provide the proof of Theorem 4.5 in both the **LSN** and **ISC** cases; the **LSN** case corresponds to Theorem 4.1. In accordance with the discussion in §4.1, our nonasymptotic convergence rate upper bound (4.25) for the expected squared gradient norm consists of the addition of two terms. The first term, $\frac{\sigma_*^2}{T}$, corresponds to the *nonimprovable statistical error* depending on the noise variance at the minimizer. The second term, which is equivalent to $\frac{\|\nabla F(\theta_0)\|_2^2 T_0{}^2}{T^2}$, corresponds to the *bias* or *optimization error* that indicates the polynomial forgetting from the initialization. Theorem 4.5 copes with a wide range of step sizes η: fixing the number of online samples T, (4.25) asserts that the optimal asymptotic risk $\frac{\sigma_*^2}{T}$ for the squared gradient holds up to an absolute constant whenever $T \gtrsim \frac{1}{\eta\mu} \vee \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2\sigma_*^2}$.

Converting the convergence rate bound in (4.25), we can achieve a tight upper bound on the sample complexity to achieve a statistical estimator of θ^* with gradient norm bounded by $O(\varepsilon)$:[7]

$$C_{4.1}(\varepsilon) = \max\left\{ \frac{74}{\eta_{max}\mu} \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{56\sigma_*^2}{\varepsilon^2} \right\}$$
$$\asymp \begin{cases} \max\left\{ \left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right) \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2} \right\}, & \text{for the } \mathbf{LSN} \text{ case,} \\ \max\left\{ \frac{L_{max}}{\mu} \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2} \right\}, & \text{for the } \mathbf{ISC} \text{ case.} \end{cases} \tag{4.26}$$

In above, the step size $\eta = \eta_{max}$ is optimized as in (4.22) for the **LSN** and (4.23) for the **ISC** case, separately, and where the asymptotics holds as ε tends to zero while σ_* is bounded away from zero. In both cases, the leading-order term of $C_{4.1}(\varepsilon)$ in either case is $\asymp \frac{\sigma_*^2}{\varepsilon^2}$ which matches the optimal statistical error up to universal constants, first among comparable literature in both cases.

Detailed proof.

The rest of this subsection devotes to prove Theorem 4.5. It is straightforward to show first (4.25) automatically holds for $T < T_0$ since for these T, $\theta_T = \theta_0$ and hence $\mathbb{E}\|\nabla F(\theta_T)\|_2^2 = \mathbb{E}\|\nabla F(\theta_0)\|_2^2$, so we only need to prove the result for $T \geq T_0$.

We first define ω_{max} which is a key quantity in our analysis in this section for both cases, as follows

$$\omega_{max} := \begin{cases} \frac{2\ell_\Xi^2}{\mu^2}, & \text{for } \mathbf{LSN} \text{ case,} \\ \frac{2L_{max}}{\mu}, & \text{for } \mathbf{ISC} \text{ case.} \end{cases} \tag{4.27}$$

A central object in our analysis is the iteration of *tracking error*, defined as

$$z_t := v_t - \nabla F(\theta_{t-1}), \qquad \text{for } t \geq T_0. \tag{4.28}$$

At a high level, this proof involves analyzing the evolution of the quantities v_t and z_t, and then bounding the norm of the gradient $\nabla F(\theta_{t-1})$ using their combination. From the updates (4.5), we can identify a martingale difference structure for the quantity tz_t: its difference decomposes as the sum of *pointwise stochastic noise*, $\varepsilon_t(\theta_{t-1})$, and the *incurred displacement noise*, $(t-1)\left[\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right]$. The expression of the martingale structure is expressed as

[7] Indeed, we choose T in Eq. (4.12) to be sufficiently large such that it satisfies the inequalities $T \geq T_0 = \lceil \frac{24}{\eta\mu} \rceil$, as well as $\frac{2700\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} \leq \frac{\varepsilon^2}{2}$ and $\frac{28\sigma_*^2}{T} \leq \frac{\varepsilon^2}{2}$. Here and on, we assume without loss of generality that $\varepsilon^2 \leq \|\nabla F(\theta_0)\|_2^2$. It is then straightforward to see that (4.13) serves as a tight sample complexity upper bound.

$$
\begin{aligned}
tz_t = t\left(v_t - \nabla F(\theta_{t-1})\right) &= \varepsilon_t(\theta_{t-1}) + (t-1)(v_{t-1} - \nabla F(\theta_{t-2})) + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) \\
&= \varepsilon_t(\theta_{t-1}) + (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})).
\end{aligned} \tag{4.29}
$$

Unwinding this relation recursively yields the decomposition

$$
tz_t - T_0 z_{T_0} = \sum_{s=T_0+1}^{t} \varepsilon_s(\theta_{s-1}) + \sum_{s=T_0+1}^{t} (s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})). \tag{4.30}
$$

We now turn to the proofs of the three auxiliary lemmas that allow us to control the relevant quantities and the main theorem, as follows:

Lemma 4.6 (Recursion involving z_t). *Under the conditions of Theorem 4.5, for all $t \geq T_0 + 1$, we have*

$$
t^2\mathbb{E}\|z_t\|_2^2 \leq (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + 2(t-1)^2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2. \tag{4.31a}
$$

On the other hand, for $t = T_0$, we have

$$
{T_0}^2\mathbb{E}\|v_{T_0}\|_2^2 - {T_0}^2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 = {T_0}^2\mathbb{E}\|z_{T_0}\|_2^2 = T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2. \tag{4.31b}
$$

See §4.4.1 for the proof of this claim. Note we have $z_{T_0} = v_{T_0} - \nabla F(\theta_0)$ which is simply the arithmetic average of T_0 i.i.d. noise terms at θ_0, $\varepsilon_1(\theta_0), \ldots, \varepsilon_{T_0}(\theta_0)$.

Our next auxiliary lemma characterizes the evolution of the sequence $(v_t : t \geq T_0)$ in terms of the quantity $\mathbb{E}\|v_t\|_2^2$.

Lemma 4.7 (Evolution of v_t). *Under the settings of Theorem 4.5, for any $\eta \in (0, \eta_{max}]$, we have*

$$
t^2\mathbb{E}\|v_t\|_2^2 - 2t\mathbb{E}\langle v_t, \nabla F(\theta_{t-1})\rangle + \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 = \mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^2, \tag{4.32a}
$$

and

$$
\begin{aligned}
\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^2 \leq (1-\eta\mu)\cdot(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 \\
- 2(t-1)^2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right\|_2^2,
\end{aligned} \tag{4.32b}
$$

for all $t \geq T_0 + 1$.

See §4.4.2 for the proof of this claim.

Our third auxiliary lemma bounds the second moment of the stochastic noise.

Lemma 4.8 (Second moment of pointwise stochastic noise). *Under the conditions of Theorem 4.5, we have*

$$
\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 \leq \omega_{max}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2, \qquad \textit{for all } t \geq T_0 + 1. \tag{4.33}
$$

See §4.4.3 for the proof of this claim.

Equipped with these three auxiliary results, we are now ready to prove Theorem 4.5.

Proof (Proof of Theorem 4.5). Our proof proceeds in two steps.
Step 1. We begin by applying the Cauchy-Schwarz and Young inequalities to the inner product $\langle v_t, \nabla F(\theta_{t-1})\rangle$. Doing so yields the upper bound

$$2t\langle v_t, \nabla F(\theta_{t-1})\rangle \le 2\left[t\|v_t\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2\right] \le \eta\mu t^2\|v_t\|_2^2 + \frac{1}{\eta\mu}\|\nabla F(\theta_{t-1})\|_2^2.$$

Taking the expectation of both sides and applying the bound (4.32a) from Lemma 4.7 yields

$$\begin{aligned}(1-\eta\mu)t^2\mathbb{E}\|v_t\|_2^2 - \frac{1-\eta\mu}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 &\le t^2\mathbb{E}\|v_t\|_2^2 - 2t\mathbb{E}\langle v_t, \nabla F(\theta_{t-1})\rangle + \mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2\\ &\le (1-\eta\mu)\cdot(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2\\ &\quad - 2(t-1)^2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2.\end{aligned}$$

Moreover, since we have $\eta \le \eta_{max} \le \frac{1}{4\mu}$ under condition (4.11), we can multiply both sides by $(1-\eta\mu)^{-1}$, which lies in $[1, \frac{3}{2}]$. Doing so yields the bound

$$t^2\mathbb{E}\|v_t\|_2^2 - \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \le (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 3\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 - 2(t-1)^2\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2.$$

Combining with the bound (4.31a) from Lemma 4.6 gives

$$\begin{aligned}&t^2\mathbb{E}\|z_t\|_2^2 + t^2\mathbb{E}\|v_t\|_2^2 - (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 - (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2\\ &\qquad\le 5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2.\end{aligned}$$

By telescoping this inequality from T_0+1 to T, we find that

$$\begin{aligned}&T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2 - T_0{}^2\mathbb{E}\|z_{T_0}\|_2^2 - T_0{}^2\mathbb{E}\|v_{T_0}\|_2^2\\ &\qquad\le \sum_{t=T_0+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right].\end{aligned} \tag{4.34}$$

Next, applying the result (4.31b) from Lemma 4.6 yields

$$\begin{aligned}&\frac{T^2}{2}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2\\ &\le T_0{}^2\mathbb{E}\|z_{T_0}\|_2^2 + T_0{}^2\mathbb{E}\|v_{T_0}\|_2^2 + \sum_{t=T_0+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right]\end{aligned}$$

$$= {T_0}^2\|\nabla F(\theta_0)\|_2^2 + 2T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2 + 5\sum_{t=T_0+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2.$$

Following some algebra, we find that

$$\begin{aligned}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le{}& \frac{2{T_0}^2\|\nabla F(\theta_0)\|_2^2 + 4T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2}{T^2}\\ &+ \frac{10}{T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{2}{\eta\mu T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2. \quad (4.35)\end{aligned}$$

Combining inequality (4.35) with the bound (4.33) from Lemma 4.8 gives

$$\begin{aligned}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le{}& \frac{2{T_0}^2\|\nabla F(\theta_0)\|_2^2 + 4T_0\left[\omega_{max}\mathbb{E}\|\nabla F(\theta_0)\|_2^2 + 2\sigma_*^2\right]}{T^2}\\ &+ \frac{10}{T^2}\sum_{t=T_0+1}^{T}\left[\omega_{max}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2\right] + \frac{2}{\eta\mu T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\\ \le{}& \frac{(4\omega_{max}+2T_0)T_0\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{10\omega_{max}+2\mu^{-1}\eta^{-1}}{T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{20\sigma_*^2}{T},\end{aligned}$$

concluding the following key gradient bound that controls the evolution of the gradient norm $\|\nabla F(\theta_{T-1})\|_2$:

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le \frac{1}{T^2}\left\{\alpha_1\mathbb{E}\|\nabla F(\theta_0)\|_2^2 + \alpha_2\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right\} + \frac{20\sigma_*^2}{T}, \quad (4.36)$$

where $\alpha_1 := (4\omega_{max}+2T_0)\,T_0$ and $\alpha_2 := 10\omega_{max} + \frac{2}{\eta\mu}$.
Step 2. Based on the estimation bound (4.36), the proof of Theorem 4.5 relies on a bootstrapping argument in order to remove the dependence of the right-hand side of Eq. (4.36) on the quantity $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$. Let $T^* \ge T_0+1$ be arbitrary. Telescoping the bound (4.36) over the iterates $T = T_0+1,\ldots,T^*$ yields

$$\sum_{T=T_0+1}^{T^*}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le \alpha_1\underbrace{\sum_{T=T_0+1}^{T^*}\frac{\|\nabla F(\theta_0)\|_2^2}{T^2}}_{Q_1} + \underbrace{\sum_{T=T_0+1}^{T^*}\frac{\alpha_2}{T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2}_{Q_2} + \underbrace{\sum_{T=T_0+1}^{T^*}\frac{20\sigma_*^2}{T}}_{Q_3}.$$

Let us deal with each of these quantities in turn, making use of the integral inequalities

$$\sum_{T=T_0+1}^{T^*}\frac{1}{T^2} \overset{(i)}{\le} \int_{T_0}^{T^*}\frac{d\tau}{\tau^2} \le \frac{1}{T_0}, \qquad \text{and} \qquad \sum_{T=T_0+1}^{T^*}\frac{1}{T} \overset{(ii)}{\le} \int_{T_0}^{T^*}\frac{d\tau}{\tau} = \log\left(\frac{T^*}{T_0}\right). \quad (4.37)$$

We clearly have

$$Q_1 \leq \frac{\alpha_1}{T_0} \|\nabla F(\theta_0)\|_2^2 \;=\; (4\omega_{max} + 2T_0)\; \|\nabla F(\theta_0)\|_2^2 .$$

Moreover, by using the fact that $T^* \geq T$, interchanging the order of summation, and then using inequality (4.37)(i) again, we have

$$\begin{aligned} Q_2 \leq \sum_{T=T_0+1}^{T^*} \frac{\alpha_2}{T^2} \sum_{t=T_0+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 &= \sum_{t=T_0+1}^{T^*} \Big(\sum_{T=T_0+1}^{T^*} \frac{\alpha_2}{T^2} \Big) \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \\ &\leq \frac{\alpha_2}{T_0} \sum_{t=T_0+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2. \end{aligned}$$

Finally, turning to the third quantity, we have $Q_3 \leq 20\sigma_*^2 \log\big(\frac{T^*}{T_0}\big)$, where we have used inequality (4.37)(ii). Putting together the pieces yields the upper bound

$$\sum_{T=T_0+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq (4\omega_{max} + 2T_0)\|\nabla F(\theta_0)\|_2^2 + \frac{\alpha_2}{T_0} \sum_{t=T_0+1}^{T^*} \mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 20\sigma_*^2 \log\left(\frac{T^*}{T_0}\right).$$

Eqs. (4.11) imply that, for either case under consideration, we have the bound $\omega_{max} \leq \frac{1}{\eta\mu}$, and, since $0 < \eta\mu \leq \frac{1}{4} < 1$, we have from (4.11) that $T_0 = \left\lceil \frac{24}{\eta\mu} \right\rceil \leq \frac{1}{\eta\mu}$, resulting in

$$4\omega_{max} + 2T_0 \leq \frac{4}{\eta\mu} + 2\left(\frac{1}{\eta\mu}\right) = \frac{54}{\eta\mu},$$

where we have the choice of burn-in time T_0 from Eq. (4.11). Similarly, we have $\alpha_2 = 10\omega_{max} + \frac{2}{\eta\mu} \leq \frac{12}{\eta\mu} \leq \frac{T_0}{2}$. Putting together the pieces yields

$$\frac{1}{2} \sum_{t=T_0+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \leq \frac{54}{\eta\mu} \mathbb{E}\|\nabla F(\theta_0)\|_2^2 + 20\sigma_*^2 \log\left(\frac{T^*}{T_0}\right). \tag{4.38}$$

Now substituting the inequality (4.38) back into the earlier bound (4.36) with $T^* = T$ allows us to obtain a bound on $\mathbb{E}\|\nabla F(\theta_{T-1})\|_2$. In particular, for any $T \geq T_0 + 1$, we have

$$\begin{aligned} \mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 &\leq \frac{54T_0}{\eta\mu} \cdot \frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{T_0}{T^2} \cdot \frac{1}{2} \sum_{t=T_0+1}^{T} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{20\sigma_*^2}{T} \\ &\leq \frac{54T_0}{\eta\mu} \cdot \frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{T_0}{T^2}\left[\frac{54}{\eta\mu}\|\nabla F(\theta_0)\|_2^2 + 20\sigma_*^2 \log\left(\frac{T}{T_0}\right)\right] + \frac{20\sigma_*^2}{T} \\ &\leq \frac{2(54)\,T_0}{\eta\mu} \cdot \frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{20\sigma_*^2}{T}\left[1 + \frac{T_0}{T}\log\left(\frac{T}{T_0}\right)\right]. \end{aligned}$$

Using the inequality $\frac{\log(x)}{x} \leq \frac{1}{e}$, valid for $x \geq 1$, we conclude that

$$\begin{aligned}
\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 &\leq \frac{2(54)T_0}{\eta\mu} \cdot \frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{20\sigma_*^2}{T}\left[1 + \frac{T_0}{T}\log\left(\frac{T}{T_0}\right)\right] \\
&\leq \frac{108}{\eta\mu} \cdot \frac{1}{\eta\mu} \cdot \frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{20\sigma_*^2}{T}\left[1 + \frac{1}{e}\right] \\
&\leq \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 T^2} + \frac{28\sigma_*^2}{T}.
\end{aligned}$$

Shifting the subscript forward by one yields Theorem 4.5.

4.4.1 *Proof of Lemma 4.6*

The claim (4.31b) follows from the definition along with some basic probability. In order to prove the claim (4.31a), recall from the ROOT-SGD update rule for v_t in the first line of (4.5) that for $t \geq T_0 + 1$ we have:

$$t v_t = (t-1)v_{t-1} + t\nabla f(\theta_{t-1};\xi_t) - (t-1)\nabla f(\theta_{t-2};\xi_t). \tag{4.39}$$

Subtracting the quantity $t\nabla F(\theta_{t-1})$ from both sides yields

$$t z_t = (t-1)v_{t-1} + t\nabla f(\theta_{t-1};\xi_t) - (t-1)\nabla f(\theta_{t-2};\xi_t) - t\nabla F(\theta_{t-1}).$$

Thus, we arrive at the following recursion for the estimation error z_t:

$$\begin{aligned}
t z_t &= (t-1)\left[v_{t-1} - \nabla F(\theta_{t-2})\right] \\
&\quad + t\left[\nabla f(\theta_{t-1};\xi_t) - \nabla F(\theta_{t-1})\right] - (t-1)\left[\nabla f(\theta_{t-2};\xi_t) - \nabla F(\theta_{t-2})\right] \\
&= (t-1)z_{t-1} + \varepsilon_t(\theta_{t-1}) + (t-1)\left[\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right].
\end{aligned}$$

Observing that the variable $\varepsilon_t(\theta_{t-1}) + (t-1)\left[\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right]$, defines an L^2-martingale-difference sequence, we see that

$$\begin{aligned}
t^2\mathbb{E}\|z_t\|_2^2 &= \mathbb{E}\,\|(t-1)z_{t-1}\|_2^2 + \mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) + (t-1)\left[\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right]\|_2^2 \\
&\leq (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + 2(t-1)^2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2,
\end{aligned}$$

where in the last step follows from Young's inequality. Computing the constants out completes the proof of the claim (4.31a).

4.4.2 Proof of Lemma 4.7

Eq. (4.32a) follows in a straightforward manner by expanding the square and taking an expectation. As for the inequality (4.32b), from the update rule (4.5) for v_t, we have

$$\begin{aligned}tv_t - \nabla F(\theta_{t-1}) &= t\nabla f(\theta_{t-1};\xi_t) + (t-1)\left[v_{t-1} - \nabla f(\theta_{t-2};\xi_t)\right] - \nabla F(\theta_{t-1})\\ &= (t-1)v_{t-1} + (t-1)\left[\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right] + \varepsilon_t(\theta_{t-1}).\end{aligned}$$

Using this relation, we can compute the expected squared Euclidean norm as

$$\begin{aligned}&\mathbb{E}\|tv_t - \nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\left\|(t-1)v_{t-1} + (t-1)\left[\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right] + \varepsilon_t(\theta_{t-1})\right\|_2^2\\ &\quad= \mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^2 + \mathbb{E}\left\|(t-1)\left[\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right] + \varepsilon_t(\theta_{t-1})\right\|_2^2\\ &\qquad\qquad + 2\mathbb{E}\left\langle (t-1)v_{t-1}, (t-1)\left[\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right] + \varepsilon_t(\theta_{t-1})\right\rangle.\end{aligned}$$

Further rearranging yields

$$\begin{aligned}&\mathbb{E}\|tv_t - \nabla F(\theta_{t-1})\|_2^2 = (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2(t-1)^2\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\|_2^2\\ &\qquad + 2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})\right\|_2^2 + 2(t-1)^2\mathbb{E}\left\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\rangle. \quad (4.40)\end{aligned}$$

We split the remainder of our analysis into two cases, corresponding to the **LSN** case or the **ISC** case. The difference in the analysis lies in how we handle the term $\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle$.

Analysis in the LSN case:

From L-Lipschitz smoothness of F in Assumption 1, we have

$$\begin{aligned}\left\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle &= -\frac{1}{\eta}\left\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle\\ &\le -\frac{1}{\eta L}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2.\end{aligned} \quad (4.41)$$

Now consider the inner product term $\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle$ in Eq. (4.40). We split it into two terms, and upper bound them using equations (4.43) and (4.41) respectively. Doing so yields:

$$\begin{aligned}&\mathbb{E}\|tv_t - \nabla F(\theta_{t-1})\|_2^2\\ &\le (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2(t-1)^2\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\|_2^2\\ &\quad + 2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})\right\|_2^2 + 2(t-1)^2\mathbb{E}\left\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle\\ &\le (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2(t-1)^2\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 + 2(t-1)^2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right\|_2^2\end{aligned}$$

$$
\begin{aligned}
&+2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2-\frac{3\eta\mu}{2}(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2-\frac{1}{2\eta L}(t-1)^2\mathbb{E}\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\|_2^2\\
\le&\left(1-\frac{3\eta\mu}{2}\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2+4(t-1)^2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^2\\
&-2(t-1)^2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^2\\
\le&\left(1-\frac{3\eta\mu}{2}+4\eta^2\ell_\Xi^2\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2-2(t-1)^2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^2.
\end{aligned}
$$

From the condition (4.23), we have $1-\frac{3}{2}\eta\mu+4\eta^2\ell_\Xi^2\le 1-\eta\mu$, which completes the proof.

Analysis in the ISC case:

We deal with the last summand in the last line of Eq. (4.40), where we use the iterated law of expectation to achieve

$$
\begin{aligned}
\mathbb{E}\left\langle v_{t-1},\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\rangle&=\mathbb{E}\left\langle v_{t-1},\mathbb{E}\left[\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\mid\mathscr{F}_{t-1}\right]\right\rangle\\
&=\mathbb{E}\left\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle.
\end{aligned}
$$

The update rule for v_t implies that $v_{t-1}=-\frac{\theta_{t-1}-\theta_{t-2}}{\eta}$ for all $t\ge T_0+1$. The following analysis uses various standard inequalities (c.f. §2.1 in [142]) that hold for individually convex and L_{max}-Lipschitz smooth functions. First, we have

$$
\begin{aligned}
\left\langle v_{t-1},\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\rangle&=-\frac{1}{\eta}\left\langle\theta_{t-1}-\theta_{t-2},\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\rangle\\
&\le-\frac{1}{\eta L_{max}}\left\|\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\|_2^2,
\end{aligned}
\tag{4.42}
$$

where the inequality follows from the Lipschitz condition. On the other hand, the μ-strong convexity of F implies that

$$
\begin{aligned}
\left\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle&=-\frac{1}{\eta}\left\langle\theta_{t-1}-\theta_{t-2},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle\\
&\le-\frac{\mu}{\eta}\|\theta_{t-1}-\theta_{t-2}\|_2^2=-\eta\mu\|v_{t-1}\|_2^2.
\end{aligned}
\tag{4.43}
$$

Plugging the bounds (4.42) and (4.43) into Eq. (4.40) yields

$$
\begin{aligned}
&\mathbb{E}\|tv_t-\nabla F(\theta_{t-1})\|_2^2\\
\le&(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+2(t-1)^2\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\|_2^2+2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})\right\|_2^2\\
&+(t-1)^2\mathbb{E}\left\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle+(t-1)^2\mathbb{E}\left\langle v_{t-1},\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\rangle\\
\le&(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+2(t-1)^2\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t)\right\|_2^2+2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&- \eta\mu(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 - \frac{1}{\eta L_{max}}(t-1)^2\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \\
&\leq (1-\eta\mu)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 - 2(t-1)^2\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2,
\end{aligned}
$$

where in the last inequality relies on the fact that $\eta \in (0, \frac{1}{4L_{max}}]$ (see Eq. (4.23)), leading to the bound (4.32b).

4.4.3 Proof of Lemma 4.8

We again split our analysis into two cases, corresponding to the **LSN** and **ISC** cases. Recall that the main difference is whether the Lipschitz stochastic noise condition holds (cf. Assumption 3), or the functions are individually convex and smooth (cf. Assumption 8).

Analysis in the LSN case:

From the ℓ_Ξ-Lipschitz smoothness of the stochastic gradients (Assumption 3) and the μ-strong-convexity of F (Assumption 1), we have

$$
\begin{aligned}
\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 &\leq 2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\
&\leq 2\ell_\Xi^2\mathbb{E}\|\theta_{t-1} - \theta^*\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\
&\leq \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2,
\end{aligned}
\tag{4.44}
$$

which establishes the claim.

Analysis in the ISC case:

Using Assumption 8 and standard inequalities for L_{max}-smooth and convex functions yields

$$
f(\theta^*;\xi) + \langle \nabla f(\theta^*;\xi), \theta\rangle + \frac{1}{2L_{max}}\|\nabla f(\theta;\xi) - \nabla f(\theta^*;\xi)\|_2^2 \leq f(\theta;\xi).
$$

Taking expectations in this inequality and performing some algebra[8] yields

$$
\begin{aligned}
\mathbb{E}\|\nabla f(\theta;\xi) - \nabla f(\theta^*;\xi)\|_2^2 &= 2L_{max}\langle \mathbb{E}[\nabla f(\theta^*;\xi)], \theta\rangle + \mathbb{E}\|\nabla f(\theta;\xi) - \nabla f(\theta^*;\xi)\|_2^2 \\
&\leq 2L_{max}\mathbb{E}\left[f(\theta;\xi) - f(\theta^*;\xi)\right]
\end{aligned}
$$

[8] In performing this algebra, we assume exchangeability of gradient and expectation operators, which is guaranteed because the function $x \mapsto \nabla f(x;\xi)$ is L_{max}-Lipschitz for a.s. ξ.

$$= 2L_{max}\left[F(\theta) - F(\theta^*)\right].$$

Recall that $\nabla F(\theta^*) = 0$ since θ^* is a minimizer of F. Using this fact and the μ-strong convexity condition, we have $F(\theta) - F(\theta^*) \le \frac{1}{2\mu}\|\nabla F(\theta)\|_2^2$. Substituting back into our earlier inequality yields

$$\mathbb{E}\|\nabla f(\theta;\xi) - \nabla f(\theta^*;\xi)\|_2^2 \le \frac{L_{max}}{\mu}\|\nabla F(\theta)\|_2^2.$$

We also note that[9]

$$\begin{aligned}
\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 &= \mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t) - [\nabla F(\theta_{t-1}) - \nabla F(\theta^*)]\|_2^2 \\
&\le \mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)\|_2^2 \\
&\le \frac{L_{max}}{\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2.
\end{aligned}$$

Finally, applying the argument of (4.44) yields the claim (4.33).

4.5 Comparison to related works

In this section, we provide a careful comparison of our convergence results to those for stochastic first-order gradient algorithms. For all nonasymptotic results, we compare our algorithm results with that of vanilla stochastic gradient descent, possibly equipped with iteration averaging and variance-reduced stochastic first-order optimization algorithms. In the Lipschitz continuous Hessian case, we can achieve asymptotic unity. We compare our ROOT-SGD convergence result with comparative work along with the following discussions in three aspects:

Comparison with classical results on SGD and its acceleration

SGD is known to be worst-case optimal for optimizing smooth and strongly-convex objectives up to a constant pre-factor. By way of contrast, our convergence metric in use, oracle query model and assumption on stochasticity are fundamentally different. Despite this, a recent work due to [145] building upon earlier analysis surveyed by [28] makes a comparable noise assumption that allows the noise variance to grow at most quadratically with the distance to optimality and applies to SGD. [145] shows that for appropriate diminishing step-sizes η_t we can conclude a guarantee of $\mathbb{E}\|\theta_T^{\mathrm{SGD}} - \theta^*\|_2^2 \lesssim \frac{\sigma_*^2}{\mu^2 T}$. With additional smoothness and noise assumptions we aim to achieve fine-grained non-asymptotic and asymptotic local asymp-

[9] This proof strategy is forklore and appears elsewhere in the variance-reduction literature; see, e.g., the proof of Theorem 1 in [91], and also adopted by [145, 146].

totic minimax optimality *with unity pre-factor*. It is straightforward to observe that the convergence rate bound of SGD under shared assumptions is in no regime better than that of ROOT-SGD presented in (4.13).

We turn to compare our ROOT-SGD convergence result with the existing arts on stochastic accelerated gradient descent for strongly convex objectives [76, 77]. With an appropriate multi-epoch design, their guarantees on the objective gap are worst-case optimal for optimizing smooth and strongly-convex objectives in terms of the dependency on *all* terms of condition number L/μ and a uniform upper bound on the noise variance. Our preliminary nonasymptotic guarantee for single-epoch ROOT-SGD in Theorem 4.1, in contrast, does not require uniform boundedness on the variance and depends solely on the variance at the minimizer θ^*.[10] That being said, our result cannot not admit an accelerated rate in terms of condition number $\sqrt{L/\mu}$ even with the help of multi-epoch designs. It is an important direction of future research to incorporate acceleration mechanism into our framework so as to achieve all-regime optimality.

Comparison with near-optimal guarantees in gradient norm

[7] develops a multi-epoch variant of SGD with averaging (under the name SGD3) via recursive regularization techniques and achieved a near-optimal rate for attaining an estimator of $O(\varepsilon)$-gradient norm. Our assumptions are not comparable in general: on the one hand, we assume a second-moment version of stochastic Lipschitz assumption (assumption 3), which makes it possible to establish guarantees that depends on the noise variance σ_* at the optimum θ^*; on the other hand, [7] makes no assumption on the modulus of continuity for the stochastic gradient, while their bound depends on a uniform upper bound σ on the noise variance. Besides, it is worth noticing that as $\varepsilon \to 0^+$, the leading-order term in their bound scales as $\frac{\sigma^2}{\varepsilon^2}\log^3\left(\frac{L}{\mu}\right)$, which is sub-optimal by a polylogarithmic factor even if the uniform boundedness assumption on the variance is satisfied. In a subsequent work, [72] applies the idea of recursive regularization to AC-SA [76] and achieves an accelerated rate to find an approximate minimizer with ε-gradient norm, while a lower bound analysis provided further justifies the necessity of a multiplicative logarithmic factor in the nonstrongly convex, local oracle setting.

It is also worth-mentioning the (Inexact) SARAH algorithm and its analysis developed by [146], which also achieves a near-optimal complexity upper bound of $O\left(\frac{\sigma_*^2}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)+\frac{L_{max}}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$ to obtain an approximate minimizer with ε-gradient norm. Note that the setting for that result is slightly different (their setting is dubbed as the **ISC** case in our extended analysis of Theorem 4.5 in §4.4). The algorithm of [146] requires random output and burn-in batches that is inversely dependent on the desired accuracy ε, yielding a logarithmic pre-factor on top of the statistical error corresponding to the Cramér-Rao lower bound; in comparison, our preliminary

[10] When measuring the risk via gradient norm, the optimal risk is characterized by the gradient noise variance σ_* at θ^*.

single-epoch ROOT-SGD result has a leading-order term in complexity bound that removes a logarithmic factor, attaining an optimal non-asymptotic guarantee up to a nonunity pre-factor.

Nonasymptotic guarantees matching local asymptotic minimax with near-unity pre-factor

[11] For SGD with PRJ averaging procedure, [15] present a convergence rate that provides a useful point of comparison, although the assumptions are different (no Lipschitz gradient, bounded variance). In particular, when choosing the step-size $\eta_t = Ct^{-\overline{\alpha}}$ for $\overline{\alpha} \in (1/2, 1)$, [15] show that the following bound holds true for the averaged iterates $\overline{\theta}_T$ for the PRJ:

$$\sqrt{\mathbb{E}\left\|\overline{\theta}_T - \theta^*\right\|_2^2} - \sqrt{\frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{T}} \le \frac{c_0}{T^{2/3}},$$

which corresponds to an $O(n^{-7/6})$ additional term in the squared estimation error metric ((4.19a) in Corollary 4.3). Here, the constant c_0 depends on the initial distance to optimum, smoothness and strong convexity parameters of second- and third-order derivatives, as well as higher-order moments of the noise. [189, 74] further improves the higher-order term from $O(n^{-7/6})$ to $O(n^{-5/4})$. The convergence rate of (single-loop) ROOT-SGD is similar to SGD with PRJ averaging procedure in the nature of the leading term and the high-order terms, but our convergence rate bound of ROOT-SGD is comparatively cleaner and easier to interpret.

The work by [73] proposes the Streaming SVRG algorithm that provides nonasymptotic guarantees in terms of the objective gap. Under a slightly different setting where smoothness and convexity assumptions are imposed on the individual function, their objective gap bound asymptotically matches the optimal risk achieved by the empirical risk minimizer under an additional self-concordance condition, with a multiplicative constant that can be made arbitrarily small. In particular, via our notations their results take the following form:

$$\mathbb{E}\left[F(\widehat{\theta}_n) - F(\theta^*)\right] \le \left(1 + \frac{5}{b}\right)\frac{1}{2n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right) + \text{high-order terms},$$

where they require $n \ge b^{2p+3}$ for some $p \ge 2$. In order to achieve the sharp pre-factor, the additional term in this bound is at least $\Omega(n^{-8/7})$, a worse rate than our Corollary 4.3. Additionally, to get the corresponding nonasymptotic guarantees under such a setting, their bound requires a scaling condition $T \gtrsim \frac{L_{max}^2}{\mu^2}$ where L_{max} denotes the smoothness of the individual function, which is larger than our burn-in sample size. Without the self-concordance condition, the convergence rate bound of Streaming SVRG suffers from an extra multiplicative factor belonging to the

[11] For convenience we include all comparable results in Table 4.1.

Algorithm	Assumption	Additional Term	Reference
PRJ	Hessian Lipschitz	$O\left(\frac{1}{n^{7/6}}\right)$	[15]
PRJ	Hessian Lipschitz	$O\left(\frac{1}{n^{5/4}}\right)$	[189, 74]
Streaming SVRG	Self-concordant	multiplicative[12]	[73]
ROOT-SGD	Hessian Lipschitz	$O\left(\frac{1}{n^{3/2}}\right)$	(Work in this chapter)

Table 4.1 Comparison of our results with comparative work. For the unity pre-factor nonasymptotic result, we only characterize the additional term to the optimal risk.

interval $\left[1, \frac{L_{max}}{\mu}\right]$, and its leading-order term thereby admits a dependency on the condition number worse than SGD.

4.6 Future directions

We have shown that ROOT-SGD enjoys favorable asymptotic and nonasymptotic behavior for solving the stochastic optimization problem (0.1) in the smooth, strongly convex case. With this result in hand, several promising future directions arise. First, it is natural to extend the results for ROOT-SGD to non-strongly convex and nonconvex settings, for both nonasymptotic and asymptotic analyses. Second, it would also be of significant interest to investigate both the nonasymptotic bounds and asymptotic efficiency of the variance-reduced estimator of ROOT-SGD in Nesterov's acceleration setting, in the hope of achieving all regime optimality in terms of the sample complexity to the stochastic first-order oracle. Finally, for statistical inference using online samples, the near-unity nonasymptotic and asymptotic results presented in this chapter can potentially yield confidence intervals and other inferential assertions for the use of ROOT-SGD estimators.

Chapter 5
ROOT-SGD with Adaptive, Diminishing Stepsize for Statistically Efficient Stochastic Optimization

We revisit ROOT-SGD, a novel method for stochastic optimization to bridge the gap between stochastic optimization and statistical efficiency with sharp convergence guarantees. Our method integrates a well-designed *diminishing stepsize strategy*, addressing key challenges in optimization, and providing robust theoretical guarantees and practical improvements. We demonstrate that ROOT-SGD with a diminishing stepsize achieves optimal convergence rates while maintaining computational efficiency. By dynamically adapting the stepsize sequence, ROOT-SGD ensures improved stability and precision throughout the optimization process. The results offer valuable insights into developing advanced algorithms that are both computationally efficient and statistically robust.

Keywords:

Stochastic Optimization, ROOT-SGD Algorithm, Diminishing Stepsize, Statistical Efficiency, Non-Asymptotic Convergence Rates

5.1 Introduction

Continuing our discussions in before, given a function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ that is differentiable as a function of its first argument, consider the unconstrained minimization problem

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{for a function of the form } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big] \tag{5.1}$$

Here the expectation is taken over a random vector $\xi \in \Xi$ with distribution $\mathbb{P}$. Throughout this chapter, we consider the case where F is strongly convex and smooth. Suppose that we have access to an oracle that generates samples $\xi \sim \mathbb{P}$. Let θ^* denote the minimizer of F, we defined the matrices $H^* := \nabla^2 F(\theta^*)$ and $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top\right]$. Under certain regularity assumptions, given

$(\xi_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, the following asymptotic limit holds true for the exact minimizer of empirical risk:

$$\widehat{\theta}_n^{\mathrm{ERM}} := \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n f(\theta, \xi_i) \quad \text{satisfies } \sqrt{n}\left(\widehat{\theta}_n^{\mathrm{ERM}} - \theta^*\right) \xrightarrow{d} \mathscr{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right) \tag{5.2}$$

Furthermore, the asymptotic distribution (5.2) is known to be *locally optimal*—see [173] and [57] for the precise statements about the optimality claim. The question naturally arises: ***can a stochastic optimization algorithm, taking the sample ξ_i as input in its i-th iteration without storing it, achieve the optimal guarantee as in equation*** (5.2)***?***

An affirmative answer to this question at least qualitatively, is provided by the seminal work by [150, 149, 155]. In particular, they show that by taking the Cesáro-average of the stochastic gradient descent (SGD) iterates, one can obtain an optimal estimator that achieves locally minimax limit (5.2), as the number of samples grows to infinity. This algorithm lays the foundations of online statistical inference [37, 165] and fine-grained error guarantees for stochastic optimization algorithms [136, 53]. However, the gap still exists between the averaged SGD algorithm and the exact minimizer of empirical risk, both asymptotically and non-asymptotically. The following questions remain unresolved:

- The asymptotic properties of the estimators produced by the Polyak-Ruppert algorithm are derived under the Lipschitz or Hölder condition of the Hessian matrix $\nabla^2 F$, at least with respect to the global optimum θ^* in all existing literature (see, e.g., [150, 57]). However, the asymptotic guarantee (5.2) for the exact minimizer holds true as long as the matrix-valued function $\nabla^2 F$ is *continuous* at θ^*, along with mild moment assumptions (see, e.g., [173]). On a historical note, the mis-match in the assumptions is particularly undesirable, given a large portion of literature is devoted to identify the optimal smoothness conditions required for the asymptotic normality of M-estimators to admit [106, 173]. ***Is there a (single-loop) stochastic optimization algorithm that achieves the asymptotic guarantee*** (5.2) ***under the mildest smoothness conditions including that the Hessian is continuous but not Hölder continuous at its global optimum?***

- On the non-asymptotic side, one would hope to prove a finite-sample upper bound for the estimator produced by the stochastic optimization algorithm under proper smoothness condition, which matches the exact behavior of the asymptotic Gaussian limit (5.2) with additional terms that decays faster as $n \to +\infty$. For example, under the one-point Hessian Lipschitz condition, [136, 189, 74] established bounds in the form of

$$\mathbb{E}\left\|\widehat{\theta}_n^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \leq \frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + \text{high order terms} \tag{5.3}$$

for the Polyak-Ruppert estimator $\widehat{\theta}_n^{\mathrm{PRJ}}$. Under the optimal trade-off, the higher-order terms in their bound scale at the order $O(n^{-7/6})$ and $O(n^{-5/4})$, respectively. Compared to the rates for the M-estimator, these bounds on the additional term do not appear to be sharp or optimal. Under suitable Lipschitz conditions, the natural scaling for the additional term would scale as $O(n^{-3/2})$ (see the discussion following Theorem 5.4 for details). For quadratic objectives, an argument similar to [114] allows one to achieve an $O(n^{-3/2})$ higher-order term

$$\mathbb{E}\left\|\widehat{\theta}_n^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \leq \frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + O\left(\frac{1}{n^{3/2}}\right) \tag{5.4}$$

with a sharp dependency on problem-specific constants. However, the design requires prior knowledge of the total number of observations n, which can limit its practicality.[1] *The question of whether an algorithm exists that is agnostic to n remains open.*

We answer both questions affirmatively using ROOT-SGD with a diminishing stepsize strategy. In the following, we describe the algorithm and explain the connection and differences between our results and [114].

The ROOT-SGD algorithm with varying stepsizes

For the stochastic optimization problem in the strongly-convex and mean-squared smooth setup, [114] recently proposed a stochastic approximation algorithm named *Recursive One-Over-T SGD*, or ROOT-SGD for short. To recap at each iteration $t = 1, 2, \ldots$ ROOT-SGD performs the following steps:

- receives an sample $\xi_t \sim \mathbb{P}$, and
- performs the updates

$$v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right) \tag{5.5a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t \tag{5.5b}$$

for a suitably chosen sequence $\{\eta_t\}_{t=1}^{\infty}$ of positive stepsizes.

For the purposes of stabilizing the iterates, Algorithm (5.5) is initialized with a *burn-in* phase of length $T_0 > 1$, in which only the v variable is updated with the θ variable held fixed. Given some initial vector $\theta_0 \in \mathbb{R}^d$, we set $\theta_t = \theta_0$ for all $t = 1, \ldots, T_0$, and compute

$$v_t = \frac{1}{t}\sum_{s=1}^{t} \nabla f(\theta_0, \xi_s) \qquad \text{for all } t = 1, \ldots, T_0$$

[1] This concept is also known as the *anytime property* in some literature, such as [74]. We will demonstrate that the standard *doubling strategy* in optimization is inadequate and that a carefully designed multi-loop strategy is necessary.

The last iterate θ_t is used as the output of the algorithm.

[114] analyzed this algorithm when it is run with a constant stepsize, and showed that ROOT-SGD simultaneously achieves non-asymptotic convergence rates and asymptotic normality with a near-optimal covariance. While the asymptotic limit includes the optimal quantity, it also includes an additional term due to the stepsize choice. In this chapter, we provide a sharper analysis that yields non-asymptotic bounds matching the asymptotic behavior in its leading-order term, with higer-order additional terms being sharp and state-of-the-art. Our work is also motivated by the practical question of stepsize schedule in ROOT-SGD. The asymptotic and non-asymptotic guarantees are established for a spectrum of rate of decaying stepsizes. The optimal trade-off between fast convergence and well-behaved limiting variance is also addressed, leading to the optimal choice of stepsize sequences under different regimes. In significance, our diminishing stepsize sequence requires no prior knowledge of n in advance.

Building upon the proof techniques in the non-asymptotic bounds of [114], our work provide fine-grained guarantees for ROOT-SGD, addressing both aforementioned questions immediately before introducing ROOT-SGD with affirmative answers. A key technical novelty is a two-time-scale characterization of the iterates (5.5) for a diminishing stepsize strategy. This allows us to effectively bound various cross terms in the error decomposition, yielding better bounds than those obtained by naïve application of Young's inequality. In addition, we also propose an improved re-starting schedule for the multi-loop algorithm, achieving exponential forgetting of the initial condition without affecting the statistical efficiency on its leading order term.

5.1.1 Contribution and organization

Let us summarize the contributions of this chapter:

- On the asymptotic side, we show in Theorem 5.1 that ROOT-SGD with a wide range of diminishing stepsize sequence converges asymptotically to the optimal Gaussian limit as $n \to +\infty$. Notably, this result only requires strong convexity, smoothness, and a set of noise moment assumptions standard in asymptotic statistics. The result does not require any higher-order smoothness other than the continuity of Hessian matrix at θ^*, another standard condition for asymptotic normality. To our knowledge, this provides a first result for a stochastic approximation algorithm that enjoys asymptotic optimality without additional smoothness conditions and the prior knowledge of n.

- On the contrary, we show that without additional smoothness conditions, a constant stepsize variant of Polyak-Ruppert algorithm fails to converge at a desirable rate, for any feasible scalings of stepsize and burn-in time choices. This manifests the difference in asymptotics between variance-reduced methods and

Polyak-Ruppert averaging methods. The result is stated in Theorem 5.2 serving as complementary to the asymptotic Theorem 5.1.

- Under the same set of assumptions, in Theorem 5.3, we establish a non-asymptotic gradient norm upper bound with the optimal leading term that exactly matches the optimal asymptotic risk, plus a higher-order term that scales as $O(n^{-4/3})$. When restarting is employed with an appropriate schedule, the resulting upper bound measured in gradient norm is of unity prefactor (arbitrarily close to 1) of the optimal asymptotic risk, with exponentially-decaying additional terms.
- In addition, when the one-point Hessian Lipschitz at the global optimum θ^* and certain fourth-moment conditions are assumed, in Theorem 5.4, we show an upper bound on the mean-squared error (MSE) in the form of (5.3). Taking an optimal trade-off leads to a higher-order term that scales as $O(n^{-3/2})$ as $n \to +\infty$ with a sharp problem-specific prefactor, and such a bound is achieved without the prior knowledge of n. With some efforts, we also establish a similar upper bound on the excess risk in Theorem 5.5.

5.1.2 Additional related works

Gradient descent and stochastic gradient descent methods have gained unprecedented popularity in the past decade amidst the era of big data [24, 30, 28], driven by the rapid growth of deep learning applications [80]. These methods excel in handling large-scale datasets due to their efficient processing of online samples. A myriad of variants have emerged from both theoretical advancements and practical needs, including variance-reduced methods [107, 91, 48], momentum-accelerated methods [140, 20], second-order methods [51, 143], adaptive gradient methods [56, 96], iteration averaging [155, 150], and coordinate descent [184], among others. The Polyak-Ruppert iteration averaging method [150, 149, 155] and its generalized form [102] have been shown to enhance robustness with respect to step size selection, achieving asymptotic normality with optimal covariance matching local minimax optimality [206, 57]. Recent studies have further explored the non-asymptotic behavior of stochastic gradient descent with iteration averaging [136, 189, 16, 14, 71, 74, 54, 53]. In the studies of linear regression and stochastic approximation, [201, 88, 87] have analyzed the ”tail-averaging” technique, achieving exponential forgetting and optimal statistical risk simultaneously. [103] investigates the Ruppert-Polyak averaging method for general linear stochastic approximation, which extends beyond optimization algorithms to applications in reinforcement learning. Under more stringent noise conditions, [134] establishes Gaussian limit and concentration inequalities for constant stepsize algorithms, with related advancements discussed in [115].

The weak convergence result from [150] has recently been generalized to functional weak convergence by [108] and [117] within the framework of i.i.d. online convex stochastic optimization. However, applying this to nonlinear stochastic approximation with Markovian data introduces several challenges that need addressing [55, 93, 137, 187, 119, 154, 157]. Referenced works beyond this overview delve deeper into topics such as asymptotic normality, statistical inference using gradient-based methods, and variants thereof [170, 169, 116, 121, 162, 94, 34, 161, 197, 90, 207, 40, 138, 127, 130, 135, 181, 118, 185, 204, 36, 85, 205].

The asymptotic efficiency of variance-reduced stochastic approximation methods has been relatively underexplored in research. [73] introduces an online variant of the SVRG algorithm [91] and establishes a non-asymptotic upper bound on excess risk, aligning its leading term with optimal asymptotics under specific self-concordant conditions on the objective function. [11] proposes *Implicit Gradient Transportation* (IGT) to reduce algorithmic variance. In the context of reinforcement learning for policy evaluation, [95, 133] provides an instance-dependent non-asymptotic upper bound on ℓ_∞ estimation error for variance-reduced stochastic approximation algorithms, matching the risk of the optimal Gaussian limit up to constant or logarithmic factors. Central to our study, [114] introduces the ROOT-SGD algorithm that achieves local minimax optimality. This algorithm can be viewed as an online variant of SARAH [144] and connects with extrapolation-smoothing methods like (N)IGT and STORM [11, 46, 45]. In a different approach, [141, 186, 110] propose dual averaging for the regularized or proximal case.[2] ROOT-SGD distinguishes itself by averaging past stochastic gradients with proper de-bias corrections, achieving both statistical efficiency and non-asymptotic high-order terms.

Organization:

This chapter is organized as follows. §5.2 describes the asymptotic normality results of ROOT-SGD and also the sub-optimality of Polyak-Ruppert averaging under the Hessian continuity assumption at the optimum. §5.3 state the non-asymptotic upper bound results on the gradient norm and also the estimation error. We prove the non-asymptotic upper bounds with sharp pre-factors in §5.5. In §5.4, we prove the asymptotic results, establishing optimality of ROOT-SGD and sub-optimality of Polyak-Ruppert averaging without high-order smoothness conditions. We finalize the chapter with some discussions in §5.6.

Notations:

Given a pair of vectors $u, v \in \mathbb{R}^d$, we write $\langle u, v\rangle = \sum_{j=1}^d u_j v_j$ for the inner product, and $\|v\|_2$ for the Euclidean norm. For a matrix M, the operator norm is defined as $|\!|\!|M|\!|\!|_{\mathrm{op}} := \sup_{\|v\|_2=1} \|Mv\|_2$. For scalars $a, b \in \mathbb{R}$, we adopt the shorthand notation $a \wedge b := \min(a,b)$ and $a \vee b := \max(a,b)$. Throughout the chapter, we use the σ-

[2] See also [57, 172] for manifold first-order optimization methods.

fields $\mathscr{F}_t := \sigma(\xi_1, \xi_2, \cdots, \xi_t)$ for any $t \geq 0$. Due to the burn-in period T_0 introduced before, the stochastic processes are indexed from time $t = T_0$. Given vector-valued martingales $(X_t)_{t \geq T_0}, (Y_t)_{t \geq T_0}$ adapted to the filtration $(\mathscr{F}_t)_{t \geq T_0}$, we use the following notation for cross variation for $t \geq T_0$:

$$[X,Y]_t := \sum_{s=T_0+1}^{t} \langle X_t - X_{t-1}, Y_t - Y_{t-1} \rangle$$

We also define $[X]_t := [X,X]_t$ to be the quadratic variation of the process $(X_t)_{t \geq T_0}$.

5.2 Asymptotic results

In this section, we present the asymptotic guarantees for ROOT-SGD and a counter-example for the Polyak-Ruppert algorithm, both under weak smoothness assumptions. We first describe the assumptions on the objective function F and associated stochastic oracles. We define the noise term

$$\varepsilon(\theta;\xi) = \nabla_\theta f(\theta;\xi) - \nabla F(\theta) \tag{5.6}$$

for each $\theta \in \mathbb{R}^d$. We also use the shorthand notation $\varepsilon_t(\theta) := \varepsilon(\theta;\xi_t)$. Throughout this section and the next non-asymptotic section, we make the following assumptions:

Assumption 9 *The population objective function F is μ-strongly-convex and L-smooth.*

Assumption 10 *The noise function $\theta \mapsto \nabla_\theta f(\theta, \xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\,\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\|_2^2 \leq \ell_\Xi^2\,\|\theta_1 - \theta_2\|_2^2 \qquad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d \tag{5.7}$$

Assumption 11 *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has a positive definite covariance matrix; hence $\sigma_*^2 := \mathbb{E}\,\|\nabla f(\theta^*;\xi)\|_2^2$ is positive and finite.*

Assumption 12 *The Hessian matrix $\nabla^2 F(\theta)$ is continuous at the optimum θ^*, i.e.,*

$$\lim_{\theta \to \theta^*} |||\nabla^2 F(\theta) - \nabla^2 F(\theta^*)|||_{op} = 0$$

Assumption 10 (sometimes referred to as *mean-squared-smoothness*) as well as Assumptions 11 and 12 are standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see, e.g., [173], Theorem 5.21). They are satisfied by a broad class of statistical models and estimators. Note that we assume only the continuity of Hessian matrix at θ^*, without assuming any bounds on its modulus of continuity. This requires merely slightly more than second-order smoothness, and is usually considered as the minimal assumption needed in the general setup. The weak condition manifests the difference between ROOT-SGD and Polyak-Ruppert averaging procedure.

The strong convexity and smoothness Assumption 9 is a global condition stronger than those typically used in the asymptotic analysis of M-estimators. They are needed for the fast convergence of the optimization algorithm, and makes it possible to establish non-asymptotic bounds. Finally, we note that in making Assumption 10, we separate the stochastic smoothness of the noise $\varepsilon(\theta,\xi)=\nabla f(\theta,\xi)-\nabla F(\theta)$ with the smoothness of the population-level objective itself. The magnitude of ℓ_Ξ and L is not comparable in general. This flexibility allows, for example, mini-batch algorithms where the population-level Lipschitz constant L remains the same but the parameter ℓ_Ξ decreases with batch-size. This setting is called *Lipschitz stochastic noise* (LSN) in [114], which requires weaker conditions than the *individual smooth and convex* (ISC) setting in their chapter.

5.2.1 Asymptotic normality

Under the conditions above, we are ready to state our asymptotic guarantees.

Theorem 5.1. *Under Assumptions 9, 10 and 11, there exists universal constants $c,c_1>0$, such that for any $\alpha\in(0,1)$, ROOT-SGD with burn-in time $T_0=c(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2})$ and stepsize sequence $\eta_t=\frac{1}{\mu T_0{}^{1-\alpha}t^\alpha}$ for $t\geq T_0$ satisfies the asymptotic limit:*

$$\sqrt{T}(\theta_T-\theta^*)\xrightarrow{d}\mathcal{N}\left(0,(H^*)^{-1}\Sigma^*(H^*)^{-1}\right)$$

where $H^*:=\nabla^2F(\theta^*)$ and $\Sigma^*:=\mathbb{E}(\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top)$.

See §5.4.2 for the proof of this theorem. En route to the proof of this asymptotic guarantee, we establish non-asymptotic bounds on the second moments of the processes $(\theta_t,v_t,z_t)_{t\geq T_0}$, where a central object in our analysis is the *tracking error process*:

$$z_t:=v_t-\nabla F(\theta_{t-1})\qquad\text{for } t\geq T_0 \tag{5.8}$$

We first establish the following (non-sharp) bound on the moments of processes z_t and v_t. Despite the worse multiplicative constants, this bound serves as a starting point of the *sharp* inequalities with the constant being unity.

Proposition 5.1. *Under Assumptions 9, 10, and 11, there exist universal constants $c_1,c_2,C>0$, using burn-in time $T_0\geq C(\frac{\ell_\Xi^2}{\mu^2}+\frac{L}{\mu})$, if the step sequence is non-increasing, and $\frac{c_1}{\mu t}<\eta_t<c_2(\frac{\mu}{\ell_\Xi^2}\wedge\frac{1}{L})$ when $t>T_0$. We have the following bounds for any $T\geq 2T_0\log T_0$:*

$$\mathbb{E}\,\|z_T\|_2^2 \leq C\left(\frac{\sigma_*^2}{T} + \frac{\ell_\Xi^2 T_0 \log T}{\mu^2 T^2}\,\|\nabla F(\theta_0)\|_2^2\right) \qquad \textit{and} \quad \mathbb{E}\,\|v_T\|_2^2 \leq C\left(\frac{\sigma_*^2}{\mu \eta_T T^2} + \frac{T_0}{\mu^2 T^3 \eta_T^2}\,\|\nabla F(\theta_0)\|_2^2\right)$$

See §5.4.1 for the proof of this claim.

A few remarks are in order. First, we note that this limiting distribution is locally asymptotically optimal (see, e.g., [57]). This result for diminishing stepsize sequence is complementary to the constant-stepsize result in the chapter [114], where the asymptotic covariance is inflated by a stepsize-dependent matrix.[3] Moreover, our method achieves optimal asymptotic covariance in a single loop and is agnostic to the knowledge of n in advance, enhancing its practicality. Theorem 5.4 allows for flexible choice of stepsize decaying rate $\alpha \in (0,1)$, albeit requiring knowledge about the structural parameters (L, ℓ_Ξ, μ). This requirement, on the other hand, can be relaxed with some efforts: given a stepsize sequence $\eta_t = h_0 t^{-\alpha}$ for some $h_0 > 0$ and arbitrary constant burn-in time, the iterates may suffer from exponential blow-up for constant number of steps, but will eventually decay at the desired rate, leading to the same asymptotic results. We omit this for simplicty. In contrast to the asymptotic guarantees by the Polyak-Ruppert averaging scheme [150, 155], Theorem 5.1 requires no quantitative Lipschitz or Hölder assumptions on the Hessian matrix ∇F, while requiring a stochastic continuity condition (Assumption 10) on the stochastic gradient. As we will see in the next sub-section, in contrast to our guarantees, the Polyak-Ruppert procedure is asymptotically sub-optimal for a function within the given class.

5.2.2 Asymptotic sub-optimality of Polyak-Ruppert averaging

In this section, we explicitly construct a problem instance under above set-up, for which Polyak-Ruppert procedure fails to converge to the optimal asymptotic distribution. In conjunction with Theorem 5.1, this exhibits an asymptotic separation between Polyak-Ruppert averaging and ROOT-SGD.

Specifically, we consider the following tail-averaged SGD estimator:

$$\theta_t = \theta_{t-1} - \eta_t \nabla f(\theta, \xi) \qquad \text{for } t = 1, 2, \ldots \tag{5.9a}$$

$$\overline{\theta}_T = \frac{1}{T - T_0} \sum_{t=T_0}^{T-1} \theta_t \tag{5.9b}$$

We consider a simple special case where the stepsize sequence is constant and fixed in advance, depending on the number of iterations in the algorithm. For the algorithm with T iterations, we consider stepsize $\eta_t = \eta = \eta_0 T^{-\alpha}$ for some constant $\eta_0 > 0$ and $t = 1, 2, \ldots$. This simplification makes the iterate (5.9a) a time-

[3] In the meantime, the asymptotic normality result for multi-loop ROOT-SGD in [114] admits a triangular array format ($n \to \infty$, $\eta \to 0$ with $\frac{\eta n}{\log(\eta^{-1})} \to \infty$), which can be difficult to interpret and impractical for practitioners, and undesirably necessitates knowledge of n aprior.

homogeneous Markov process, which is amendable to our analysis. Such a simplification has been employed in existing literature [14, 53], and the constant-stepsize algorithm usually behaves qualitatively similar to the one with diminishing stepsize $\eta_t = \eta_0 t^{-\alpha}$.

The following theorem shows the asymptotic sub-optimality of the estimator (5.9), even if started from the optimum, for any choice of burn-in period and step size.

Theorem 5.2. *There exists a function $F : \mathbb{R} \to \mathbb{R}$ that satisfies Assumptions 9 and 12 with constants $(\mu = 1, L = 2)$ and noise model $f(\cdot, \xi)$ satisfying Assumptions 10 and 11 with constants $(\ell_\Xi = 0, \sigma_* = 2)$. For any $\alpha \in [0,1)$, $\beta \in [0,1)$ and $\eta_0 > 0, S_0 > 0$, the procedure (5.9) starting from $\theta_0 = \theta^*$, with step size $\eta = \eta_0 T^{-\alpha}$ and burn-in time $T_0 = S_0 T^\beta$ leads to the following limit:*

$$\lim_{T \to +\infty} T \cdot \mathbb{E} \left\| \overline{\theta}_T - \theta^* \right\|_2^2 = +\infty \tag{5.10}$$

See §5.4.3 for the proof of this theorem.

Note that Theorem 5.2 shows that without the Hessian Lipschitz condition, the Polyak-Ruppert algorithm does not even converge with the desired rate, let alone the optimal asymptotic distribution. The proof is done via an explicit construction of a pathological function. With the Hessian Lipschitz condition removed, one could construct a strongly convex and smooth function, whose second derivative has a *sharp spike* at the optimum θ^*. This will break the local linearization arguments for the proof of Polyak-Ruppert algorithm. By employing recent progress in the analysis of MCMC algorithms [59], we can furthermore show that this leads to large bias that cannot be corrected using averaging. On the other hand, for ROOT-SGD, not only the asymptotic guarantees in Theorem 5.1 but also the non-asymptotic bounds on the gradient norm in Theorem 5.3 works. Moreover, note that [150] considered the case where the Hessian matrix is λ-Hölder at θ^*, and allows for stepsize choice $\eta_t \propto t^{-\alpha}$ for $\alpha \in [1-\lambda, 1)$. Theorem 5.2 can be extended to show that stepsize outside this range does not yield the correct rate. The construction we exploit, on the other hand, is by driving λ to 0 so that no stepsize choice is allowed.

5.3 Non-asymptotic results

In this section, we present the non-asymptotic results. We first establish sharp bounds on the gradient norm with near-unity pre-factor on the optimal complexity term, and exponentially decaying additional term. Then, we establish an estimation error bound with the pre-factor being unity and the additional term decaying as $n^{-3/2}$. Note that the former result holds true under exactly the same assumptions as needed in §5.2, while the latter requires additional conditions, as with existing literature [74, 136].

5.3.1 Upper bounds on the gradient norm

Recall the decomposition $\nabla F(\theta_t) = v_{t+1} - z_{t+1}$, it is easy to see that Proposition 5.1 implies the following bound on the gradient norm of the last iterate:

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq c\frac{\sigma_*^2}{T} + \frac{cT_0 \log T}{\mu^2 T^2}\left(\ell_\Xi^2 + \frac{1}{\eta_T^2}\right)\|\nabla F(\theta_0)\|_2^2$$

When taking largest possible stepsize $\eta = c\big(\frac{1}{L} \wedge \frac{\mu}{\ell_\Xi^2}\big)$, this bound matches the gradient norm bound in the original ROOT-SGDchapter [114], up to logarithmic factors in the high-order term. Our bound allows a more flexible choice of diminishing stepsizes. This flexibility allows us to achieve the exact asymptotically optimal limiting covariance, as opposed to the slightly larger covariance in the constant stepsize regime [114]. More importantly, this allows us to tune the stepsize sequence in order to address the optimal trade-off between fast convergence and small variance in the asymptotic limit. Note that the pre-factor in the leading term σ_*^2/T is *not* unity. However, owing to the inherent martingale structure in the process $(z_t)_{t \geq T_0}$,[4] one could extract the main part of the variance and bound the additional parts using Proposition 5.1. The multiplicative constant in such bounds will only contribute to the high-order terms in the final conclusion. See Theorem 5.3 and its proofs for details.

Note that the bounds in Proposition 5.1 depends on the initial condition $\|\nabla F(\theta_0)\|_2^2$ with polynomially-decaying factor T^{-2} and $T^{-3}\eta_T^{-2}$. For the algorithm ROOT-SGD, this cannot be avoided in general, as the stochastic gradients from initial rounds are being counted in the averaging process. On the other hand, this issue can be easily mitigated by *re-starting* the process for a few epochs. In Algorithm 4, we present a cold-start version of the algorithm. The algorithm consists of B short epochs and one long epoch. Each short epoch only uses constant number of data points, while the long epoch uses the rest of data points.

Theorem 5.3. *Under above set-up, given $\alpha \in (0,1)$, there exists constants $c_1 > 0$ depending only on α, such that the iterates* (5.5) *with any burn-in time $T_0 \geq c\big(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu}\big)$ and stepsize sequence $\eta_t = \frac{1}{c\mu {T_0}^{1-\alpha} t^\alpha}$ satisfies the bound:*

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \left(1 + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2} \wedge \alpha}\right) \cdot \frac{\sigma_*^2}{T} + c\log T \cdot \left(\frac{T_0}{T}\right)^{2 \wedge \frac{5-3\alpha}{2}} \|\nabla F(\theta_0)\|_2^2 \tag{5.11a}$$

Furthermore, for $B > \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2} \vee 1\right)$, the multi-loop estimator produced by Algorithm 4 satisfies the bound:

[4] It can be shown that the process $(tz_t)_{t \geq T_0}$ is a martingale adapted to the natural filtration (see §E.1 for details).

Algorithm 4 ROOT-SGD with cold start

```
Require: Burn-in time $T_0$, stepsize sequence $(\eta_t)_{t \geq T_0}$, number of restart epochs $B$, initial point $\theta_0$
Ensure: The last-iterate estimator $\widehat{\theta}_n$
Set initial point for first epoch $\theta_0^{(1)} = \theta_0$
for $b = 1, 2, \cdots, B$ do
    Run ROOT-SGD with burn-in time $T_0$, initial point $\theta_0^{(b)}$ and stepsize $\eta_t := \frac{c}{\mu T_0}$ for $T^\flat := cT_0 \log T_0$ iterations, and obtain the sequence $\left(\theta_t^{(b)}\right)_{t=T_0+1}^{T^\flat}$
    Set the initial point $\theta_0^{(b+1)} := \theta_{T^\flat}^{(b)}$ for the next round
end for
Run ROOT-SGD for $T := n - BT^\flat$ rounds with stepsize sequence $(\eta_t)_{t \geq T_0}$ and burn-in period $T_0$, and output the last iterate $\widehat{\theta}_n := \theta_T^{(B+1)}$
```

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_n)\right\|_2^2 \leq \left(1 + c\left(\frac{T_0}{n}\right)^{\frac{1-\alpha}{2} \wedge \alpha} \log^2 n\right) \frac{\sigma_*^2}{n} \tag{5.11b}$$

See §5.5.1 for the proof of this theorem.

A few remarks are in order. First, by taking $\alpha = 1/3$, for any constant $\omega \in (0,1)$, we can obtain an MSE bound on the gradient for the multi-loop estimator.

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_n)\right\|_2^2 \leq (1+\omega)\frac{\mathrm{Tr}(\Sigma^*)}{n} \qquad \text{for } n \geq \frac{c}{\omega^3}\left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right) \log^3 \frac{T_0}{\omega} \tag{5.12}$$

In other words, we obtain a near-optimal bound on the gradient norm with $(1+\omega)$ pre-factor compared to the asymptotic optimal limit, as long as the sample size is larger than the threshold $O\big(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\big)$, up to log factors. We remark that this threshold is also sharp: the term $O\big(\frac{L}{\mu}\big)$ is the number of iterations needed for gradient descent, while the $O\big(\frac{\ell_\Xi^2}{\mu^2}\big)$ term is the smallest sample size needed to distinguish the quadratic function $\frac{\mu}{2}\|x\|_2^2$ from the constant function 0, under the noise Assumption 10. This establish a gradient-norm result complementary to the sub-optimality gap bound in [73]. The gradient norm bound does not require the self-concordant condition needed in [73], and achieves a sharper convergence rate in terms of both the $(1+\omega)$ factor and the initial condition.[5]

With a potentially sub-optimal choice of $\alpha \in (0,1)$, one would get a worse exponent in the dependency of n on ω in the bound (5.12), while the rest parts of the bound remain unchanged. If ω is taken as a constant, the near-optimal bounds are available for the entire range of parameter $\alpha \in (0,1)$. Finally, we note that the bound (5.12) lead to an $\widetilde{O}(n^{-4/3})$ bound on the additional term, achieved by the stepsize choice $\eta_t = \frac{1}{c\mu T_0{}^{2/3} t^{1/3}}$. This rate and step-size choice, however, is not always optimal. In particular, as we will see in the next section, with the one-point

[5] The dependency on $\|\nabla F(\theta_0)\|_2$ decays exponentially fast and is omitted for simplicity.

Hessian Lipschitz condition on the objective function F, we can obtain an improved $\widetilde{O}(n^{-3/2})$ bound on the additional term.

5.3.2 Upper bounds on the estimation error

To obtain a precise upper bound for the estimation error $\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2$ that matches the asymptotic limit, we need the following one-point Hessian Lipschitz condition, as a quantitative counterpart of the continuity Assumption 12:

Assumption 12′ *There exists $L_2 > 0$, such that for any $\theta \in \mathbb{R}^d$, we have:*

$$|||\nabla^2 F(\theta) - \nabla^2 F(\theta^*)|||_{op} \leq L_2 \left\|\theta - \theta^*\right\|_2$$

Note that some form of quantitative description on the modulus of continuity of the Hessian matrix at θ^* is necessary to get any bound on the estimation error that scales as $\frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)$. If the Hessian can change sharply in a neighborhood of θ^*, the Hessian at this specific point will become irrelevant. Here, we make a standard one-point Hessian Lipschitz condition, while it is easy to extend our analysis to the case with one-point Hölder conditions.

We also need the following stronger fourth moment conditions for technical reasons. Note that these conditions are also exploited in prior works [136, 74].

Assumption 10′ *The noise function $\theta \mapsto \nabla_\theta f(\theta, \xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\left\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\right\|_2^4 \leq \ell_\Xi^4 \left\|\theta_1 - \theta_2\right\|_2^4 \qquad \textit{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d \tag{5.13}$$

Assumption 11′ *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has bounded fourth moment: $\widetilde{\sigma_*}^4 := \mathbb{E}\left\|\nabla f(\theta^*;\xi)\right\|_2^4$ is finite.*

By Hölder's inequality, it is clear that the constants in Assumptions 10′ and 11′ are larger than their second-moment counterparts, i.e., $\ell_\Xi \leq \ell_\Xi$ and $\sigma_* \leq \widetilde{\sigma_*}$.

Under the fourth moment conditions, we can establish the following fourth-moment bounds for the processes z_t and v_t, analogous to the second-moment results in Proposition 5.1.

Proposition 5.2. *Under Assumptions 9, 10′, and 11′, there exist universal constants $c_1, c_2, C > 0$, using burn-in time $T_0 \geq C(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu})$, if the step sequence is non-increasing, and $\frac{c_1}{\mu t} < \eta_t < c_2(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L})$ when $t > T_0$. We have the following bounds for any $T \geq 2T_0 \log T_0$:*

$$\mathbb{E}\left\|z_T\right\|_2^4 \leq C\left(\frac{\widetilde{\sigma_*}^2}{T} + \frac{\ell_\Xi^2 T_0 \log T}{\mu^2 T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right)^2 \qquad \textit{and} \quad \mathbb{E}\left\|v_T\right\|_2^4 \leq C\left(\frac{\widetilde{\sigma_*}^2}{\mu \eta_T T^2} + \frac{T_0}{\mu^2 T^3 \eta_T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right)^2$$

See §5.5.2 for the proof of this claim.

Compared to Proposition 5.1, the variance parameters (σ_*, ℓ_Ξ) are replaced with their fourth-moment counterparts $(\widetilde{\sigma_*}, \ell_\Xi)$. These fourth-moment estimates are utilized to control the error induced by approximation the estimation error $\theta_T - \theta^*$ using the pre-conditioned gradient $(H^*)^{-1}\nabla F(\theta_T)$. As with the case of Proposition 5.1, these terms appear only in the high-order terms of Theorem 5.4.

Now we are ready to present our main theorem, which provides the MSE bounds on the estimation error $\theta_T - \theta^*$, with the sharp pre-factor. To state the theorem, we define the following auxiliary quantities that appears in the high-order terms:

$$\mathscr{H}_T^{(\nabla)} := \log T \cdot \frac{\sigma_*^2}{T} \left(\frac{T_0}{T}\right)^{\alpha \wedge 1-\alpha} + \log T \cdot \|\nabla F(\theta_0)\|_2^2 \left(\frac{T_0}{T}\right)^{2 \wedge \frac{7}{2} - 2\alpha} \tag{5.14a}$$

$$\widetilde{r}_T := \frac{\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{T_0}{T}} \cdot \left(\|\nabla F(\theta_0)\|_2^4\right)^{1/4} \qquad \text{and} \tag{5.14b}$$

$$\mathscr{H}_n^{(\sigma)} := \frac{\sigma_*^2 \log^2 n}{\lambda_{\min}(H^*)^2 n} \left(\frac{T_0}{n}\right)^{\alpha \wedge 1-\alpha} + \frac{L_2 \widetilde{\sigma_*}^3 \log^2 n}{\lambda_{\min}(H^*)\mu^3 n^{3/2}} + \frac{L_2^2 \widetilde{\sigma_*}^4 \log^2 n}{\lambda_{\min}(H^*)^2 \mu^4 n^2} \tag{5.14c}$$

The term $\mathscr{H}_T^{(\nabla)}$ is part of the high-order term that appears in the bound for the gradient norm. It is indeed the upper bound for the *superfluous* part of the noise in the processes $(z_t)_{t \geq T_0}$ and $(v_t)_{t \geq T_0}$, without taking into account the cross term $\mathbb{E}\langle z_t, v_t \rangle$. The quantity $\widetilde{r}_T$ is a coarse upper bound on the convergence rate $\|\theta_t - \theta^*\|_2$ in terms of the fourth moment. In combination with the one-point Hessian Lipschitz Assumption 12′, this quantity controls the additional *linearization error* induced by relating the non-asymptotic behavior of the gradient to the iterates. Finally, the term $\mathscr{H}_n^{(\sigma)}$ is used to characterize the high-order terms for the error in the multi-loop estimator produced by Algorithm 4.

Theorem 5.4. *Under Assumptions 9, 10′, 11′ and 12′, there exists universal constant $c, c_1 > 0$, for burn-in-time $T_0 = c\left(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu}\right)$ and stepsize $\eta_t = \frac{c_1}{\mu {T_0}^{1-\alpha} t^\alpha}$ for $t \geq T_0$, we have the following bounds holding true for $t \geq 2T_0 \log T_0$:*

$$\mathbb{E}\|\theta_T - \theta^*\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\mathscr{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)^2} + \frac{cL_2\widetilde{r}_T^3}{\lambda_{\min}(H^*)} + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)^2} \tag{5.15a}$$

Furthermore, for $B \geq \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2} \vee 1\right)$, the multi-loop estimator by Algorithm 4 satisfies the bound

$$\mathbb{E}\left\|\widehat{\theta}_n - \theta^*\right\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{n} + c\mathscr{H}_n^{(\sigma)} \tag{5.15b}$$

See §5.5.3 for the proof of this theorem.

A few remarks are in order. First, we note that the asymptotically optimal $\frac{1}{n}\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)$ variance is achieved with the exact pre-factor 1. Taking the optimal stepsize choice with $\alpha = 1/2$, the high order term scales as $O(n^{-3/2})$ in both bounds (5.15a) and (5.15b). This is made possible by the stochastic Lipschitz condition for the gradient noise, and strictly improves existing bounds of $O(n^{-7/6})$ in the chapter [136] and the $O(n^{-5/4})$ bound in the chapter [189, 74]. It is easy to see that the bound (5.15b) is obtained by removing the terms depending on the initial condition, up to logarithmic factors in the additional term. This is natural because the initial condition is forgotten exponentially fast in the first B restarting epochs of Algorithm 4. Finally, when taking the optimal parameter $\alpha = 1/2$, the three high-order terms in the expression of $\mathscr{H}_n^{(\sigma)}$ have a clean interpretation.

- The first term $\widetilde{O}\left(\frac{\sigma_*^2\sqrt{T_0}}{\lambda_{\min}(H^*)^2 n^{3/2}}\right)$ characterizes the additional gradient noise collected in a neighborhood of θ^*. Since θ^* itself is unknown, the best possible estimator naturally take the average of gradient noise in a neighborhood around θ^* of radius $O\big(\frac{\sigma_*}{\mu\sqrt{n}}\big)$, which is the rate for estimating θ^*. Under Assumption 10, the variance for gradient noise at $\theta \in \mathbb{B}\big(\theta^*, \frac{\sigma_*}{\mu\sqrt{n}}\big)$, pre-conditioned with Hessian H^*, scales as:

$$\begin{aligned}
&\mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta)\right\|_2^2\\
&\leq \mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta^*)\right\|_2^2\\
&\quad + 2\sqrt{\mathbb{E}\left\|(H^*)^{-1}\big(\varepsilon_t(\theta)-\varepsilon_t(\theta^*)\big)\right\|_2^2\cdot\mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta^*)\right\|_2^2} + \mathbb{E}\left\|(H^*)^{-1}\big(\varepsilon_t(\theta)-\varepsilon_t(\theta^*)\big)\right\|_2^2\\
&\leq \mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big) + 2\frac{\ell_\Xi\sigma_*}{\lambda_{\min}(H^*)^2}\left\|\theta-\theta^*\right\|_2 + \frac{\ell_\Xi^2}{\lambda_{\min}(H^*)^2}\left\|\theta-\theta^*\right\|_2^2\\
&= \mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big) + O\left(\frac{\sigma_*^2\ell_\Xi}{\mu\lambda_{\min}^2(H^*)n^{3/2}}\right)
\end{aligned}$$

  The above derivations is tight in the worst case. Compared to the term $\widetilde{O}\left(\frac{\sigma_*^2\sqrt{T_0}}{\lambda_{\min}(H^*)^2 n^{3/2}}\right)$ in our bound (5.15b), the difference is that we replace $\frac{\ell_\Xi^2}{\mu^2}$ with $T_0 = c\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)$ and is optimal up to a polylogarithmic factor when $\frac{L}{\mu} \lesssim \frac{\ell_\Xi^2}{\mu^2}$.

- The rest two terms involves the one-point Hessian-Lipschitz parameter L_2. A natural linearization argument in the neighborhood of θ^* on the (generally non-linear) gradient function leads to these terms. In particular, simple calculus yields the following bounds:

$$\left\|(H^*)^{-1}\nabla F(\theta) - (\theta-\theta^*)\right\|_2 \leq \frac{L_2}{\lambda_{\min}(H^*)}\left\|\theta-\theta^*\right\|_2^2$$

Substituting with the $\mathbb{L}^4$ convergence rate for the iterates $\theta_T - \theta^*$ yields the bound on this linearization error, which matches the latter two terms in $\mathscr{H}_n^{(\sigma)}$.

The arguments in the proof of Theorem 5.4 indeed applies to any function that is *locally quadratic* around θ^*. Applying it to the function F itself, we arrive at the following theorem:

Theorem 5.5. *Under the same setup as in Theorem 5.4, we have the following bounds on the excess risk:*

$$\mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) \leq \frac{\operatorname{Tr}\left(\Sigma^*(H^*)^{-1}\right)}{2T} + \frac{c\mathscr{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)} + cL_2\widetilde{r}_T^3 + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)} \tag{5.16a}$$

and for the multi-loop estimator $\widehat{\theta}_n$ with $B \geq \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2} \vee 1\right)$, we have that

$$\mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) \leq \frac{\operatorname{Tr}\left(\Sigma^*(H^*)^{-1}\right)}{2n} + c\lambda_{\min}(H^*) \cdot \mathscr{H}_n^{(\sigma)} \tag{5.16b}$$

See §5.5.4 for the proof of this theorem.

Note that under the one-point Hessian-Lipschitz Assumption 12′, the leading-order term $\frac{\operatorname{Tr}\left(\Sigma^*(H^*)^{-1}\right)}{2n}$ is the asymptotic risk under the limiting Gaussian distribution. The high-order terms in Theorem 5.5 differ from those in Theorme 5.4 by a factor of $\lambda_{\min}(H^*)$. This bound replaces the self-concordance assumption in [73] with a less structural one-point Hessian-Lipschitz condition. Theorem 5.5 and their results are not comparable in general, as they are based on different assumptions. When taking the optimal trade-off, Theorem 5.5 leads to an $O(n^{-3/2})$ high-order term in addition to the sharp leading-order one. This result matches the bounds for ERM in [73], and improves the bounds for streaming SVRG in [73] in terms of the rate of convergence for the additional term.

5.4 Proof of asymptotic results

In this section, we present the proofs for the asymptotic results, Theorem 5.1 and Theorem 5.2. The former guarantees the asymptotic normality of ROOT-SGD under our assumptions, while the latter shows an example that satisfies our assumptions but makes Polyak-Ruppert algorithm fail asymptotically. En route our proof, in §5.4.1 we present the proof of Proposition 5.1, the non-asymptotic convergence rates for the process $(v_t)_{t\geq T_0}$ and $(z_t)_{t\geq T_0}$. This serves as the basic building block for the fine-grained asymptotic and non-asymptotic guarantees.

5.4.1 Proof of Proposition 5.1

Our main technical tools are the following two lemmas, which bound the second moments of v_t and z_t based on other parameters.

Lemma 5.6. *Under Assumption 9, 10, 11, when* $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{\ell_\Xi^2}$, *we have:*

$$\mathbb{E}\|v_t\|_2^2 \leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\|v_{t-1}\|_2^2 + \frac{26}{\mu\eta_{t-1}t^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2\sigma_*^2}{t^2}$$

For the process z_t, we have the following lemma which leads to an $O(1/\sqrt{t})$ bound.

Lemma 5.7. *Under Assumptions 9, 10 and 11, for* $t \geq 1$, *we have:*

$$\mathbb{E}\|z_t\|_2^2 \leq \frac{T_0{}^2\|z_0\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + \frac{2\ell_\Xi^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}\mathbb{E}\|\nabla F(\theta_s)\|_2^2 + \frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}s^2\eta_s^2\mathbb{E}\|v_s\|_2^2$$

The proofs of the Lemmas are postponed to Section E.1.1 and Section E.1.2 respectively. Given these lemmas, we now give a proof of this proposition.

We first note that for any $t \geq 2$ and $\eta_t < \frac{1}{2L}$, we have:

$$\begin{aligned}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 &\leq 2\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\mathbb{E}\|\nabla F(\theta_t) - \nabla F(\theta_{t-1})\|_2^2 \\ &\leq 2\mathbb{E}\|v_t - z_t\|_2^2 + 2L^2\eta_t^2\mathbb{E}\|v_t\|_2^2 \leq 6\mathbb{E}\|v_t\|_2^2 + 4\mathbb{E}\|z_t\|_2^2\end{aligned}$$

Therefore, by Lemma 5.6, if t and η_{t-1} satisfies $t\eta_{t-1}\mu > \frac{1}{4C}$, we obtain:

$$\begin{aligned}\mathbb{E}\|v_t\|_2^2 &\leq \left(1-\frac{\eta_{t-1}\mu}{2}\right)\left(1-\frac{1}{t}\right)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{C}{\mu\eta_{t-1}t^2}(\mathbb{E}\|v_{t-1}\|_2^2 + \mathbb{E}\|z_{t-1}\|_2^2) + \frac{2\sigma_*^2}{t^2} \\ &\leq \left(1-\frac{\eta_{t-1}\mu}{4}\right)\left(1-\frac{1}{t}\right)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{2C}{t^2\mu\eta_{t-1}}\mathbb{E}\|z_{t-1}\|_2^2 + \frac{2\sigma_*^2}{t^2}\end{aligned}$$

Consequently, we obtain:

$$t^2\mathbb{E}\|v_t\|_2^2 \leq (1-c\eta_{t-1}\mu)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{2C}{\mu\eta_{t-1}}\mathbb{E}\|z_{t-1}\|_2^2 + 2\sigma_*^2 \qquad (5.17)$$

for a universal constant $C > 0$.

Similarly, by Lemma 5.7, if s satisfies $s\eta_{s-1}\mu > \frac{1}{4C}$ for any $s > T_0$, we have:

$$\mathbb{E}\|z_t\|_2^2 \leq \frac{T_0{}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}(\mathbb{E}\|z_s\|_2^2 + \mathbb{E}\|v_s\|_2^2) + \frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}s^2\eta_s^2\mathbb{E}\|v_s\|_2^2$$

$$\leq \frac{{T_0}^2 \mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}\mathbb{E}\left\|z_s\right\|_2^2 + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}s^2\eta_s^2\mathbb{E}\left\|v_s\right\|_2^2 \tag{5.18}$$

for a universal constant $C' > 0$.

Note that the bounds (5.17) and (5.18) give recursive upper bounds on the second moments of the processes $(z_t)_{t\geq T_0}$ and $(v_t)_{t\geq T_0}$, i.e., they bound the quantities $\mathbb{E}\left\|z_t\right\|_2^2$ and $\mathbb{E}\left\|v_t\right\|_2^2$ based on their history. In the following, we solve the recursive inequalities.

We define the following quantities for $T \geq T_0$:

$$W_T := T^2\mathbb{E}\left\|v_T\right\|_2^2 \quad \text{and} \quad H_T := \sup_{T_0\leq t\leq T} t\mathbb{E}\left\|z_t\right\|_2^2$$

First, for any $T > T_0$, by taking the supremum in Eq (5.18) over $t \in [T_0, T]$, we obtain the following bound:

$$\begin{aligned}\sup_{T_0\leq t\leq T} t\mathbb{E}\left\|z_t\right\|_2^2 &\leq T_0\mathbb{E}\left\|z_{T_0}\right\|_2^2 + 2\sigma_*^2 + C\frac{\ell_\Xi^2}{\mu^2}\sup_{T_0\leq t\leq T}\frac{1}{t}\sum_{s=T_0}^{t-1}\mathbb{E}\frac{s\left\|z_s\right\|_2^2}{s} + C'\ell_\Xi^2\sup_{T_0\leq t\leq T}\frac{1}{t}\sum_{s=T_0}^{t-1}\eta_{t-1}^2 s^2\mathbb{E}\left\|v_s\right\|_2^2\\ &\leq T_0\mathbb{E}\left\|z_{T_0}\right\|_2^2 + 2\sigma_*^2 + C\frac{\ell_\Xi^2}{\mu^2}\sup_{T_0\leq t\leq T}\left(\frac{1}{t}\sum_{s=T_0}^{t}\frac{1}{s}\right)\cdot\sup_{T_0\leq t\leq T} t\mathbb{E}\left\|z_t\right\|_2^2 + C'\ell_\Xi^2\sup_{T_0\leq t\leq T}\frac{1}{t}\sum_{s=T_0}^{t-1}\eta_{t-1}^2 s^2\mathbb{E}\left\|v_s\right\|_2^2\end{aligned}$$

For $T_0 > 2C\frac{\ell_\Xi^2}{\mu^2}$, we have:

$$C\frac{\ell_\Xi^2}{\mu^2}\sup_{T_0\leq t\leq T}\frac{1}{t}\sum_{s=T_0}^{t}\frac{1}{s} \leq \frac{C\ell_\Xi^2}{\mu^2 T_0} < \frac{1}{2}$$

So we can discard the term involving z_t itself in the right hand side of the above bound at a price of factor 2:

$$H_T \leq 2H_{T_0} + 4\sigma_*^2 + 2C'\ell_\Xi^2\sup_{T_0\leq t\leq T}\frac{1}{t}\sum_{s=T_0}^{t-1}\eta_{s-1}^2 W_s \tag{5.19a}$$

On the other hand, the bound (5.17) implies the bound:

$$W_T \leq (1 - c\eta_{T-1}\mu)W_{T-1} + \frac{C}{T\mu\eta_{T-1}}H_{T-1} + 2\sigma_*^2 \tag{5.19b}$$

for universal constants $c, C > 0$.

The solution to above recursive relations are given by the following lemma:

Lemma 5.8. *For a pair of sequences $(H_t)_{t\geq T_0}$ and $(W_t)_{t\geq T_0}$ satisfying the recursive relation* (5.19a) *with non-increasing stepsize sequence $(\eta_t)_{t\geq T_0}$. Assuming that $(H_t)_{t\geq T_0}$ is non-decreasing, there exists universal constants $c > 0$, such that for $T \geq T_0$, we have the bound:*

$$H_T \leq c\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \eta_{T_0}}{\mu} W_{T_0} + H_{T_0}\right) \qquad \textit{and} \tag{5.20a}$$

$$W_T \leq \frac{c}{\eta_T \mu}\sigma_*^2 + c\left(\frac{T_0}{T\mu^2\eta_{T-1}^2} + e^{-\mu \sum_{t=T_0+1}^{T} \eta_t} {T_0}^2\right) W_{T_0} \tag{5.20b}$$

See Section E.1.3 for the proof of this lemma. Taking this lemma as given, we now proceed with the proof of this proposition.

First, we note that the exponent in the bound (5.20b) satisfies the bound:

$$\mu \sum_{t=T_0+1}^{T} \eta_t \geq c_1 \sum_{t=T_0+1}^{T} \frac{1}{t} \geq c_1 \log \frac{T}{T_0}$$

For $c_1 \geq 2$ and $\eta_T \leq \frac{c'}{\mu T_0}$, we have that $\frac{T_0}{T\mu^2\eta_{T-1}^2} \geq e^{-\mu \sum_{t=T_0+1}^{T} \eta_t} {T_0}^2$. So the bound (5.20b) implies that:

$$\mathbb{E}\left\|v_t\right\|_2^2 \leq \frac{c\sigma_*^2}{\mu\eta_t t^2} + \frac{cT_0}{t^3\eta_t^2\mu^2}\mathbb{E}\left\|v_{T_0}\right\|_2^2$$

For the process z_t, by substituting the bounds in Lemma 5.8 into Eq (5.18), for stepsize $\eta_t < \frac{1}{\mu T_0}$, we obtain:

$$\begin{aligned}
\mathbb{E}\left\|z_t\right\|_2^2 &\leq \frac{{T_0}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2 H_t}{\mu^2 t^2}\left(\sum_{s=T_0}^{t-1}\frac{1}{s}\right) + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}\eta_s^2 W_s \\
&\leq \frac{{T_0}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2 \log t}{\mu^2 t^2}\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0\eta_{T_0}}{\mu}\mathbb{E}\left\|v_{T_0}\right\|_2^2\right) + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=T_0+1}^{t}\eta_s\frac{\sigma_*^2}{\mu} + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}\frac{T_0}{s\mu^2}\mathbb{E}\left\|v_{T_0}\right\|_2^2 \\
&\leq c\frac{\sigma_*^2}{t} + c\frac{{T_0}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + c\frac{\ell_\Xi^2 T_0 \log t}{\mu^2 t^2}\mathbb{E}\left\|v_{T_0}\right\|_2^2
\end{aligned}$$

For the initial conditions at burn-in period, we have:

$$\begin{aligned}
\mathbb{E}\left\|z_{T_0}\right\|_2^2 &= \frac{1}{{T_0}^2}\mathbb{E}\left\|\sum_{t=0}^{T_0}\varepsilon_t(\theta_0)\right\|_2^2 \leq \frac{\sigma_*^2 + \ell_\Xi^2\left\|\theta_0 - \theta^*\right\|_2^2}{T_0} \\
\mathbb{E}\left\|v_{T_0}\right\|_2^2 &\leq 2\left\|\nabla F(\theta_0)\right\|_2^2 + \mathbb{E}\left\|z_{T_0}\right\|_2^2 \leq 2\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{2(\sigma_*^2 + \ell_\Xi^2\left\|\theta_0 - \theta^*\right\|_2^2)}{T_0}
\end{aligned}$$

Note that $\left\|\theta_0 - \theta^*\right\|_2^2 \leq \frac{1}{\mu^2}\left\|\nabla F(\theta_0)\right\|_2^2$ and $T_0 > \frac{\ell_\Xi^2}{\mu^2}$, we have $\frac{\ell_\Xi^2\left\|\theta_0 - \theta^*\right\|_2^2}{T_0} \leq \mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2$. For $T \geq 2T_0 \log T_0$, we also have:

$$\left(\frac{T_0}{T^3\eta_T^2\mu^2} + \frac{\ell_\Xi^2 T_0 \log T}{T^2\mu^2}\right)\frac{\sigma_*^2}{T_0} \leq \frac{3\sigma_*^2}{T} \qquad \text{and} \qquad \frac{{T_0}^2}{T^2}\cdot\frac{\sigma_*^2}{T_0} \leq \frac{\sigma_*^2}{T}$$

Putting them together, we have the bounds:

$$\mathbb{E}\,\|z_T\|_2^2 \le C\left(\frac{\sigma_*^2}{T} + \frac{\ell_\Xi^2 T_0 \log T}{\mu^2 T^2}\,\|\nabla F(\theta_0)\|_2^2\right) \qquad \text{and} \quad \mathbb{E}\,\|v_T\|_2^2 \le C\left(\frac{\sigma_*^2}{\mu\eta_T T^2} + \frac{T_0}{\mu^2 T^3 \eta_T^2}\,\|\nabla F(\theta_0)\|_2^2\right)$$

which complete the proof of this proposition.

5.4.2 *Proof of Theorem 5.1*

By Proposition 5.1, for $t \ge T_0$, taking $\eta_t = \frac{1}{\mu {T_0}^{1-\alpha} t^\alpha}$, there exist constants $a_1, a_2 > 0$ depending on the problem-specific parameters $(\mu, L, \ell_\Xi, \sigma_*, \theta_0, \alpha)$ but independent of t, such that for $t \ge 2T_0 \log T_0$, we have the bounds:

$$\begin{aligned}\mathbb{E}\,\|v_t\|_2^2 &\le a_1\left(\frac{1}{t^2\eta_t} + \frac{1}{t^3\eta_t^2} + \frac{1}{t^2}\right) \le \frac{3a_1}{t^{2-\alpha}}\\ \mathbb{E}\,\|z_t\|_2^2 &\le \frac{a_2}{t} + \frac{a_2 \log t}{t^2} \le \frac{2a_2}{t}\end{aligned}$$

and consequently, we have:

$$\mathbb{E}\,\|\theta_t - \theta^*\|_2^2 \le \frac{1}{\mu}\mathbb{E}\,\|\nabla F(\theta_t)\|_2^2 \le \frac{2}{\mu}(\mathbb{E}\,\|v_{t+1}\|_2^2 + \|z_{t+1}\|_2^2) \le \frac{2}{\mu^2}\left(\frac{3a_1}{t^{2-\alpha}} + \frac{2a_2}{t}\right) \le \frac{a_3}{t}$$

for a constant $a_3 = \frac{6}{\mu^2}(a_1 + a_2) < +\infty$.

For the martingale Ψ_t, we note that:

$$\mathbb{E}\,\|\Psi_t\|_2^2 = \sum_{s=T_0}^{t}(s-1)^2\mathbb{E}\,\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\|_2^2 \le \sum_{s=T_0}^{t}(s-1)^2\ell_\Xi^2\mathbb{E}\,\|\theta_{s-1} - \theta_{s-2}\|_2^2$$

$$\le \sum_{s=T_0}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\,\|v_{s-1}\|_2^2 \le \frac{1}{\mu^2{T_0}^{2-2\alpha}}\sum_{s=0}^{t-1}s^{2-2\alpha}\cdot\frac{3a_1}{s^{2-\alpha}} \le \frac{3a_1}{(1-\alpha)\mu^2{T_0}^{2-2\alpha}}t^{1-\alpha}$$

Define the process $N_t := \sum_{s=1}^{t}\varepsilon_s(\theta^*)$. We note that:

$$\mathbb{E}\,\|M_t - N_t\|_2^2 = \sum_{s=1}^{t}\mathbb{E}\,\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\|_2^2 \le \ell_\Xi^2\sum_{s=1}^{t}\mathbb{E}\,\|\theta_s - \theta^*\|_2^2 \le \ell_\Xi^2 a_3 \log t$$

Putting together the pieces, we obtain:

$$\begin{aligned}t\mathbb{E}\left\|z_t - \frac{1}{t}N_t\right\|_2^2 &\le \frac{3}{t}\,\|z_0\|_2^2 + \frac{3}{t}\mathbb{E}\,\|\Psi_t\|_2^2 + \frac{3}{t}\mathbb{E}\,\|M_t - N_t\|_2^2\\ &\le \frac{3}{t}\,\|z_0\|_2^2 + \frac{3a_1 t^{1-\alpha}}{(1-\alpha)\mu^2{T_0}^{2-2\alpha}t} + \frac{3}{t}\cdot\ell_\Xi^2 C\log t \to 0 \qquad (5.21)\end{aligned}$$

Note that N_t is sum of i.i.d. random vectors. By standard CLT, we have:

$$\frac{N_T}{\sqrt{T}} \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$$

The second moment bound (5.21) implies that:

$$\left\| \sqrt{T} z_T - \frac{N_T}{\sqrt{T}} \right\|_2 \xrightarrow{p} 0$$

Combining these results with Slutsky's theorem, we find that

$$\sqrt{T} z_T \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$$

Note that $\nabla F(\theta_{t-1}) = v_t - z_t$. Since we have the bound $\mathbb{E}\|v_t\|_2^2 \leq \frac{3a_1}{t^{2-\alpha}}$ for $\alpha \in (0,1)$, it is easy to see that $\sqrt{T} v_T \xrightarrow{p} 0$. Consequently, by Slutsky's theorem, we obtain:

$$\sqrt{T} \cdot \nabla F(\theta_T) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$$

Finally, we note that for $\theta \in \mathbb{R}^d$, there is:

$$\begin{aligned}
\|\nabla F(\theta) - H^*(\theta - \theta^*)\|_2 &= \left\| \int_0^1 \nabla^2 F(\theta^* + \gamma(\theta - \theta^*))(\theta - \theta^*) d\gamma - H^*(\theta - \theta^*) \right\|_2 \\
&\leq \int_0^1 \|\!|\nabla^2 F(\theta^* + \gamma(\theta - \theta^*)) - H^*|\!\|_{\text{op}} \cdot \|\theta - \theta^*\|_2 \, d\gamma \\
&\leq \|\theta - \theta^*\|_2 \cdot \sup_{\|\theta' - \theta^*\|_2 \leq \|\theta - \theta^*\|_2} |\!|\!|\nabla^2 F(\theta') - H^*|\!|\!|_{\text{op}}
\end{aligned}$$

Therefore, since $F \in C^2$, we have:

$$\lim_{\theta \to \theta^*} \frac{\|\nabla F(\theta) - H^*(\theta - \theta^*)\|_2}{\|\theta - \theta^*\|_2} = 0$$

By Assumption 9, we have $\|\nabla F(\theta) - \nabla F(\theta^*)\|_2 \geq \mu \|\theta - \theta^*\|_2$, plugging into above bounds, we obtain $\lim_{\theta \to \theta^*} \frac{\|\nabla F(\theta) - H^*(\theta - \theta^*)\|_2}{\|\nabla F(\theta)\|_2} = 0$.

Therefore, since $\sqrt{T} \cdot \nabla F(\theta_T) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$, we have $\sqrt{T} \|\nabla F(\theta_T) - H^*(\theta_T - \theta^*)\|_2 \xrightarrow{p} 0$. This leads to $\sqrt{T} H^*(\theta_T - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$, and consequently,

$$\sqrt{T}(\theta_T - \theta^*) \xrightarrow{d} \mathcal{N}(0, (H^*)^{-1} \Sigma^* (H^*)^{-1})$$

which finishes the proof.

5.4.3 Proof of Theorem 5.2

The proof is by explicit construction of a function (and associated noise) satisfying the Assumptions 9, 10, 11 and 12, for which the Polyak-Ruppert procedure fails.

Consider the following function:

$$F(x) := \begin{cases} x^2 - \frac{1}{2}\int_0^x \frac{z\,dz}{\log(e+|z|^{-1})} & x \geq 0 \\ x^2 - \frac{1}{4}\int_0^x \frac{z\,dz}{\log(e+|z|^{-1})} & x < 0 \end{cases}$$

Some algebra yields:

$$F'(x) = \begin{cases} 2x - \frac{x}{2\log(e+|x|^{-1})} & x \geq 0 \\ 2x - \frac{x}{4\log(e+|x|^{-1})} & x < 0 \end{cases}$$

and

$$F''(x) = \begin{cases} 2 - \frac{1}{2\log(e+|x|^{-1})} - \frac{1}{2\log^2(e+|x|^{-1})\cdot(e|x|+1)} & x \geq 0 \\ 2 - \frac{1}{4\log(e+|x|^{-1})} - \frac{1}{4\log^2(e+|x|^{-1})\cdot(e|x|+1)} & x < 0 \end{cases}$$

Clearly, F is twice continuously differentiable everywhere on $\mathbb{R}$, satisfying the bound for any $x \in \mathbb{R}$:

$$1 \leq F''(x) \leq 2$$

It is easy to see that F has an unique minimizer 0, with $H^* = F''(0) = 2$.

We consider an additive Gaussian noise model

$$f(\theta, \xi_t) := F(\theta) - \sqrt{2}\langle \xi_t, \theta \rangle \qquad \text{where } \xi_t \sim \mathcal{N}(0,1)$$

Clearly, the noise model satisfies Assumption 10 and 11 with constants $\sigma_* = \sqrt{2}$ and $\ell_\Xi = 0$.

Now we consider the SGD update rule on function F:

$$\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t) + \sqrt{2}\eta\xi_{t+1}$$

Given $\eta = \eta_0 T^{-\alpha}$, we consider the following re-scaled function:

$$\forall x > 0 \qquad F_\eta(x) := \eta^{-1} F\left(\sqrt{\eta} x\right) \tag{5.22}$$

Clearly, F_η is a strongly-convex and smooth function, with $1 \leq F''_\eta(x) \leq 2$. Denote $\psi_t := \theta_t / \sqrt{\eta}$ and $\overline{\psi}_T := \overline{\theta}_T / \sqrt{\eta}$. The SGD iterates can be re-written as

$$\psi_{t+1} = \psi_t - \eta \nabla F_\eta(\psi_t) + \sqrt{2\eta}\xi_{t+1}$$

We also define the re-scaled function $\delta_\eta(x) := \frac{1}{\sqrt{\eta}}\delta(x\sqrt{\eta})$. Clearly we have the relation $\delta_\eta(x) = 2x - \nabla F_\eta(x)$. We denote $\pi_\eta^{(t)} := \mathscr{L}(\psi_t)$, the probability law of the iterate ψ_t.

This is an instance of *unadjusted Langevin algorithm* (ULA) on the function F_η, which is known to converge to an approximation to the target density $\pi_\eta \propto e^{-F_\eta}$. More precisely, the following non-asymptotic error bounds are known from the chapter [59] (for notational simplicity, we suppress the dependency on the strong convexity and smoothness parameter, as well as the problem dimension, as they are all universal constants in above problem):

Proposition 5.3 (Special case of [59], Theorem 5). *Under above setup, we have the following bound for $k = 1, 2\cdots$*

$$\mathscr{W}_2^2(\pi_\eta^{(k)}, \pi_\eta) \leq 2e^{-c_1\eta k}\big(\|\psi_0\|_2^2 + 1\big) + c_2\eta \tag{5.23a}$$

for constants $c_1, c_2 > 0$ independent of η, k and ψ_0.

The mean-square error bounds for estimation expectation of a Lipschitz functional is also given by [59].

Proposition 5.4 (Special case of [59], Eq (27) and Theorem 15). *Under above setup, given any Lipschitz test function h, let $\overline{h}_{T_0,T} := \frac{1}{T-T_0}\sum_{t=T_0}^{T-1} h(\psi_t)$, the following bounds hold true:*

$$\left(\mathbb{E}\left[\overline{h}_{T_0,T}\right] - \mathbb{E}_{\pi_\eta}\left[h(X)\right]\right)^2 \leq \frac{\|h\|_{\mathrm{Lip}}^2}{T-T_0}\sum_{t=T_0}^{T}\mathscr{W}_2^2(\pi_\eta^{(t)}, \pi_\eta) \tag{5.23b}$$

$$\mathrm{var}\left(\overline{h}_{T_0,T}\right) \leq c\frac{\|h\|_{\mathrm{Lip}}^2}{\eta(T-T_0)} \tag{5.23c}$$

for a universal constant $c > 0$.

Note that $\psi_0 = 0$. So we have the following bound on the sum of squares of Wasserstein distance

$$\sum_{k=T_0}^{T}\mathscr{W}_2^2(\pi_\eta^{(k)}, \pi_\eta) \leq 2\sum_{k=T_0}^{T} e^{-c_1\eta k} + c_2(T-T_0)\eta \leq \frac{2}{c_1\eta} + c_2(T-T_0)\eta$$

Substituting into the MSE bound in Proposition 5.4, for any choice of burn-in parameter $\beta \in [0,1)$, we have the bound:

$$\mathbb{E}\left(\overline{\psi}_T - \mathbb{E}_{\pi_\eta}[X]\right)^2 \leq c\left(\eta + \frac{1}{\eta(T-T_0)}\right) \leq c'T^{-\min(\alpha, 1-\alpha)} \tag{5.24}$$

where the constants $c, c' > 0$ can depend on $\|\theta_0\|_2$ and η_0, but are independent of T.

It remains to study the stationary distribution π_η. The following lemma characterizes the size of bias under the stationary distribution π_η.

Lemma 5.9. *For the* 1*-dimensional probability distribution* π_η *defined above, we have that*

$$\mathbb{E}_{\pi_\eta}[X] \geq c \cdot \left(\log \frac{1}{\eta}\right)^{-1}$$

for a universal constant $c > 0$.

Combining the bound (5.24) and Lemma 5.9, we arrive at the lower bound:

$$\mathbb{E}\left[\overline{\psi}_T^2\right] \geq \frac{c_1}{\log^2 T} - \frac{c_2}{T^{\min(\alpha, 1-\alpha)}}$$

for constants $c_1, c_2 > 0$ that are independent of T.

Recovering the original scaling, we obtain the lower bound for the Polyak-Ruppert estimator:

$$\mathbb{E}\left\|\overline{\theta}_T - \theta^*\right\|_2^2 \geq \frac{c_1'}{T^\alpha \log^2 T} - \frac{c_2'}{T^{\min(2\alpha, 1)}}$$

Taking the limit, we have:

$$\lim_{T \to +\infty} T \cdot \mathbb{E}\left\|\overline{\theta}_T - \theta^*\right\|_2^2 = +\infty$$

which completes the proof of this theorem.

Proof (Proof of Lemma 5.9). Denote the normalization constant:

$$Z_\eta := \int e^{-F_\eta(x)} dx$$

Since $x^2 \leq F(x) \leq 2x^2$ for any $x \in \mathbb{R}$, we have the bound $\sqrt{\pi/2} \leq Z_\eta \leq \sqrt{\pi}$ for any choice of $\eta > 0$. By definition, we have the expression:

$$\mathbb{E}_{\pi_\eta}[X] = Z_\eta^{-1} \int_0^{+\infty} x \left(e^{-F_\eta(x)} - e^{-F_\eta(-x)}\right) dx$$

Note that $F(x) \leq F(-x)$ for any $x \geq 0$. So we have that $\mathbb{E}_{\pi_\eta}[X] \geq 0$, and the following bound holds:

$$\mathbb{E}_{\pi_\eta}[X] \geq \frac{1}{\sqrt{\pi}} \int_1^2 \left(e^{-F_\eta(x)} - e^{-F_\eta(-x)}\right) dx$$

Given $x \in [1, 2]$ fixed, we lower bound the difference in the density function as follows:

$$e^{-F_\eta(x)} - e^{-F_\eta(-x)} = e^{-x^2}\left(e^{1/2 \int_0^x \delta_\eta(z) dz} - e^{1/4 \int_0^x \delta_\eta(z) dz}\right) \geq \frac{e^{-4}}{4} \int_0^x \delta_\eta(z) dz$$

$$\geq \frac{e^{-4}}{4} \int_{1/2}^{1} \frac{z}{\log\left(e + (z\sqrt{\eta})^{-1}\right)} dz \geq \frac{e^{-4}}{8} \cdot \frac{1}{\log\left(e + \frac{2}{\sqrt{\eta}}\right)}$$

Integrating with $x \in [1,2]$, we arrive at the lower bound:

$$\mathbb{E}_{\pi_\eta}[X] \geq c \cdot \left(\log \frac{1}{\eta}\right)^{-1}$$

for universal constant $c > 0$.

5.5 Proof of the non-asymptotic bounds with sharp pre-factor

In this section, we present the proofs for Theorem 5.3, Theorem 5.4 and Theorem 5.5. These three results provide upper bounds on three different metrics (gradient norm, iterate distance, and function value), with the leading-order term exactly matching the optimal normal limit, and sharp high-order terms. We present the proof of Proposition 5.2 en route (in §5.5.2), the higher-moment non-asymptotic convergence rates for the process $(v_t)_{t \geq T_0}$ and $(z_t)_{t \geq T_0}$ that is analogous to Proposition 5.1.

5.5.1 Proof of Theorem 5.3

We first establish the results for the single-loop algorithm, and then use it to prove the results with the re-starting loops.

Throughout the proof, we use the following notations for the risk functions

$$r_v(t) := \left(\mathbb{E}\|v_t\|_2^2\right)^{1/2} \qquad \text{and} \quad r_\theta(t) := \frac{1}{\mu}\left(\mathbb{E}\|\nabla F(\theta_T)\|_2^2\right)^{1/2}$$

Clearly, by the strong convexity Assumption 9, we have the bound $\mathbb{E}\|\theta_T - \theta^*\|_2^2 \leq r_\theta(t)^2$.

We start by observing the following decomposition:

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 = \mathbb{E}\|z_{T+1}\|_2^2 + \mathbb{E}\|v_{T+1}\|_2^2 - 2\mathbb{E}\langle z_{T+1}, v_{T+1}\rangle \tag{5.25}$$

The following lemma provides sharp bounds on the leading-order term $\mathbb{E}\|z_{T+1}\|_2^2$.

Lemma 5.10. *Under above set-up, for $T \geq 2T_0 \log T_0$ and any $G \in \mathbb{R}^{d\times d}$, the following bounds hold true for the process $(z_t)_{t \geq T_0}$:*

$$\mathbb{E}\|Gz_T\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left(G\Sigma^* G^\top\right) + c|||G|||_{op}^2 \mathscr{H}_T^{(z)} \tag{5.26a}$$

where the high order term $\mathscr{H}_T^{(z)}$ is defined as

$$\mathscr{H}_T^{(z)} := c\left(\sqrt{\frac{T_0}{T}} + \frac{{T_0}^{\alpha}}{T^{\alpha}}\right)\frac{\sigma_*^2}{T} + c\frac{{T_0}^2 \log T}{T^2}\left(1 + \frac{T^{2\alpha-3/2}}{{T_0}^{2\alpha-3/2}}\right)\|\nabla F(\theta_0)\|_2^2 \quad (5.26b)$$

See §E.1.4 for the proof of this lemma.

Invoking Proposition 5.1, we have the bound for v_T:

$$\mathbb{E}\|v_T\|_2^2 \leq c\left(\frac{\sigma_*^2}{\mu\eta_T T^2} + \frac{T_0}{\mu^2 T^3 \eta_T^2}\|\nabla F(\theta_0)\|_2^2\right)$$

For the stepsize choice $\eta_t = \frac{1}{\mu {T_0}^{1-\alpha} t^{\alpha}}$, we have the bound

$$\mathbb{E}\|v_T\|_2^2 \leq c\frac{{T_0}^{1-\alpha}}{T^{1-\alpha}} \cdot \frac{\sigma_*^2}{T} + c\frac{{T_0}^{3-2\alpha}}{T^{3-2\alpha}} \cdot \|\nabla F(\theta_0)\|_2^2 \tag{5.27}$$

Combining the bounds (5.26a) and (5.27) and substituting into the decomposition (5.25), we arrive at the following bound by applying Young's inequality:

$$\begin{aligned}
\mathbb{E}\|\nabla F(\theta_T)\|_2^2 &\leq \mathbb{E}\|z_{T+1}\|_2^2 + \mathbb{E}\|v_{T+1}\|_2^2 + 2\sqrt{\mathbb{E}\|z_{T+1}\|_2^2} \cdot \sqrt{\mathbb{E}\|v_{T+1}\|_2^2} \\
&\leq \left(1 + \left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}}\right) \cdot \mathbb{E}\|z_{T+1}\|_2^2 + \left(1 + \left(\frac{T}{T_0}\right)^{\frac{1-\alpha}{2}}\right)\mathbb{E}\|v_{T+1}\|_2^2 \\
&\leq \frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T \cdot \|\nabla F(\theta_0)\|_2^2
\end{aligned}$$

which proves the first claim (5.11a).

Now we turn to the proof of multi-loop results. By applying the one-loop result to each short epoch, we have the bound for $b = 1, 2, \cdots, B$:

$$\begin{aligned}
\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^2 &\leq \frac{\sigma_*^2}{T^\flat} + c\left(\frac{T_0}{T^\flat}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T^\flat} + c\left(\frac{T_0}{T^\flat}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T \cdot \mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2 \\
&\overset{(i)}{\leq} \frac{2\sigma_*^2}{T^\flat} + \frac{1}{2}\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2
\end{aligned}$$

In step (i), we use the fact that $T^\flat \geq 2cT_0 \log T_0$ and that $2 \wedge \frac{5-3\alpha}{2} > 1$ for $\alpha \in (0,1)$.

Solving the recursion, we arrive at the bound:

$$\mathbb{E}\left\|\nabla F\left(\theta_0^{(B+1)}\right)\right\|_2^2 \leq \frac{4\sigma_*^2}{T_0} + 2^{-B}\|\nabla F(\theta_0)\|_2^2$$

Substituting this initial condition into the bound (5.11a), we obtain the final bound:

$$\mathbb{E}\left\|\nabla F\big(\theta_T^{(B+1)}\big)\right\|_2^2 \leq \frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T\cdot\left(\frac{4\sigma_*^2}{T_0} + 2^{-B}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$

Taking $B \geq \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2}\vee 1\right)$ and substituting with $T = n - BT^\flat$, we arrive at the conclusion:

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_n)\right\|_2^2 \leq \left(1 + c\left(\frac{T_0}{n}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\log^2 n\right)\frac{\sigma_*^2}{n}$$

which proves the bound (5.11b).

5.5.2 Proof of Proposition 5.2

Throughout the proof, we frequently use the following inequalities for the moments of stochastic gradients, which holds true for any $\theta \in \mathbb{R}^d$:

$$\mathbb{E}\left\|\nabla f(\theta,\xi_t)\right\|_2^4 \leq 27\widetilde{\sigma_*}^4 + 27\left(1+\frac{\ell_\Xi^4}{\mu^4}\right)\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 \tag{5.28}$$

To see why this is true, we note that:

$$\begin{aligned}\mathbb{E}\left\|\nabla f(\theta,\xi_t)\right\|_2^4 &\leq 27\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 + 27\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^4 + 27\mathbb{E}\left\|\varepsilon_t(\theta)-\varepsilon_t(\theta^*)\right\|_2^4\\ &\leq 27\widetilde{\sigma_*}^4 + 27\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 + 27\ell_\Xi^4\mathbb{E}\left\|\theta-\theta^*\right\|_2^4\\ &\leq 27\widetilde{\sigma_*}^4 + 27\left(1+\frac{\ell_\Xi^4}{\mu^4}\right)\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4\end{aligned}$$

Now we turn to the proof of this proposition. Similar to the proof of Proposition 5.1, we need the following technical lemmas:

Lemma 5.11. *Under Assumption 9, 10′, 11, there exists universal constants $c, c' > 0$, when $\eta_t \leq c\big(\frac{1}{L}\wedge\frac{\mu}{\ell_\Xi^2}\big)$, we have the bound*

$$\sqrt{\mathbb{E}\left\|v_t\right\|_2^4} \leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{c'}{t^2}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}\right)$$

Lemma 5.12. *Under Assumption 10′, we have the bound*

$$\sqrt{\mathbb{E}\left\|z_t\right\|_2^4} \leq \frac{cT_0{}^2\left\|z_0\right\|_2^2}{t^2} + \frac{c\widetilde{\sigma_*}^2}{t} + \frac{c\ell_\Xi^2}{\mu^2t^2}\sum_{s=T_0}^{t-1}\sqrt{\mathbb{E}\left\|\nabla F(\theta_s)\right\|_2^4} + \frac{c\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1}s^2\eta_s^2\sqrt{\mathbb{E}\left\|v_s\right\|_2^4}$$

See Section E.1.5 and E.1.6 for the proofs of the two lemmas. Taking these two lemmas as given, we now proceed with the proof of the proposition.

The rest of proof goes in parallel with the proof of Proposition 5.1. We first note that:

$$\sqrt{\mathbb{E}\|\nabla F(\theta_t)\|_2^4} \le 4\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 4\sqrt{\mathbb{E}\|\nabla F(\theta_t) - \nabla F(\theta_{t-1})\|_2^4}$$
$$\le 4\sqrt{\mathbb{E}\|z_t\|_2^2} + 4\sqrt{\mathbb{E}\|v_t\|_2^2} + 4(\eta_t L)^4\sqrt{\mathbb{E}\|v_t\|_2^4} \le 4\sqrt{\mathbb{E}\|z_t\|_2^4} + 6\sqrt{\mathbb{E}\|v_t\|_2^4}$$

Substituting into the bounds in Lemma 5.11 and 5.12, and defining the quantities $H_T := \sup_{T_0 \le t \le T} t\sqrt{\mathbb{E}\|z_t\|_2^4}$, $W_T := T^2\sqrt{\mathbb{E}\|v_T\|_2^4}$, we arrive at the following recursive inequalities:

$$H_T \le 2H_{T_0} + 4\widetilde{\sigma_*}^2 + 2C'\ell_\Xi^2 \sup_{T_0 \le t \le T} \frac{1}{t}\sum_{s=T_0}^{t-1} \eta_{s-1}^2 W_s \tag{5.29a}$$

$$W_T \le (1 - c\eta_{T-1}\mu)W_{T-1} + \frac{C}{T\mu\eta_{T-1}} H_{T-1} + 2\widetilde{\sigma_*}^2 \tag{5.29b}$$

Invoking Lemma 5.8 by replacing (ℓ_Ξ, σ_*) with $(\ell_\Xi, \widetilde{\sigma_*})$, we obtain the following bounds:

$$H_T \le c\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \eta_{T_0}}{\mu} W_{T_0} + H_{T_0}\right) \qquad \text{and}$$
$$W_T \le \frac{c}{\eta_T \mu}\sigma_*^2 + c\left(\frac{T_0}{T\mu^2\eta_{T-1}^2} + e^{-\mu\sum_{t=T_0+1}^T \eta_t} {T_0}^2\right) W_{T_0}$$

For the initial conditions, by applying Khintchine's inequality as well as Young's inequality, we note that:

$$\mathbb{E}\left\|z_{T_0}\right\|_2^4 = \frac{1}{{T_0}^4}\mathbb{E}\left\|\sum_{t=1}^{T_0} \varepsilon_t(\theta_0)\right\|_2^4 \le \frac{1}{{T_0}^4}\mathbb{E}\left(\sum_{t=1}^{T_0}\|\varepsilon_t(\theta_0)\|_2^2\right)^2 \le 8\frac{1}{{T_0}^2}\left(\widetilde{\sigma_*}^4 + \ell_\Xi^4\|\theta_0 - \theta^*\|_2^4\right)$$
$$\mathbb{E}\left\|v_{T_0}\right\|_2^4 \le 8\mathbb{E}\|\nabla F(\theta_0)\|_2^4 + \mathbb{E}\left\|z_{T_0}\right\|_2^4 \le 8\|\nabla F(\theta_0)\|_2^4 + 8\frac{1}{{T_0}^2}\left(\widetilde{\sigma_*}^4 + \ell_\Xi^4\|\theta_0 - \theta^*\|_2^4\right)$$

Following exactly the same arguments as in the proof of Proposition 5.1, we arrive at the desired bounds.

5.5.3 *Proof of Theorem 5.4*

We define the quantities $r_\theta(t)$ and $r_v(t)$ the same as in the proof of Theorem 5.3. Furthermore, we denote the following quantities:

$$\widetilde{r}_v(t) := \left(\mathbb{E}\|v_t\|_2^4\right)^{1/4} \qquad \text{and} \qquad \widetilde{r}_\theta(t) := \frac{1}{\mu}\left(\mathbb{E}\|\nabla F(\theta_t)\|_2^4\right)^{1/4}$$

Clearly, by the strong convexity Assumption 9, we have the bound $\mathbb{E}\|\theta_T - \theta^*\|_2^4 \leq r_\theta(t)^4$.

We also note the following decomposition of the gradient:

$$\nabla F(\theta_T) = \int_0^1 \nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big)(\theta_T - \theta^*)d\gamma$$

which leads to the following bound under Assumption 12$'$:

$$\begin{aligned}\big\|(H^*)^{-1}\nabla F(\theta_T) - (\theta_T - \theta^*)\big\|_2 &\leq \int_0^1 \big\|(H^*)^{-1}\left(\nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\big\|_2 d\gamma \\ &\leq \frac{L_2}{\lambda_{\min}(H^*)}\|\theta_T - \theta^*\|_2^2 \leq \frac{L_2}{\lambda_{\min}(H^*)\mu^2}\|\nabla F(\theta_T)\|_2^2 \end{aligned} \tag{5.30}$$

We can then upper bound the mean-squared error using the processes $(z_t)_{t\geq T_0}$ and $(v_t)_{t\geq T_0}$:

$$\begin{aligned}\mathbb{E}\|\theta_T - \theta^*\|_2^2 &\leq \mathbb{E}\left(\big\|(H^*)^{-1}\nabla F(\theta_T)\big\|_2 + \frac{L_2}{\mu^2\lambda_{\min}(H^*)}\|\nabla F(\theta_T)\|_2^2\right)^2 \\ &\leq \mathbb{E}\big\|(H^*)^{-1}\big(v_{T+1} - z_{T+1}\big)\big\|_2^2 + 2\frac{L_2}{\lambda_{\min}(H^*)}\widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)^2}\widetilde{r}_\theta^4(T)\end{aligned} \tag{5.31}$$

The leading-order term in the bound (5.31) admits the following decomposition:

$$\mathbb{E}\big\|(H^*)^{-1}\big(z_{T+1} - v_{T+1}\big)\big\|_2^2 = \mathbb{E}\big\|(H^*)^{-1}z_{T+1}\big\|_2^2 + \mathbb{E}\big\|(H^*)^{-1}v_{T+1}\big\|_2^2 - 2\mathbb{E}\left[\langle (H^*)^{-1}z_T,\, (H^*)^{-1}.v_T\rangle\right]$$

In the following, we bound the three terms in above equation, respectively. Invoking Lemma 5.10 with $G = (H^*)^{-1}$, we have the bound:

$$\mathbb{E}\big\|(H^*)^{-1}z_{T+1}\big\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{\alpha\wedge\frac{1}{2}} + \frac{c\|\nabla F(\theta_0)\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \tag{5.32a}$$

For the process v_t, Proposition 5.1 yields the following upper bound:

$$\mathbb{E}\big\|(H^*)^{-1}v_{T+1}\big\|_2^2 \leq \frac{1}{\lambda_{\min}(H^*)^2}\mathbb{E}\|v_{T+1}\|_2^2 \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{3-2\alpha} \tag{5.32b}$$

The bound for the cross term is given by the following lemma:

Lemma 5.13. *Under above set-up, for $T \geq cT_0\log T_0$, for any $d\times d$ deterministic matrix G, the following bound holds true:*

$$\left|\mathbb{E}\left[\langle Gz_t, Gv_t\rangle\right]\right| \leq c|||G|||_{op}^2 \left(\frac{T_0}{t}\right)^{1-\alpha} \left(\frac{\sigma_*^2}{t} + \left(\frac{T_0}{t}\right)^{2-\alpha} \|\nabla F(\theta_0)\|_2^2\right) \log t$$
$$+ c\frac{|||G|||_{op}^2 L_2}{\mu^2} \left(\frac{T_0}{t}\right)^{\frac{1-\alpha}{2}} \left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2} \log^2 t \, \|\nabla F(\theta_0)\|_2^3\right)$$

See §E.1.7 for the proof of this lemma.

Substituting with $G = (H^*)^{-1}$, we obtain the bound for the cross term:

$$\left|\mathbb{E}\left[\langle (H^*)^{-1}z_t, (H^*)^{-1}v_t\rangle\right]\right| \leq \frac{c\sigma_*^2 \log T}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2 \log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{3-2\alpha}$$
$$+ \frac{cL_2\widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)^2\mu^2 T^{3/2}}\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\|\nabla F(\theta_0)\|_2^3 \log^2 T}{\lambda_{\min}(H^*)^2\mu^2}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha} \tag{5.32c}$$

For the rest two terms in the expression (5.31), we invoke Proposition 5.2, and obtain the rate:

$$\widetilde{r}_\theta(T) \leq \frac{c\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{c\sqrt{\log T}}{\mu}\|\nabla F(\theta_0)\|_2 \left(\frac{T_0}{T}\right)^{1\wedge 3/2-\alpha} \tag{5.32d}$$

Combining the bounds (5.32a)-(5.32d) and substituting into the decomposition (5.31), we arrive at the bound

$$\mathbb{E}\|\theta_T - \theta^*\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2 \log T}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{\alpha\wedge 1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2 \log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha}$$
$$+ \frac{cL_2\widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)\mu^3 T^{3/2}} + \frac{cL_2\mathbb{E}\|\nabla F(\theta_0)\|_2^3 \log^2 T}{\lambda_{\min}(H^*)\mu^3}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha\wedge 3}$$
$$+ \frac{cL_2^2\widetilde{\sigma_*}^4}{\lambda_{\min}(H^*)^2\mu^4 T^2} + \frac{cL_2^2\mathbb{E}\|\nabla F(\theta_0)\|_2^4 \log^2 T}{\lambda_{\min}(H^*)^2\mu^4}\left(\frac{T_0}{T}\right)^{6-4\alpha\wedge 4}$$

Noting that $\frac{7}{2} - 2\alpha \wedge 3 \geq \frac{3}{2}$ and $6 - 4\alpha \wedge 4 \geq 4$, we complete the proof of the bound (5.15a).

Now we turn to the proof of the multi-loop result (5.15b). Invoking Proposition 5.1 and 5.2 and noting that $\|\nabla F(\theta_t)\|_2 \leq \|z_{t+1}\|_2 + \|v_{t+1}\|_2$, we obtain the bound for $T^\flat \geq cT_0 \log T_0$:

$$\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^2 \leq \frac{1}{2}\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat} \quad \text{and}$$
$$\sqrt{\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^4} \leq \frac{1}{2}\sqrt{\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat}$$

Solving the recursion, we have that:

$$\mathbb{E}\left\|\nabla F\big(\theta_0^{(B+1)}\big)\right\|_2^2 \leq 2^{-B}\mathbb{E}\left\|\nabla F\big(\theta_0\big)\right\|_2^2 + \frac{2c\sigma_*^2}{T^\flat} \qquad \text{and}$$

$$\sqrt{\mathbb{E}\left\|\nabla F\big(\theta_0^{(B+1)}\big)\right\|_2^4} \leq 2^{-B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{2c\widetilde{\sigma_*}^2}{T^\flat}$$

Taking $B \geq \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2} \vee 1\right)$ and substituting into the bound (5.15a), we have the following guarantee for the multi-loop estimator:

$$\mathbb{E}\left\|\widehat{\theta}_n - \theta^*\right\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{n} + \frac{c\sigma_*^2\log^2 n}{\lambda_{\min}(H^*)^2 n}\left(\frac{T_0}{n}\right)^{\alpha\wedge 1-\alpha}$$
$$+ \frac{cL_2\widetilde{\sigma_*}^3\log^2 n}{\lambda_{\min}(H^*)\mu^3 n^{3/2}} + \frac{cL_2^2\widetilde{\sigma_*}^4\log^2 n}{\lambda_{\min}(H^*)^2\mu^4 n^2}$$

which completes the proof.

5.5.4 Proof of Theorem 5.5

Applying second-order Taylor expansion with integral remainder, for any $\theta \in \mathbb{R}^d$, we note the following identity.

$$F(\theta) = F(\theta^*) + \langle\theta - \theta^*, \nabla F(\theta^*)\rangle + (\theta - \theta^*)^\top \int_0^1 \nabla^2 F\big(\gamma\theta + (1-\gamma)\theta^*\big)d\gamma\cdot(\theta - \theta^*)$$

Noting that $\nabla F(\theta^*) = 0$ and invoking Assumption 12′, we have that:

$$F(\theta) \leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + \|\theta - \theta^*\|_2 \cdot \int_0^1 |||\nabla^2 F\big(\gamma\theta + (1-\gamma)\theta^*\big) - H^*|||_{\mathrm{op}} d\gamma \cdot \|\theta - \theta^*\|_2$$
$$\leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + L_2\|\theta - \theta^*\|_2^3 \tag{5.33}$$

Similar to Eq (5.30), we have the bound:

$$\left\|(H^*)^{1/2}(\theta - \theta^*) - (H^*)^{-1/2}\nabla F(\theta)\right\|_2 \leq \int_0^1 \left\|(H^*)^{-1/2}\left(\nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\right\|_2 d\gamma$$
$$\leq \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}}\|\theta_T - \theta^*\|_2^2 \leq \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}\mu^2}\|\nabla F(\theta_T)\|_2^2$$

Denote the residual $q_t := (H^*)^{1/2}(\theta_t - \theta^*) - (H^*)^{-1/2}\nabla F(\theta_t)$. Substituting into the bound (5.33), we have that:

$$\mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) \leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}\nabla F(\theta) + q_T\right\|_2^2 + L_2\mathbb{E}\|\theta_T - \theta^*\|_2^3$$

$$
\begin{aligned}
&\leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}(z_{T+1}+v_{T+1})\right\|_2^2 + 2L_2\widetilde{r}_\theta^3(T) + \mathbb{E}\left\|q_T\right\|_2^2 \\
&\leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}(z_{T+1}+v_{T+1})\right\|_2^2 + 2L_2\widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)}\widetilde{r}_\theta^4(T)
\end{aligned}
$$

Invoking Proposition 5.1, Lemma 5.10 and E.1 with $G = (H^*)^{-1/2}$, we have the bounds

$$
\begin{aligned}
\mathbb{E}\left\|(H^*)^{-1/2}z_{T+1}\right\|_2^2 &\leq \frac{\mathrm{Tr}\left(\Sigma^*(H^*)^{-1}\right)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{\alpha\wedge\frac{1}{2}} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \\
\mathbb{E}\left\|(H^*)^{-1/2}v_{T+1}\right\|_2^2 &\leq \frac{\mathbb{E}\left\|v_{T+1}\right\|_2^2}{\lambda_{\min}(H^*)} \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{3-2\alpha}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\left\langle(H^*)^{-1/2}z_t,\,(H^*)^{-1/2}v_t\right\rangle\right] &\leq \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{3-2\alpha} \\
&\quad + \frac{cL_2\widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)\mu^2T^{3/2}}\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\left\|\nabla F(\theta_0)\right\|_2^3\log^2 T}{\lambda_{\min}(H^*)\mu^2}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha}
\end{aligned}
\tag{5.34}
$$

Putting them together, we arrive at the bound

$$
\begin{aligned}
\mathbb{E}\left[F(\theta_T)-F(\theta^*)\right] \leq \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right)}{2T} &+ \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{\alpha\wedge1-\alpha} \\
&+ \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} + cL_2\widetilde{r}_T^3 + c\frac{L_2^2}{\mu}\widetilde{r}_T^4
\end{aligned}
$$

for the quantity $\widetilde{r}_T := \frac{\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{T_0}{T}}\cdot\left(\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\right)^{1/4}$.

For the multi-loop algorithm, applying the same argument on the initial gradient norm as in the proof of Theorem 5.4, we arrive at the desired bound.

5.6 Discussion

In this chapter, we revisited the problem of stochastic optimization for strongly convex and smooth M-estimators, focusing on the ROOT-SGD algorithm with diminishing stepsize. We established both sharp asymptotic and non-asymptotic results, demonstrating that ROOT-SGD converges asymptotically to the optimal normal limit under minimal smoothness conditions that guarantee asymptotic normality. In contrast, we provided a counter-example showing that the Polyak-Ruppert averaging procedure is asymptotically sub-optimal under the same conditions.

On the non-asymptotic side, we derived upper bounds on the gradient norm, estimation error, and excess risk, where the leading term matches the asymptotic risk with near-unity pre-factor, and high-order terms decay exponentially. Additionally, with a one-point Hessian-Lipschitz condition, we established that the additional terms decay at a rate of $O(n^{-3/2})$, achieving optimality without requiring prior knowledge of the sample size.

Our findings extend to broader optimization scenarios, suggesting potential applications in non-strongly convex, non-convex, and stochastic approximation problems. Future research could explore these methods in Markovian and distributed data settings, opening new avenues for further development.

Bibliography

References

1. A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58):3235–3249, 2012.
2. M. Ai, F. Wang, J. Yu, and H. Zhang. Optimal subsampling for large-scale quantile regression. *Journal of Complexity*, 62:101512, 2021.
3. P. Alagidede and T. Panagiotidis. Stock returns and inflation: Evidence from quantile regressions. *Economics Letters*, 117(1):283–286, 2012.
4. R. Alhamzawi. Bayesian analysis of composite quantile regression. *Statistics in Biosciences*, 8(2):358–373, 2016.
5. D. E. Allen, P. Gerrans, R. Powell, and A. K. Singh. Quantile regression: its application in investment analysis. *Jassa*, (4):7–12, 2009.
6. Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):8194–8244, 2017.
7. Z. Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems*, pages 1157–1167, 2018.
8. X. Amatriain. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2):37–48, 2013.
9. J. Angrist, V. Chernozhukov, and I. Fernández-Val. Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563, 2006.
10. Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, A. Sekhari, and K. Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.
11. S. Arnold, P.-A. Manzagol, R. B. Harikandeh, I. Mitliagkas, and N. Le Roux. Reducing the variance in online optimization by transporting past gradients. *Advances in Neural Information Processing Systems*, 32:5391–5402, 2019.
12. H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.
13. R. Babanezhad, M. O. Ahmed, A. Virani, M. Schmidt, J. Konečnỳ, and S. Sallinen. Stop wasting my gradients: Practical SVRG. In *Advances in Neural Information Processing Systems*, volume 28, pages 2251–2259, 2015.
14. F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

15. F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.
16. F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 26:773–781, 2013.
17. I. U. Badshah. Quantile regression analysis of the asymmetric return-volatility relation. *Journal of Futures Markets*, 33(3):235–265, 2013.
18. R. R. Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3):577–580, 1966.
19. R. R. Bahadur and R. R. Rao. On deviations of the sample mean. *The Annals of Mathematical Statistics*, 31(4):1015–1027, 1960.
20. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
21. A. Beknazaryan, H. Sang, and Y. Xiao. Cramér type moderate deviations for random fields. *Journal of Applied Probability*, 56(1):223–245, 2019.
22. A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.
23. B. Bercu and A. Rouault. Sharp large deviations for the ornstein–uhlenbeck process. *Theory of Probability & Its Applications*, 46(1):1–19, 2002.
24. D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1989.
25. Bharti and A. Kumar. Exploring herding behaviour in indian equity market during covid-19 pandemic: impact of volatility and government response. *Millennial Asia*, 13(3):513–531, 2022.
26. V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
27. L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.
28. L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
29. L. Bottou and Y. Le Cun. Large scale online learning. In *Advances in Neural Information Processing Systems*, volume 16, pages 217–224. MIT Press, 2004.
30. S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
31. M. Buchinsky. The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of applied econometrics*, 13(1):1–30, 1998.
32. N. Cesa-Bianchi and F. Orabona. Online learning algorithms. *Annual review of statistics and its application*, 8(1):165–190, 2021.
33. S.-M. Chang. *A stationary stochastic approximation algorithm for estimation in the glmm*. North Carolina State University, 2008.
34. H. Chen, W. Lu, and R. Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.
35. L. Chen and Y. Zhou. Quantile regression in big data: A divide and conquer based strategy. *Computational Statistics & Data Analysis*, 144:106892, 2020.
36. X. Chen, Z. Lai, H. Li, and Y. Zhang. Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods. *Journal of the American Statistical Association*, pages 1–24, 2024.
37. X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.
38. X. Chen, W. Liu, X. Mao, and Z. Yang. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21(182):1–43, 2020.
39. X. Chen, W. Liu, and Y. Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

40. X. Chen, W. Liu, and Y. Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 117(540):1858–1874, 2022.
41. V. Chernozhukov and I. Fernández-Val. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2):559–589, 2011.
42. V. Chernozhukov and C. Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.
43. H. Cramér. Sur un nouveau theoreme-limite de la theorie des probabilities. *Actualites Scientifiques et Industrielles*, 736:5–23, 1938.
44. M. Crouch, A. McGregor, G. Valiant, and D. P. Woodruff. Stochastic streams: Sample complexity vs. space complexity. In *24th Annual European Symposium on Algorithms (ESA 2016)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2016.
45. A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pages 2260–2268. PMLR, 2020.
46. A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in Neural Information Processing Systems*, 32:15210–15219, 2019.
47. J. Dagpunar. An easily implemented generalised inverse Gaussian generator. *Communications in Statistics-Simulation and Computation*, 18(2):703–710, 1989.
48. A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27:1646–1654, 2014.
49. A. Défossez and F. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *International Conference on Artificial Intelligence and Statistics*, volume 38, pages 205–213. PMLR, 2015.
50. A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer Berlin, Heidelberg, 2nd edition, 1998.
51. J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28(126):549–560, 1974.
52. A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.
53. A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020.
54. A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017.
55. T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg. Finite-time analysis of stochastic gradient descent under Markov randomness. *arXiv preprint arXiv:2003.10973*, 2020.
56. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.
57. J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.
58. J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49(1):21–48, 2021.
59. A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.
60. P. Englund and Y. M. Ioannides. House price dynamics: an international empirical perspective. *Journal of housing economics*, 6(2):119–136, 1997.
61. V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.
62. X. Fan, I. Grama, and Q. Liu. Sharp large deviation results for sums of independent random variables. *Science China Mathematics*, 58:1939–1958, 2015.
63. X. Fan, I. Grama, Q. Liu, and Q.-M. Shao. Self-normalized cramér type moderate deviations for stationary sequences and applications. *Stochastic Processes and their Applications*, 130(8):5124–5148, 2020.

64. X. Fan, H. Hu, and X. Ma. Cramér moderate deviations for the elephant random walk. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(2):023402, 2021.
65. X. Fan, H. Hu, and L. Xu. Normalized and self-normalized Cramér-type moderate deviations for the Euler-Maruyama scheme for the SDE. *Science China Mathematics*, pages 1–16, 2024.
66. Y. Fan, J.-S. Li, and N. Lin. Residual projection for quantile regression in vertically partitioned big data. *Data Mining and Knowledge Discovery*, 37(2):710–735, 2023.
67. C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *Advances in Neural Information Processing Systems*, pages 686–696, 2018.
68. S. Firpo, N. M. Fortin, and T. Lemieux. Unconditional quantile regressions. *Econometrica*, 77(3):953–973, 2009.
69. B. Fitzenberger, R. Koenker, J. Machado, and B. Melly. Economic applications of quantile regression 2.0. *Empirical Economics*, 62(1):1–6, 2022.
70. B. Fitzenberger, R. Koenker, and J. A. Machado. *Economic applications of quantile regression*. Springer Science & Business Media, 2013.
71. N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.
72. D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.
73. R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, pages 728–763. PMLR, 2015.
74. S. Gadat and F. Panloup. Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.
75. J. Gama, R. Sebastiao, and P. P. Rodrigues. On evaluating stream learning algorithms. *Machine learning*, 90:317–346, 2013.
76. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.
77. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.
78. R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.
79. E. L. Glaeser and C. G. Nathanson. An extrapolative model of house price dynamics. *Journal of Financial Economics*, 126(1):147–170, 2017.
80. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.
81. R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.
82. P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.
83. X. He, X. Pan, K. M. Tan, and W.-X. Zhou. Scalable estimation and inference for censored quantile regression process. *The Annals of Statistics*, 50(5):2899–2924, 2022.
84. X. He, X. Pan, K. M. Tan, and W.-X. Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.
85. D. L. Huo, Y. Chen, and Q. Xie. Effectiveness of constant stepsize in Markovian LSA and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20447–20455, 2024.
86. P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, V. K. Pillutla, and A. Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, 2018.

87. P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory*, pages 545–604, 2018.
88. P. Jain, P. Netrapalli, S. M. Kakade, R. Kidambi, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.
89. Y. Ji, N. Lin, and B. Zhang. Model selection in binary and tobit quantile regression using the gibbs sampler. *Computational Statistics & Data Analysis*, 56(4):827–839, 2012.
90. Y. Jin, T. Xiao, and K. Balasubramanian. Statistical inference for Polyak-Ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.
91. R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
92. C. Joutard. Sharp large deviations in nonparametric estimation. *Nonparametric Statistics*, 18(3):293–306, 2006.
93. M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.
94. B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.
95. K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021.
96. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
97. R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.
98. J. Konečnỳ and P. Richtárik. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017.
99. H. Kozumi and G. Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.
100. A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21(155):1–52, 2020.
101. H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer, 2003.
102. H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal on Control and Optimization*, 31(4):1045–1062, 1993.
103. C. Lakshminarayanan and C. Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.
104. G. Lan, Z. Li, and Y. Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.
105. G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1-2):167–215, 2018.
106. L. Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.
107. N. Le Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.
108. S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

109. S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics*, page 105673, 2024.
110. S. Lee and S. J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(55):1705–1744, 2012.
111. L. Lei and M. Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 148–156, 2017.
112. L. Lei and M. I. Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.
113. C. J. Li. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *Available at SSRN:* `https://ssrn.com/abstract=4934923` *or* `http://dx.doi.org/10.2139/ssrn.4934923`, 2024.
114. C. J. Li. ROOT-SGD: Sharp nonasymptotics and near-optimal asymptotics in a single algorithm. *Available at SSRN:* `https://ssrn.com/abstract=4869694` *or* `http://dx.doi.org/10.2139/ssrn.4869694`, 2024.
115. G. Li, W. Wu, Y. Chi, C. Ma, A. Rinaldo, and Y. Wei. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, 70(8):5969–5999, 2024.
116. T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical inference using SGD. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
117. X. Li, J. Liang, X. Chang, and Z. Zhang. Statistical estimation and online inference via local SGD. In *Conference on Learning Theory*, pages 1613–1661. PMLR, 2022.
118. X. Li, J. Liang, X. Chen, and Z. Zhang. Stochastic approximation MCMC, online inference, and applications in optimization of queueing systems. *arXiv preprint arXiv:2309.09545*, 2023.
119. X. Li, J. Liang, and Z. Zhang. Online statistical inference for nonlinear stochastic approximation with Markovian data. *arXiv preprint arXiv:2302.07690*, 2023.
120. Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.
121. T. Liang and W. J. Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):431–456, 2019.
122. H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.
123. J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.
124. L. Ljung, G. Pflug, and H. Walk. *Stochastic Approximation and Optimization of Random Systems*. Springer, 1992.
125. L. Luo and P. X.-K. Song. Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):69–97, 2020.
126. L. Luo and P. X.-K. Song. Multivariate online regression analysis with heterogeneous streaming data. *Canadian Journal of Statistics*, 51(1):111–133, 2023.
127. Y. Luo, X. Huo, and Y. Mei. Covariance estimators for the ROOT-SGD algorithm in online learning. *arXiv preprint arXiv:2212.01259*, 2022.
128. L. Ma and L. Pohlman. Return forecasts and optimal portfolio construction: a quantile regression approach. *The European Journal of Finance*, 14(5):409–425, 2008.
129. R. Man, X. Pan, K. M. Tan, and W.-X. Zhou. A unified algorithm for penalized convolution smoothed quantile regression. *Journal of Computational and Graphical Statistics*, 33(2):625–637, 2024.
130. S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.
131. Y. Miao, Y.-X. Chen, and S.-F. Xu. Asymptotic properties of the deviation between order statistics and p-quantile. *Communications in Statistics—Theory and Methods*, 40(1):8–14, 2010.

132. H. Moon and W.-X. Zhou. High-dimensional composite quantile regression: Optimal statistical guarantees and fast algorithms. *Electronic Journal of Statistics*, 17(2):2067–2119, 2023.
133. W. Mou, K. Khamaru, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. Optimal variance-reduced stochastic approximation in Banach spaces. *arXiv preprint arXiv:2201.08518*, 2022.
134. W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997, 2020.
135. W. Mou, A. Pananjady, and M. J. Wainwright. Optimal oracle inequalities for projected fixed-point equations, with applications to policy evaluation. *Mathematics of Operations Research*, 48(4):2308–2336, 2023.
136. E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24:451–459, 2011.
137. D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.
138. J. Negrea, J. Yang, H. Feng, D. M. Roy, and J. H. Huggins. Tuning stochastic gradient algorithms for statistical inference via large-sample asymptotics. *arXiv preprint arXiv:2207.12395*, 2022.
139. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
140. Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR*, 269(3):543–547, 1983.
141. Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.
142. Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.
143. Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.
144. L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.
145. L. M. Nguyen, P. H. Nguyen, P. Richtárik, K. Scheinberg, M. Takáč, and M. van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.
146. L. M. Nguyen, K. Scheinberg, and M. Takáč. Inexact SARAH algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.
147. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.
148. N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.
149. B. T. Polyak. A new method of stochastic approximation type. *Automat. i Telemekh*, 51(7 pt. 2):937–946, 1990.
150. B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
151. S. Portnoy and R. Koenker. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.
152. R.-D. Reiss. On the accuracy of the normal approximation for quantiles. *The Annals of Probability*, pages 741–744, 1974.
153. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.
154. A. Roy and K. Balasubramanian. Online covariance estimation for stochastic gradient descent under Markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.
155. D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

156. J. Saastamoinen. Quantile regression analysis of dispersion of stock returns-evidence of herding? 2008.
157. S. Samsonov, D. Tiapkin, A. Naumov, and E. Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *Conference on Learning Theory*, pages 4511–4547. PMLR, 2024.
158. R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.
159. S. Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754, 2016.
160. S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2):567–599, 2013.
161. C. Shi, R. Song, W. Lu, and R. Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116(535):1307–1318, 2021.
162. R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.
163. J. Stachurski. *Economic dynamics: theory and computation*. MIT Press, 2009.
164. W. J. Su and Y. Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.
165. W. J. Su and Y. Zhu. HiGrad: Uncertainty quantification for online learning and stochastic approximation. *Journal of Machine Learning Research*, 24(124):1–53, 2023.
166. K. M. Tan, H. Battey, and W.-X. Zhou. Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of machine learning research*, 23(272):1–61, 2022.
167. K. M. Tan, L. Wang, and W.-X. Zhou. High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):205–233, 2022.
168. P. Tarres and Y. Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.
169. P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.
170. P. Toulis, D. Tran, and E. Airoldi. Towards stability and optimality in stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1290–1298. PMLR, 2016.
171. Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, pages 1–67, 2021.
172. N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Conference On Learning Theory*, pages 650–687. PMLR, 2018.
173. A. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.
174. A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.
175. A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer, 1996.
176. W. N. van Wieringen and H. Binder. Sequential learning of regression models by penalized estimation. *Journal of Computational and Graphical Statistics*, 31(3):877–886, 2022.
177. S. Volgushev, S.-K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019.
178. J. Wang, K. Wu, A. Sim, and S. Hwangbo. Accurate signal timing from high frequency streaming data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4852–4854. IEEE, 2017.
179. K. Wang, H. Wang, and S. Li. Renewable quantile regression for streaming datasets. *Knowledge-Based Systems*, 235:107675, 2022.

180. Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2406–2416, 2019.
181. Z. Wei, W. Zhu, and W. B. Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *arXiv preprint arXiv:2307.06915*, 2023.
182. J. Wen, S. Yang, C. D. Wang, Y. Jiang, and R. Li. Feature-splitting algorithms for ultrahigh dimensional quantile regression. *Journal of Econometrics*, page 105426, 2023.
183. B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3639–3647, 2016.
184. S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.
185. E. Xia, K. Khamaru, M. J. Wainwright, and M. I. Jordan. Instance-dependent confidence and early stopping for reinforcement learning. *Journal of Machine Learning Research*, 24(392):1–43, 2023.
186. L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.
187. C. Xie and Z. Zhang. A statistical online inference approach in averaged stochastic approximation. *Advances in Neural Information Processing Systems*, 35:8998–9009, 2022.
188. S. Xu and Y. Miao. Limit behaviors of the deviation between the sample quantiles and the quantile. *Filomat*, 25(2):197–206, 2011.
189. W. Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*, 2011.
190. J. Yang, X. Meng, and M. W. Mahoney. Quantile regression for large-scale applications. *SIAM Journal on Scientific Computing*, 36(5):S78–S110, 2014.
191. X. Yang, N. N. Narisetty, and X. He. A new approach to censored quantile regression estimation. *Journal of Computational and Graphical Statistics*, 27(2):417–425, 2018.
192. Y. Yang, H. J. Wang, and X. He. Posterior inference in bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, 84(3):327–344, 2016.
193. B. Ying and A. H. Sayed. Performance limits of stochastic sub-gradient learning, part i: Single agent case. *Signal Processing*, 144:271–282, 2018.
194. K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.
195. K. Yu and J. Stander. Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics*, 137(1):260–276, 2007.
196. K. Yu and J. Zhang. A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, 34(9-10):1867–1879, 2005.
197. L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *Advances in Neural Information Processing Systems*, 34:4234–4248, 2021.
198. L. Yu and N. Lin. Admm for penalized quantile regression in big data. *International Statistical Review*, 85(3):494–518, 2017.
199. L. Yu, N. Lin, and L. Wang. A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 26(4):935–939, 2017.
200. Q. Zhang and W. Wang. A fast algorithm for approximate quantiles in high speed data streams. In *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*, pages 29–29. IEEE, 2007.
201. T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine learning*, pages 919–926, 2004.
202. Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013.
203. D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 3921–3932, 2018.
204. W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

205. W. Zhu, Z. Lou, Z. Wei, and W. B. Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

206. Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty. Local minimax complexity of stochastic convex optimization. *Advances in Neural Information Processing Systems*, pages 3431–3439, 2016.

207. Y. Zhu and J. Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *2021 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2021.

208. H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.

Appendix C
Appendix for Chapter 3

C.1 Experimental setup

We conducted an experiment to assess the performance of the `Online-QR` algorithm in estimating regression coefficients from streaming data. For illustrative purposes, we set the quantile level $\tau = 0.5$. The experiment was structured as follows:

- **Data Generation:** We generated synthetic data consisting of $N = 100,000$ observations over $T = 200$ time blocks. Each time block contained $n = \frac{N}{T}$ observations, with predictor variables $\mathbf{X}$ sampled from a $p = 50$ dimensional multivariate normal distribution $\mathscr{N}(\mathbf{0}_p, \mathbf{V})$, where $\mathbf{V}_{jk} = 0.5^{|j-k|}$.
- **Regression Model:** The response variable $\mathbf{y}$ was generated according to the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (-1)^j$ for $j = 1, 2, \dots, p$ represents the true regression coefficients, and $\boldsymbol{\varepsilon} \sim \mathscr{N}(0,1)$ denotes the random error.
- **Algorithm Implementation:** We implemented the `Online-QR` algorithm to estimate the regression coefficients $\boldsymbol{\beta}$. The algorithm processed each time block in an online fashion, updating the estimate $\overline{\boldsymbol{\beta}}^{\text{QR}}$ using QR decomposition at each step.
- **Evaluation Metrics:** We evaluated the algorithm's performance using two metrics:

  1. **Absolute Estimation Error (AE):** Defined as the sum of absolute differences between the estimated and true regression coefficients across all dimensions.
  2. **ℓ_1-norm Distance:** The sum of absolute differences between the final estimated coefficients $\overline{\boldsymbol{\beta}}_T^{\text{QR}}$ and the true coefficients $\boldsymbol{\beta}$.

The R code used for data generation, algorithm implementation, and metric computation is shown below. Each run of the code produces varying outputs due to the random nature of data generation, ensuring robustness and generalizability of the experiment results.

```
# Load necessary library
```

```
library(MASS)

# Set parameters
N <- 100000
T <- 200
n <- N / T
p <- 50

# Generate the covariance matrix V
V <- matrix(0, nrow = p, ncol = p)
for (j in 1:p) {
  for (k in 1:p) {
    V[j, k] <- 0.5^abs(j - k)
  }
}

# Generate the covariates from the multivariate normal distribution
X <- MASS::mvrnorm(n = N, mu = rep(0, p), Sigma = V)

# Generate the regression coefficients
beta <- (-1)^(1:p)

# Generate the random error epsilon
epsilon <- rnorm(N)

# Calculate y using the linear model
y <- X %*% beta + epsilon

# Split data into T blocks
X_blocks <- array(NA, dim = c(n, p, T))
y_blocks <- matrix(NA, nrow = n, ncol = T)

for (t in 1:T) {
  start_idx <- (t - 1) * n + 1
  end_idx <- t * n
  X_blocks[,,t] <- X[start_idx:end_idx, ]
  y_blocks[,t] <- y[start_idx:end_idx]
}

# Initialize beta_QR_0
bar_beta_QR <- rep(0, p)

# Initialize AE
AE <- 0

# Algorithm for QR estimation in stream data
for (t in 1:T) {
  # (a) Store (X_t, y_t) and bar_beta_QR(t-1)
  X_t <- X_blocks[,,t]
  y_t <- y_blocks[,t]

  # (b) Compute tilde_v_t and tilde_beta_QR_t
  tilde_v_t <- rep(0, n)
  tilde_beta_QR_t <- rep(0, p)
```

```
    for (i in 1:n) {
      hat_delta_it2 <- (1/8) * (y_t[i] - X_t[i,] %*% bar_beta_QR)^2
      tilde_v_it <- hat_delta_it2 + 0.5
      tilde_v_t[i] <- tilde_v_it
    }

    X_t_star <- cbind(1, X_t)  # Add intercept to X_t
    y_t_star <- y_t

    # Compute tilde_beta_QR_t using QR solution
    tilde_beta_QR_t <- solve(t(X_t_star) %*% X_t_star) %*% t(X_t_star) %*% y_t_star

    # (c) Update bar_beta_QR_t
    if (t == 1) {
      bar_beta_QR <- tilde_beta_QR_t[-1]  # Initialize bar_beta_QR_1 without intercept
    } else {
      bar_beta_QR <- (1/t) * tilde_beta_QR_t[-1] + ((t-1)/t) * bar_beta_QR
    }

    # Compute AE for this iteration and accumulate
    AE <- AE + sum(abs(bar_beta_QR - beta))
  }

  # Compute the L1-norm distance between bar_beta_QR and beta
  L1_distance <- sum(abs(bar_beta_QR - beta))

  # Print results
  cat("Final Estimate bar_beta_QR_T:\n")
  print(c(0, bar_beta_QR))  # Add intercept back for final estimate

  cat("Absolute Estimation Error (AE):\n")
  print(AE)

  cat("L1-norm distance between bar_beta_QR and beta:\n")
  print(L1_distance)
```

This experimental setup allowed us to assess the effectiveness and accuracy of the Online-QR algorithm in handling streaming data and estimating regression coefficients under varying data volumes and dimensions.

C.1.1 Experimental results

To evaluate the performance of the Online-QR algorithm, we ran the experiment using synthetic data generated as described earlier. Below are the results from a sample run:

```
    > cat("Absolute Estimation Error (AE):\n")
    Absolute Estimation Error (AE):
```

```
> print(AE)
[1] 61.03646

> cat("L1-norm distance between bar_beta_QR and beta:\n")
L1-norm distance between bar_beta_QR and beta:
> print(L1_distance)
[1] 0.1681378
```

These metrics quantify the accuracy of our algorithm in estimating the regression coefficients under the experimental conditions.

Appendix D
Appendix for Chapter 4

In the appendix, we provide deferred proofs for theorems and lemmas in the main text organized as follows. §D.1 proves the additional results for both the nonasymptotic and the asymptotic convergence properties of our ROOT-SGD algorithm. §D.2 complements our asymptotic efficiency result in §4.3.3 and establishes an additional asymptotic result for constant-step-size ROOT-SGD. §D.3 presents auxiliary lemmas stated in §D.1. Finally, §D.4 proves necessary lemmas for the proof of Proposition 2 (in §D.2.1).

D.1 Proofs of additional nonasymptotic and asymptotic results

We provide the convergence rate analysis and the proofs of our remaining theorems in this section.

D.1.1 Intermediate result Proposition 1 and its proof

En route our proof of Theorem 4.2 we state and prove an intermediate upper-bound result for single-epoch version of ROOT-SGD. For our convenience we forgo tracking the universal constants (which can be change at each appearance) due to complications of our derivations.

Proposition 1 (Improved nonasymptotic upper bound, single-epoch ROOT-SGD) *Under Assumptions 1, 4, 5, 6, suppose that we run Algorithm 2 with step-size* $\eta \in \big(0, \frac{1}{56L} \wedge \frac{\mu}{64\ell_{\Xi}^2}\big]$. *Then for any* $T \geq 1$, *the iterate* θ_T *satisfies the bound*

$$\mathbb{E}\,\|\nabla F(\theta_T)\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left\{\frac{\ell_{\Xi}^2 \eta}{\mu} + \frac{\log T}{\eta \mu T} + \frac{\ell_{\Xi}^2 \log T}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} + \frac{C L_\gamma \widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2}$$

$$+ \frac{C\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{CL_\gamma\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}} \tag{D.1}$$

A few remarks are in order. When setting $T \to \infty$ the leading-order term $(1 + \frac{C\ell_\Xi^2\eta}{\mu})\frac{\sigma_*^2}{T}$ of the nonasymptotic bound (D.1) nearly matches the optimal statistical risk for the gradient norm with unit pre-factor when η is prescribed as positively small, and as will be seen later it matches the asymptotic Proposition 2 under a shared umbrella of assumptions. It can be observed that the dependence on the initial gradient norm $\|\nabla F(\theta_0)\|_2$ decays polynomially, which is generally unavoidable for single-epoch ROOT-SGD, as the gradient noise at the initial point θ_0 is also averaged along the iterates. However, as we will see anon, an improved guarantee can be obtained by appropriately re-starting the algorithm, leading to near-optimal guarantees in terms of the gradient norm. In addition, we note that the high-order terms of Eq. (D.1) contains terms that depend on the step-size η at opposite directions which demands a trade-off. We forgo optimizing the step-size as is the conduct in our multi-epoch result.

For the rest of §D.1.1 we prepare to prove Proposition 1. From the discussions in §4.2.2 we decomposes $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$ as the summation of three terms:

$$\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\|v_t - z_t\|_2^2 = \mathbb{E}\|v_t\|_2^2 + \mathbb{E}\|z_t\|_2^2 - 2\mathbb{E}\langle v_t, z_t\rangle. \tag{D.2}$$

En route our proof, we provide estimations for $\mathbb{E}\|tv_t\|_2^2$, $\mathbb{E}\|tz_t\|_2^2$ and $\mathbb{E}\langle tz_t, tv_t\rangle$ separately, where our main focus will be on bounding the cross term. On a very intuitive and high-level viewpoint, when comparing with the Polyak-Ruppert-Juditsky analysis, we can roughly think of the $(\eta tv_t : t \geq 0)$ process acts like a last-iterate SGD (as it is in the quadratic minimization case) and is *fast* and *small*. The tz_t process more resembles random walk at a slower rate driven by the same noise sequence. The two timescale intuitions beneath is that two fast-slow discounted random walks processes driven by the same noise has an inner product that is approximately the second moment of the fast process. In our case this results in the ”asymptotically independence” of the two processes in the sense that $\mathbb{E}\langle tz_t, tv_t\rangle$ scales as $\mathbb{E}\|tv_t\|_2^2$, so $\nabla F(\theta_{t-1}) = v_t - z_t$ is approximately of the same scale as z_t in its first and second orders.

We first introduce the following lemma which is an essential part of the proof:

Lemma D.1 (Sharp bound on v_t). *Under the setting of Theorem 4.1, there exists a universal constant $c > 0$, such that for $T \geq T_0 + 1$, we have:*

$$\mathbb{E}\|v_T\|_2^2 \leq \frac{c\sigma_*^2}{\eta\mu T^2} + \frac{c}{\eta^4\mu^4T^4}\|\nabla F(\theta_0)\|_2^2. \tag{D.3}$$

We defer the proof of Lemma D.1 to §D.1.1.1. This lemma, along with Theorem 4.1, helps conclude the following bound on z_t that has a leading-order term of near-unity pre-factor, that is, $(1 + o(1))\frac{\sigma_*}{\sqrt{t}}$:

Lemma D.2 (Sharp bound on z_t). *Under settings of Theorem 4.1, the following bounds hold true for $T \geq T_0 + 1$:*

$$\mathbb{E}\,\|z_T\|_2^2 - \frac{\sigma_*^2}{T} \leq c\left\{\frac{\ell_\Xi^2 \eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} + c\frac{\ell_\Xi \sigma_*}{\mu}\cdot\frac{T_0}{T^2}\,\|\nabla F(\theta_0)\|_2 + c\frac{{T_0}^2}{T^2}\,\|\nabla F(\theta_0)\|_2^2, \tag{D.4}$$

for some universal constant $c > 0$.

See §D.1.1.2 for the proof of this lemma.

Finally, we need the following lemma, which bounds the cross term $\mathbb{E}\langle v_t, z_t\rangle$. Under the Lipschitz condition on the Hessian matrix and additional moment conditions, this lemma provides significant sharper bound than the naïve bound obtained by applying the Cauchy-Schwartz inequality and invoking the previous two lemmas.

Lemma D.3 (Sharp bound on the cross term). *Under settings of Theorem 4.1, we have the following bound for any $T \geq T_0 + 1$:*

$$\left|\mathbb{E}\langle v_T, z_T\rangle\right| \leq c\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 T^4}\right)\log T + cL_\gamma\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right), \tag{D.5}$$

for some universal constant $c > 0$.

See §D.1.1.3 for the proof of this lemma.

Taking the aforementioned lemmas as given, we are ready to prove the sharp bound. In particular, by substituting these three lemmas into the decomposition (D.2), we have the following bound

$$\begin{aligned}
&\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^2 - \frac{\sigma_*^2}{T} = \mathbb{E}\,\|v_T - z_T\|_2^2 - \frac{\sigma_*^2}{T} = \left(\mathbb{E}\,\|z_T\|_2^2 - \frac{\sigma_*^2}{T}\right) + \mathbb{E}\,\|v_T\|_2^2 - 2\mathbb{E}\langle v_T, z_T\rangle \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + C\left(\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\,\|\nabla F(\theta_0)\|_2 + \frac{\ell_\Xi^2}{\mu^2}\cdot\frac{T_0}{T^2}\,\|\nabla F(\theta_0)\|_2^2\right) \\
&\quad + C\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right)\log T + 6C_0L_\gamma\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right) \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \underbrace{\frac{\ell_\Xi}{\mu\sqrt{T}}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \\
&\quad + C\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \underbrace{\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\,\|\nabla F(\theta_0)\|_2}\right) + C\frac{L_\gamma\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}.
\end{aligned}$$

Absorbing the bracketed cross terms into corresponding sum of the squares, this gives Eq. (D.1) and concludes Proposition 1.

D.1.1.1 Proof of Lemma D.1

Our main technical tools is the following Lemma D.4, which recursively bound the second moments of v_t:

Lemma D.4. *Under the setting of Theorem 4.1, we have the following bound for* $t \geq T_0 + 1$

$$t^2 \mathbb{E}\|v_t\|_2^2 \leq \left(1 - \frac{\eta\mu}{2}\right)(t-1)^2 \mathbb{E}\|v_{t-1}\|_2^2 + \frac{10}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 4\sigma_*^2. \tag{D.6}$$

See §D.3.3.2 for the proof of this lemma.

On the other hand, invoking Theorem 4.1, we have that

$$\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \leq \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2} + \frac{28\ \sigma_*^2}{t}, \qquad \text{for } t \geq T_0 + 1.$$

Now, to combine everything together, we conclude from (4.12) and (D.6) that

$$\begin{aligned} t^2 \mathbb{E}\|v_t\|_2^2 &\leq \left(1 - \frac{\eta\mu}{2}\right)(t-1)^2 \mathbb{E}\|v_{t-1}\|_2^2 + \frac{10}{\eta\mu}\left[\frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2} + \frac{28\ \sigma_*^2}{t}\right] + 4\sigma_*^2 \\ &\leq \left(1 - \frac{\eta\mu}{2}\right)(t-1)^2 \mathbb{E}\|v_{t-1}\|_2^2 + c\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3 t^2} + c\sigma_*^2. \end{aligned} \tag{D.7}$$

Multiplying both sides by t^2, we obtain that

$$\begin{aligned} t^4 \mathbb{E}\|v_t\|_2^2 &\leq \left(1 - \frac{\eta\mu}{2}\right) t^2 (t-1)^2 \mathbb{E}\|v_{t-1}\|_2^2 + \frac{c\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + c\sigma_*^2 t^2 \\ &\leq \left(1 - \frac{\eta\mu}{6}\right)(t-1)^4 \mathbb{E}\|v_{t-1}\|_2^2 + \frac{c\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + c\sigma_*^2 t^2, \end{aligned}$$

for time index t satisfying $t \geq T_0 \geq \frac{6}{\eta\mu}$. This gives, by solving the recursion,

$$\begin{aligned} T^4 \mathbb{E}\|v_T\|_2^2 &\leq \left(1 - \frac{\eta\mu}{6}\right)^{T-T_0} T_0{}^4 \mathbb{E}\left\|v_{T_0}\right\|_2^2 + c\sum_{t=T_0+1}^{T}\left(1 - \frac{\eta\mu}{6}\right)^{T-t}\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + \sigma_*^2 T^2\right) \\ &\leq \left(1 - \frac{\eta\mu}{6}\right)^{T-T_0} T_0{}^4 \mathbb{E}\left\|v_{T_0}\right\|_2^2 + 6c\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4} + 6c\frac{\sigma_*^2}{\eta\mu}T^2. \end{aligned}$$

It suffices to bound the initial condition $\mathbb{E}\left\|v_{T_0}\right\|_2^2$. Recall that $v_{T_0} = \frac{1}{T_0}\sum_{s=1}^{T_0}\nabla f(\theta_0; \xi_s)$, which is average of i.i.d. random vectors. It immediately follows from Assumptions 2 and 3 that:

$$\mathbb{E}\left\|v_{T_0}\right\|_2^2 \le \|\nabla F(\theta_0)\|_2^2 + \frac{2\sigma_*^2}{T_0} + \frac{2\ell_\Xi^2\,\|\nabla F(\theta_0)\|_2^2}{\mu^2 T_0}.$$

Putting them together, we complete the proof of this lemma.

D.1.1.2 Proof of Lemma D.2

Recalling that the recursive update rule of z_t reveals an underlying martingale structure

$$tz_t = (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}).$$

Adding and subtracting the $\varepsilon_t(\theta^*)$ term in the above display we express the noise increment as

$$tz_t - (t-1)z_{t-1} = \varepsilon_t(\theta^*) + \underbrace{(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)}_{=:\,\zeta_t}.$$

In words, the increment of tz_t splits into two parts: the additive part $\varepsilon_t(\theta^*)$ and the multiplicative part ζ_t. Taking expectation on the squared norm in above and using the property of square-integrable martingales, we have via further expanding the square on the right hand

$$t^2\mathbb{E}\,\|z_t\|_2^2 - (t-1)^2\mathbb{E}\,\|z_{t-1}\|_2^2 = \mathbb{E}\,\|\varepsilon_t(\theta^*) + \zeta_t\|_2^2 = \mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^2 + \mathbb{E}\,\|\zeta_t\|_2^2 + 2\mathbb{E}\langle\varepsilon_t(\theta^*),\,\zeta_t\rangle.$$

Telescoping the above equality for $t = T_0+1,\dots,T$ gives

$$T^2\mathbb{E}\,\|z_T\|_2^2 - T_0{}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^2 + \sum_{t=T_0+1}^{T}\mathbb{E}\,\|\zeta_t\|_2^2 + 2\sum_{t=T_0+1}^{T}\mathbb{E}\langle\varepsilon_t(\theta^*),\,\zeta_t\rangle.$$

By Assumption 5, we have $\mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^2 = \sigma_*^2$. For the additional noise ζ_t, Young's inequality leads to the bound

$$\begin{aligned}\mathbb{E}\,\|\zeta_t\|_2^2 &\le \left(\ell_\Xi(t-1)\sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta_{t-2}\|_2^2} + \ell_\Xi\sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta^*\|_2^2}\right)^2\\ &\le \left(\eta\ell_\Xi(t-1)\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^2} + \frac{\ell_\Xi}{\mu}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2}\right)^2\\ &\le 2\eta^2\ell_\Xi^2(t-1)^2\mathbb{E}\,\|v_{t-1}\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2.\end{aligned} \tag{D.8}$$

It remains to bound the summation of the cross term. Observing that:

$$\begin{aligned}\sum_{t=T_0+1}^{T} \mathbb{E}\langle \varepsilon_t(\theta^*),\, \zeta_t\rangle &= \sum_{t=T_0+1}^{T} \mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*) + (t-1)\left(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right)\rangle \\ &= \sum_{t=T_0+1}^{T} \Big\{ t\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle \Big\}.\end{aligned}$$

Since the random samples $(\xi_t)_{t\geq 1}$ are i.i.d. and the iterate θ_{t-2} is independent of the sample ξ_{t-1}, we have that

$$\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle = \mathbb{E}\langle \varepsilon_{t-1}(\theta^*),\, \varepsilon_{t-1}(\theta_{t-2}) - \varepsilon_{t-1}(\theta^*)\rangle.$$

Consequently, we can re-write the quantity of interests as a telescope sum, leading to the following identity:

$$\begin{aligned}\sum_{t=T_0+1}^{T} \mathbb{E}\langle \varepsilon_t(\theta^*),\, \zeta_t\rangle &= \sum_{t=T_0+1}^{T} \Big\{ t\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle \Big\} \\ &= \sum_{t=T_0+1}^{T} \Big\{ t\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle \varepsilon_{t-1}(\theta^*),\, \varepsilon_{t-1}(\theta_{t-2}) - \varepsilon_{t-1}(\theta^*)\rangle \Big\} \\ &= T\cdot\mathbb{E}\langle \varepsilon_T(\theta^*),\, \varepsilon_T(\theta_{T-1}) - \varepsilon_T(\theta^*)\rangle - T_0\cdot\mathbb{E}\langle \varepsilon_{T_0}(\theta^*),\, \varepsilon_{T_0}(\theta_{T_0-1}) - \varepsilon_{T_0}(\theta^*)\rangle.\end{aligned}$$

In order to bound the inner product terms, we invoke the Cauchy-Schwartz inequality and Assumption 3. For each $t \geq T_0$, we have that

$$\begin{aligned}|t\cdot\mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle| &\leq t\cdot\sqrt{\mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^2}\cdot\sqrt{\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2} \\ &\leq t\sigma_*\ell_\Xi\sqrt{\mathbb{E}\,\|\theta_{t-1} - \theta^*\|_2^2} \leq \frac{t\sigma_*\ell_\Xi}{\mu}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2}. \qquad \text{(D.9)}\end{aligned}$$

Plugging $t = T_0$ and $t = T$ into Eq. (D.9) separately and combining with Eq. (D.8), we have that

$$\begin{aligned}T^2\mathbb{E}\,\|z_T\|_2^2 \leq {T_0}^2\mathbb{E}\,\|z_{T_0}\|_2^2 + (T-T_0)\sigma_*^2 + 2\eta^2\ell_\Xi^2\sum_{t=T_0+1}^{T}(t-1)^2\mathbb{E}\,\|v_{t-1}\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=T_0+1}^{T}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 \\ + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + T_0\,\|\nabla F(\theta_0)\|_2\right). \qquad \text{(D.10)}\end{aligned}$$

The above bound involves the second moments of the vectors v_t and $\nabla F(\theta_t)$. We recall the following bounds from Theorem 4.1 and Lemma D.1, for each $t \geq T_0$:

$$\begin{aligned}&\mathbb{E}\,\|v_t\|_2^2 \leq \frac{c\sigma_*^2}{\eta\mu t^2} + \frac{c}{\eta^4\mu^4 t^4}\,\|\nabla F(\theta_0)\|_2^2, \quad \text{and} \\ &\mathbb{E}\,\|\nabla F(\theta_t)\|_2^2 \leq \frac{c\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2} + \frac{c\sigma_*^2}{t}.\end{aligned}$$

Substituting these bounds to Eq. (D.10), we note that

$$\sum_{t=T_0+1}^{T}(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 \le \frac{c\sigma_*^2 T}{\eta\mu}+\frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4 T_0}, \quad \text{and}$$

$$\sum_{t=T_0+1}^{T}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 \le c\sigma_*^2\log\left(\frac{T}{T_0}\right)+\frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2 T_0}.$$

Finally, for the burn-in period we note that $z_{T_0}=\sum_{s=1}^{T_0}\varepsilon_s(\theta_0)$ is a sum of T_0 i.i.d. random vectors, and hence the following estimation holds

$$\begin{aligned}
&{T_0}^2\mathbb{E}\left\|z_{T_0}\right\|_2^2 = T_0\mathbb{E}\left\|\varepsilon_1(\theta_0)\right\|_2^2\\
&\le T_0\mathbb{E}\left\|\varepsilon_1(\theta_0)-\varepsilon_1(\theta^*)\right\|_2^2+2T_0\sqrt{\mathbb{E}\left\|\varepsilon_1(\theta^*)\right\|_2^2\mathbb{E}\left\|\varepsilon_1(\theta_0)-\varepsilon_1(\theta^*)\right\|_2^2}+T_0\mathbb{E}\left\|\varepsilon_1(\theta^*)\right\|_2^2\\
&\le T_0\ell_\Xi^2\left\|\theta_0-\theta^*\right\|_2^2+2T_0\ell_\Xi\sigma_*\left\|\theta_0-\theta^*\right\|_2+T_0\sigma_*^2 \le \frac{T_0\ell_\Xi^2}{\mu^2}\left\|\nabla F(\theta_0)\right\|_2^2+\frac{2T_0\ell_\Xi\sigma_*}{\mu}\left\|\nabla F(\theta_0)\right\|_2+T_0\sigma_*^2.
\end{aligned}$$

Some algebra yields (D.4) and hence the whole Lemma D.2.

D.1.1.3 Proof of Lemma D.3

First, by Cauchy-Schwartz inequality, we can easily observe that:

$$\left|\mathbb{E}\langle v_T, z_T\rangle\right| \le \sqrt{\mathbb{E}\left\|v_T\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|z_T\right\|_2^2} \le c\Big(\frac{\sigma_*}{T\sqrt{\eta\mu}}+\frac{\left\|\nabla F(\theta_0)\right\|_2}{\eta^2\mu^2T^2}\Big)\cdot\Big(\frac{\sigma_*}{\sqrt{T}}+\frac{\left\|\nabla F(\theta_0)\right\|_2}{\eta\mu T}\Big).$$

So for $T \le cT_0\log T_0$, the conclusion of this lemma is automatically satisfied. For the rest of this section, we assume that $\frac{T}{\log T} > cT_0$ for some universal constant $c>0$.

The proof requires some bounds on the fourth moment of the stochastic process defined by the algorithm. In particular, we need the following two lemmas. The first lemma is analogous to the bound in Theorem 4.1:

Lemma D.5 (Higher-order-moment bound on $\nabla F(\theta_{t-1})$). *Suppose Assumptions 1, 5 and 6 hold. Let the step-size $\eta \le \frac{1}{56L}\wedge\frac{\mu}{64\ell_\Xi^2}$ and the burn-in time $T_0=\left\lceil\frac{24}{\eta\mu}\right\rceil$. Then for any $T \ge T_0$, the estimator θ_T produced by the ROOT-SGD algorithm satisfies the bound*

$$\left(\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4\right)^{1/2} \le \frac{140\widetilde{\sigma_*}^2}{T+1}+\frac{60\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2(T+1)^2}. \tag{D.11}$$

Proof can be found in §D.3.1.

We also need a lemma on the fourth-moment bound of v_t, analogous to Lemma D.1:

Lemma D.6 (sharp higher-order-moment bound on v_t). *Under the setting of Proposition 1 we have the following bound for $T \ge T_0+1$*

$$\sqrt{\mathbb{E}\,\|v_T\|_2^4} \leq \frac{4484\widetilde{\sigma_*}^2}{\eta\mu T^2} + \frac{1359375}{\eta^4\mu^4 T^4}\,\|\nabla F(\theta_0)\|_2^2. \tag{D.12}$$

Proof can be found in §D.3.2.

Taking these two lemmas as given, we proceed with the proof. Following the two-time-scale intuition discussed in Section 4.2.2, the process v_t moves faster than the averaging process z_t. Therefore, it is reasonable to expect the correlation between v_t and $z_{t-\widetilde{T}^*}$ to be small, for sufficiently large time window $\widetilde{T}^* > 0$. For the rest of this section, we choose the window size:

$$\widetilde{T}^* = \frac{c}{\mu\eta}\log T, \quad \text{for some universal constant } c > 0. \tag{D.13}$$

Since we have assumed without loss of generality that $\frac{T}{\log T} > cT_0 = \frac{24c}{\eta\mu}$, the window size guarantees the relation $T - \widetilde{T}^* > T/2$.

We subtract off a $(t-\widetilde{T}^*)z_{t-\widetilde{T}^*}$ term the tz_t expression above, and decompose the absolute value of the cross term $|\mathbb{E}\langle v_t, tz_t\rangle|$ as:

$$|\mathbb{E}\langle tz_t, v_t\rangle| \leq (t-\widetilde{T}^*)\underbrace{\left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, v_t\rangle\right|}_{=:I_1} + \underbrace{\left|\mathbb{E}\langle tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}, v_t\rangle\right|}_{=:I_2}. \tag{D.14}$$

For bounding the term I_2, we make use of the recursive rule of tz_t to obtain the bound

$$\mathbb{E}\left\|tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2 = \mathbb{E}\left\|\sum_{s=t-\widetilde{T}^*+1}^{t}\left\{(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})) + \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*) + \varepsilon_s(\theta^*)\right\}\right\|_2^2$$
$$\leq \sum_{s=t-\widetilde{T}^*+1}^{t}\left\{(s-1)^2\eta^2\ell_\Xi^2\mathbb{E}\,\|v_{s-1}\|_2^2 + \frac{\ell_\Xi^2}{\mu^2}\mathbb{E}\,\|\nabla F(\theta_{s-1})\|_2^2 + \sigma_*^2\right\} \leq \widetilde{T}^*\cdot\left(\sigma_*^2 + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3 t^2}\right).$$

Consequently, we have the bound

$$\left|\mathbb{E}\langle tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}, v_t\rangle\right| \leq \sqrt{\mathbb{E}\left\|tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\,\|v_t\|_2^2}$$
$$\leq c\sqrt{\widetilde{T}^*}\left(\frac{\sigma_*}{\sqrt{\eta\mu}t} + \frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2 t^2}\right)\left(\sigma_* + \frac{\|\nabla F(\theta_0)\|_2}{\eta^{3/2}\mu^{3/2}t}\right) \leq c\left(\frac{\sigma_*^2}{\eta\mu t} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 t^3}\right)\sqrt{\log t}.$$

The bound for the term I_1 in the decomposition (D.14) is given by the following analysis: law of iterated expectations gives

$$\left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, v_t\rangle\right| = \left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, \mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\rangle\right| \leq \sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2}, \tag{D.15}$$

where the last inequality comes from applying the Cauchy-Schwarz inequality.

The second moment for $z_{t-\widetilde{T}^*}$ is relatively easy to estimate using Lemma D.2. It suffices to study the conditional expectation $\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})$. We claim the following bound:

$$\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} \le \frac{cL_\gamma}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu t}} + \frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2 t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}} + \frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2 t}\right). \tag{D.16}$$

We prove this inequality at the end of this section. Taking this bound as given, we now proceed with the proof for Lemma D.3.

Bringing this back to the inequality (D.15) and by utilizing the z_t bound by Lemma D.2, we have

$$\mathbb{E}\|z_t\|_2^2 \le C\left(\frac{\sigma_*^2}{t} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2}\right),$$

and thus

$$\begin{aligned}
&\left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, v_t\rangle\right| \le \sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2^2} \\
&\le \frac{cL_\gamma}{\mu}\left(\frac{\sigma_*}{\sqrt{t}} + \frac{\|\nabla F(\theta_0)\|_2}{\eta\mu t}\right)\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu t}} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta\mu^2 t}\right) \\
&\le \frac{cL_\gamma}{\mu}\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{3/2}t^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{9/2}t^{7/2}}\right).
\end{aligned}$$

Combining the bounds for I_1 and I_2 together, we estimate the cross term as:

$$\begin{aligned}
&\left|\mathbb{E}\langle tz_t, v_t\rangle\right| \\
&\le c(t-\widetilde{T}^*)\frac{L_\gamma}{\mu}\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{3/2}t^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{9/2}t^{7/2}}\right) + c\left(\frac{\sigma_*^2}{\eta\mu t} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 t^3}\right)\sqrt{\log t}.
\end{aligned} \tag{D.17}$$

We conclude by dividing both sides of Eq. (D.17) by T and arrive at the following bound:

$$\begin{aligned}
&\left|\mathbb{E}\langle v_T, z_T\rangle\right| \\
&\le c\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 T^4}\right)\sqrt{\log T} + cL_\gamma\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right).
\end{aligned}$$

This finishes our bound on the cross term and conclude Lemma D.3.

Proof of Eq (D.16)**:**

We note the following expansion:

$$\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) = \int_0^1 \nabla^2 F\left(\lambda\theta_{t-2} + (1-\lambda)\theta_{t-1}\right)(\theta_{t-1} - \theta_{t-2})d\lambda,$$

which leads to the following bound under the Lipschitz continuity condition for the Hessians (Assumption 4):

$$\begin{aligned}
&\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - \nabla^2 F(\theta^*)(\theta_{t-1} - \theta_{t-2})\right\|_2 \\
&= \int_0^1 \left\|\left(\nabla^2 F\left(\lambda\theta_{t-2} + (1-\lambda)\theta_{t-1}\right) - \nabla^2 F(\theta^*)\right)(\theta_{t-1} - \theta_{t-2})\right\|_2 d\lambda \\
&\le \eta L_\gamma \|v_{t-1}\|_2 \int_0^1 \|\lambda(\theta_{t-2} - \theta^*) + (1-\lambda)(\theta_{t-1} - \theta^*)\|_2 d\lambda \\
&\le \eta L_\gamma \|v_{t-1}\|_2 \cdot \max\left(\|\theta_{t-1} - \theta^*\|_2, \|\theta_{t-2} - \theta^*\|_2\right). \qquad \text{(D.18)}
\end{aligned}$$

Since $H^* = \nabla^2 F(\theta^*)$ we have

$$\begin{aligned}
&t\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2 \\
&= \left\|\mathbb{E}\left((t-1)\left(v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right) + \nabla F(\theta_{t-1}) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2 \\
&= \|\mathbb{E}\left((t-1)\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2}\right)\right. \\
&\quad + \left.\left.\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) + \nabla F(\theta_{t-1}) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\big\|_2 \\
&\le \left\|\mathbb{E}\left((t-1)\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2 \\
&\quad + \left\|\mathbb{E}\left((t-1)\left(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2. \qquad \text{(D.19)}
\end{aligned}$$

Further by rearranging the terms, and dividing both sides by $(t-1)$, we obtain

$$\begin{aligned}
&\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2 \\
&\le \left\|\mathbb{E}\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2}) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \left\|\mathbb{E}\left(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathscr{F}_{t-\widetilde{T}^*}\right\|_2 \\
&\le (1-\eta\mu)\left\|\mathbb{E}\left(v_{t-1} \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \eta L_\gamma \mathbb{E}\left(\|v_{t-1}\|_2 \cdot \max\left(\|\theta_{t-1} - \theta^*\|_2, \|\theta_{t-2} - \theta^*\|_2\right) \mid \mathscr{F}_{t-\widetilde{T}^*}\right),
\end{aligned}$$

where in the last inequality we apply the result in Eq. (D.18). Next by calculating the second moment of both the RHS and the LHS of the above quantity and the Hölder's inequality, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} \\
&\le (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}(v_{t-1} \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} + \eta L_\gamma \sqrt{\mathbb{E}\left\|\mathbb{E}\left(\|v_{t-1}\|_2^2 \cdot \left(\|\theta_{t-1} - \theta^*\|_2^2 + \|\theta_{t-2} - \theta^*\|_2^2\right) \mid \mathscr{F}_{t-\widetilde{T}^*}\right)\right\|_2^2} \\
&\le (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}(v_{t-1} \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} + \eta L_1 \left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/4}\left\{\left(\mathbb{E}\|\theta_{t-1} - \theta^*\|_2^4\right)^{1/4} + \left(\mathbb{E}\|\theta_{t-2} - \theta^*\|_2^4\right)^{1/4}\right\}.
\end{aligned}$$

Recursively applying the above inequality from $t - \widetilde{T}^*$ to t and we have that

$$
\begin{aligned}
&\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} \\
&\le (1-\eta\mu)^{\widetilde{T}^*}\mathbb{E}\left\|v_{t-\widetilde{T}^*}\right\|_2^2 + \frac{L_\gamma}{\mu}\max_{t-\widetilde{T}^*\le s\le t}\left(\mathbb{E}\left\|v_s\right\|_2^4\right)^{1/4}\cdot\max_{t-\widetilde{T}^*\le s\le t}\left(\mathbb{E}\left\|\theta_{t-2}-\theta^*\right\|_2^4\right)^{1/4}.
\end{aligned}
\tag{D.20}
$$

We recall from Lemmas D.5 and D.6 the following

$$
\begin{aligned}
\left(\mathbb{E}\left\|v_T\right\|_2^4\right)^{1/2} &\le C\left(\frac{\widetilde{\sigma_*}^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right), \qquad \text{and} \\
\left(\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4\right)^{1/2} &\le C\left(\frac{\widetilde{\sigma_*}^2}{T} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2}\right).
\end{aligned}
$$

Bringing this into Eq. (D.20) and we have that

$$
\begin{aligned}
\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} &\le (1-\eta\mu)^{\widetilde{T}^*}\mathbb{E}\left\|v_{t-\widetilde{T}^*}\right\|_2^2 \\
&+ \frac{cL_\gamma}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu}(t-\widetilde{T}^*)} + \frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2(t-\widetilde{T}^*)^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t-\widetilde{T}^*}} + \frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2(t-\widetilde{T}^*)}\right).
\end{aligned}
$$

Substituting with the window size $\widetilde{T}^*$ defined in Eq (D.13), the above inequality reduces as follows:

$$
\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\widetilde{T}^*})\right\|_2^2} \le \frac{cL_\gamma}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu}t} + \frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}} + \frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2 t}\right).
$$

D.1.2 Proof of Theorem 4.2

Utilizing the intermediate Proposition 1 in §D.1.1, we now aim to improve the dependency on initialization and turn to the proof of our multi-epoch nonasymptotic result. Invoking Eq. (D.11) in Lemma D.5, we obtain for $b = 1, 2, \cdots, B$ the bound for $T^\flat \ge cT_0$:

$$
\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2 &\le \frac{1}{e^2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat}, \qquad \text{and} \\
\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^4} &\le \frac{1}{e^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat},
\end{aligned}
$$

where our setting of $T^\flat$ gives a discount factor of $1/e^2$. Solving the recursion, we arrive at the bound:

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^2 \leq \frac{c\sigma_*^2}{T^\flat} + e^{-2B}\left\|\nabla F(\theta_0)\right\|_2^2, && \text{and}\\
&\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4} \leq \frac{c\widetilde{\sigma_*}^2}{T^\flat} + e^{-2B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}.
\end{aligned}$$

Our take is $B \geq \log \frac{T^\flat\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}}{c\sigma_*^2}$ such that $e^{-2B}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 \leq \frac{\sigma_*^2}{T^\flat}$ and $e^{-2B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} \leq \frac{\widetilde{\sigma_*}^2}{T^\flat}$ both hold. Finally, we have

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^2 \leq e^{-2B}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat} \leq \frac{c'\sigma_*^2}{T^\flat}, && \text{and}\\
&\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4} \leq e^{-2B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat} \leq \frac{c'\widetilde{\sigma_*}^2}{T^\flat}, && \text{and}\\
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^3 \leq \left(\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4\right)^{\frac{3}{4}} \leq \frac{c'\widetilde{\sigma_*}^3}{(T^\flat)^{3/2}}, && \text{and}\\
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2 \leq \frac{c\sigma_*}{(T^\flat)^{1/2}},
\end{aligned}$$

where constants c, c' change from line to line. Substituting this initial condition into the bound (D.1), we obtain the final bound:

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_T^{(B+1)})\right\|_2^2 - \frac{\sigma_*^2}{T}\\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2}\\
&\quad + C\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{\ell_\Xi\sigma_*\|\nabla F(\theta_0)\|_2}{\eta\mu^2T^2}\right) + \frac{CL_\gamma\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + C\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{L_\gamma\widetilde{\sigma_*}^3}{\eta^2\mu^4T^{7/2}}\right)\\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2},
\end{aligned}$$

which proves the bound (4.16).

Finally, substituting T by the final epoch length $n - BT^\flat$ and adopt similar reasoning as the previous one, we arrive at the conclusion:

$$\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{n} \leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu n^{1/2}} + \frac{\log n}{\eta\mu n} + \frac{\log\left(\frac{n}{T_0}\right)\ell_\Xi^2}{\mu^2 n}\right)\frac{\sigma_*^2}{n} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}n^2},$$

which proves the bound (4.16). Plugging η as given by $\eta = \frac{c}{\ell_\Xi n^{1/2}} \wedge \frac{1}{4L}$ with $c = 0.49$, we have

$$
\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 &\le C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu n^{1/2}} + \frac{\log n}{\eta\mu n} + \frac{\log\left(\frac{n}{T_0}\right)\ell_\Xi^2}{\mu^2 n}\right)\frac{\sigma_*^2}{n} + \frac{CL_\gamma\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}n^2} \\
&\le C\left(\log\left(\frac{en}{T_0}\right)\left(\frac{\ell_\Xi}{\mu\sqrt{n}}\right) + \frac{L}{\mu n}\right)\frac{\sigma_*^2}{n} + \frac{C(\ell_\Xi^{1/2}n^{1/4} + L^{1/2})L_\gamma\widetilde{\sigma_*}^3}{\mu^{5/2}n^2}.
\end{aligned}
$$

This concludes (4.17) and hence Theorem 4.2.

D.1.3 Proof of Corollary 4.3

The proof consists of two parts: bounds on the mean-squared error $\mathbb{E}\,\|\theta_T - \theta^*\|_2^2$ and bounds on the expected objective gap $\mathbb{E}\big[F(\theta_T) - F(\theta^*)\big]$. Two technical lemmas are needed in the proofs for both cases.

The first lemma is analogous to Lemma D.2, which provides a sharp bound on Gz_t for any matrix $G \in \mathbb{R}^{d\times d}$.

Lemma D.7. *Under settings of Theorem 4.1, for any matrix $G \in \mathbb{R}^{d\times d}$, the following bounds hold true for $T \ge T_0 + 1$:*

$$
\begin{aligned}
\mathbb{E}\,\|Gz_T\|_2^2 \le \frac{1}{T}\operatorname{Tr}\left(G\Sigma^* G^\top\right) + c|||G|||_{op}^2\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} \\
+ c|||G|||_{op}^2\left\{\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\,\|\nabla F(\theta_0)\|_2 + \frac{{T_0}^2}{T^2}\,\|\nabla F(\theta_0)\|_2^2\right\}. \quad \text{(D.21)}
\end{aligned}
$$

for some universal constant $c > 0$.

The second lemma is analogous to Lemma D.3, and provides sharp bound on the cross term $\mathbb{E}\langle Gz_t, Gv_t\rangle$.

Lemma D.8. *Under settings of Theorem 4.1, we have the following bound for any $T \ge T_0 + 1$:*

$$
\begin{aligned}
\left|\mathbb{E}\big\langle Gv_T, Gz_T\big\rangle\right| \le c|||G|||_{op}^2\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 T^4}\right)\log T \\
+ c|||G|||_{op}^2 L_1\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right), \quad \text{(D.22)}
\end{aligned}
$$

for some universal constant $c > 0$.

See §D.3.4 for the proof of both lemmas.

Taking these two lemmas as given, we now proceed with the proof of Corollary 4.3.

D.1.3.1 Proof of the MSE bound (4.19a)

We start with the following decomposition of the gradient:

$$\nabla F(\theta_T) = \int_0^1 \nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big)(\theta_T - \theta^*)d\rho,$$

which leads to the following bound under Assumption 4:

$$\begin{aligned}\big\|(H^*)^{-1}\nabla F(\theta_T) - (\theta_T - \theta^*)\big\|_2 &\le \int_0^1 \big\|(H^*)^{-1}\left(\nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\big\|_2 d\rho \\ &\le \frac{L_1}{\lambda_{\min}(H^*)}\|\theta_T - \theta^*\|_2^2 \le \frac{L_1}{\lambda_{\min}(H^*)\mu^2}\|\nabla F(\theta_T)\|_2^2. \end{aligned} \tag{D.23}$$

We can then upper bound the mean-squared error using the processes $(z_t)_{t\ge T_0}$ and $(v_t)_{t\ge T_0}$:

$$\begin{aligned}\mathbb{E}\|\theta_T - \theta^*\|_2^2 &\le \mathbb{E}\left(\big\|(H^*)^{-1}\nabla F(\theta_T)\big\|_2 + \frac{L_1}{\mu^2\lambda_{\min}(H^*)}\|\nabla F(\theta_T)\|_2^2\right)^2 \\ &\le \mathbb{E}\big\|(H^*)^{-1}\big(v_{T+1} - z_{T+1}\big)\big\|_2^2 + \frac{2L_1}{\lambda_{\min}(H^*)^2\mu^2}\mathbb{E}\|\nabla F(\theta_T)\|_2^3 \\ &\quad + \frac{L_1^2}{\lambda_{\min}(H^*)^2\mu^4}\mathbb{E}\|\nabla F(\theta_T)\|_2^4. \end{aligned} \tag{D.24}$$

The first term in the bound (D.24) admits the following decomposition:

$$\begin{aligned}&\mathbb{E}\big\|(H^*)^{-1}\big(z_{T+1} - v_{T+1}\big)\big\|_2^2 \\ &= \mathbb{E}\big\|(H^*)^{-1}z_{T+1}\big\|_2^2 + \mathbb{E}\big\|(H^*)^{-1}v_{T+1}\big\|_2^2 - 2\mathbb{E}\langle (H^*)^{-1}z_{T+1}, (H^*)^{-1}v_{T+1}\rangle.\end{aligned}$$

Note that the re-starting scheme in Algorithm 3 gives the initial conditions:

$$\mathbb{E}\|\nabla F(\theta_0)\|_2^2 \le \frac{c\sigma_*^2}{T_0}, \qquad \text{and} \qquad \left(\mathbb{E}\|\nabla F(\theta_0)\|_2^4\right)^{1/2} \le \frac{c\widetilde{\sigma}_*^2}{T_0}. \tag{D.25}$$

Using these initial conditions, and invoking the Lemma D.7 with test matrix $G = (H^*)^{-1}$, we obtain the bound:

$$\mathbb{E}\left\|(H^*)^{-1}z_T\right\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right) + \frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{T_0}{T}\right\}\frac{\sigma_*^2\log T}{T}.$$

Similarly, invoking Lemma D.8 with test matrix $G = (H^*)^{-1}$, we have that:

$$\left|\mathbb{E}\left\langle (H^*)^{-1}v_T, (H^*)^{-1}z_T\right\rangle\right| \leq \frac{c\sigma_*^2\sqrt{\log T}}{\lambda_{\min}(H^*)^2\eta\mu T^2} + \frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2}.$$

For the term $\mathbb{E}\left\|(H^*)^{-1}v_T\right\|_2^2$, Lemma D.1 along with the initial condition yields:

$$\mathbb{E}\left\|(H^*)^{-1}v_T\right\|_2^2 \leq \frac{1}{\lambda_{\min}(H^*)^2}\mathbb{E}\left\|v_T\right\|_2^2 \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2\mu\eta T^2}.$$

Collecting above bounds, we conclude that

$$\begin{aligned}\mathbb{E}\left\|(H^*)^{-1}\nabla F(\theta_T)\right\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right) + \frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\mu\eta T}\right\}\frac{\sigma_*^2\log T}{T}\\ + \frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2}.\end{aligned} \tag{D.26}$$

In order to bound the last two terms of the decomposition (D.24), we recall from Lemma D.5 and the initial condition (D.25) that:

$$\left(\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4\right)^{1/2} \leq \frac{c\widetilde{\sigma_*}^2}{T}.$$

Combining with Eq. (D.26) and substituting into the decomposition (D.24), we conclude that:

$$\begin{aligned}\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right) + \frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\eta\mu T}\right\}\frac{\sigma_*^2\log T}{T}\\ + \frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\left\{\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2} + \frac{\widetilde{\sigma_*}^3}{T^{3/2}} + \frac{L_1}{\mu^2}\frac{\widetilde{\sigma_*}^4}{T^2}\right\}.\end{aligned}$$

Note in the last line, the second $O(T^{-3/2})$ term is always no smaller than the previous first term. Taking $T = n - BT^\flat$ with $n \geq 2BT^\flat$, some algebra then completes the proof of the desired bound.

D.1.3.2 Proof of the objective gap bound (4.19b)

Applying second-order Taylor expansion with integral remainder, for any $\theta \in \mathbb{R}^d$, we note the following identity.

$$F(\theta) = F(\theta^*) + \langle \theta - \theta^*, \nabla F(\theta^*) \rangle + (\theta - \theta^*)^\top \int_0^1 \nabla^2 F\big(\rho\theta + (1-\rho)\theta^*\big) d\rho \cdot (\theta - \theta^*).$$

Noting that $\nabla F(\theta^*) = 0$ and invoking assumption 4, we have that:

$$\begin{aligned}
&F(\theta) \\
&\leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + \|\theta - \theta^*\|_2 \cdot \int_0^1 \left|\!\left|\!\left| \nabla^2 F\big(\rho\theta + (1-\rho)\theta^*\big) - H^* \right|\!\right|\!\right|_{\mathrm{op}} d\rho \cdot \|\theta - \theta^*\|_2 \\
&\leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + L_1 \|\theta - \theta^*\|_2^3 .
\end{aligned} \tag{D.27}$$

Analogous to Eq. (D.23), we have the bound:

$$\begin{aligned}
&\left\| (H^*)^{1/2}(\theta_T - \theta^*) - (H^*)^{-1/2}\nabla F(\theta_T) \right\|_2 \\
&\leq \int_0^1 \left\| (H^*)^{-1/2} \left(\nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big) - H^*\right)(\theta_T - \theta^*) \right\|_2 d\rho \\
&\leq \frac{L_1}{\sqrt{\lambda_{\min}(H^*)}} \|\theta_T - \theta^*\|_2^2 \leq \frac{L_1}{\mu^2 \sqrt{\lambda_{\min}(H^*)}} \|\nabla F(\theta_T)\|_2^2 .
\end{aligned}$$

Denote the residual $q_t := (H^*)^{1/2}(\theta_t - \theta^*) - (H^*)^{-1/2}\nabla F(\theta_t)$. Substituting into the bound (D.27), we have that:

$$\begin{aligned}
&\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left\| (H^*)^{-1/2}\nabla F(\theta) + q_T \right\|_2^2 + L_1 \mathbb{E}\|\theta_T - \theta^*\|_2^3 \\
&\leq \frac{1}{2}\mathbb{E}\left\| (H^*)^{-1/2}\nabla F(\theta_T) \right\|_2^2 + \frac{1}{\sqrt{\lambda_{\min}(H^*)}}\mathbb{E}\Big[\|q_t\|_2 \cdot \|\nabla F(\theta_T)\|_2\Big] + \frac{1}{2}\mathbb{E}\|q_t\|_2^2 + \frac{L_1}{\mu^3}\mathbb{E}\|\nabla F(\theta_T)\|_2^3 .
\end{aligned} \tag{D.28}$$

For the first term, by applying Lemma D.8 and D.7 with $G = (H^*)^{-1/2}$, we can obtain the following bound analogous to Eq. (D.26):

$$\begin{aligned}
\mathbb{E}\left\| (H^*)^{-1/2}\nabla F(\theta_T) \right\|_2^2 \leq \frac{1}{2T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right) + \frac{c}{\lambda_{\min}(H^*)}\left\{ \frac{\ell_\Xi^2 \eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\mu\eta T} \right\} \frac{\sigma_*^2 \log T}{T} \\
+ \frac{cL_1}{\lambda_{\min}(H^*)\mu^2} \cdot \frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2} T^2} .
\end{aligned}$$

For the rest of the terms, we recall that Lemma D.5 with the initial condition (D.25) gives the bound $\big(\mathbb{E}\|\nabla F(\theta_T)\|_2^4\big)^{1/2} \leq \frac{c\widetilde{\sigma_*}^2}{T}$. Substituting into the decomposition (D.27), we obtain that:

$$\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \leq \frac{1}{2T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right) + \frac{c}{\lambda_{\min}(H^*)}\left\{ \frac{\ell_\Xi^2 \eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\mu\eta T} \right\} \frac{\sigma_*^2 \log T}{T}$$

$$+\frac{cL_1}{\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{\mu T^{3/2}}+\frac{L_1^2}{\mu^4}\cdot\frac{\widetilde{\sigma_*}^4}{\lambda_{\min}(H^*)T^2}.$$

Noting that $T = n - BT^\flat$ with $n \geq 2BT^\flat$, we completes the proof of the desired bound.

D.1.4 Proof of Theorem 4.4

Here we provide a two-step proof of Theorem 4.4. We continue to adopt the v_t—z_t decomposition as earlier used, and we proceed with the proof in two steps:

Step 1:

We first claim the following single-epoch result, Eq. (D.29), that under the setting of Theorem 4.4 along with $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$, the single-epoch estimator produced by Algorithm 2 with burn-in time $T_0 = \left\lceil \frac{24}{\eta\mu} \right\rceil$, as $T \to \infty$, $\eta \to 0$ such that $\eta T \to \infty$ satisfies the following convergence in probability:

$$\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*) \xrightarrow{p} 0. \tag{D.29}$$

Taking this as given, we now combine Eq. (D.29) with our multi-epoch design Algorithm 3 we can essentially assume without loss of generality that $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$. Under the current scaling condition, the final long epoch in Algorithm 3 will be triggered with length $T = n - T^\flat B$, and hence we apply Eq. (D.3) so for some $C \leq 56$ we have the initial condition holds: $\mathbb{E}\|\nabla F(\theta_0^{(\eta)})\|_2^2 \leq \frac{C\sigma_*^2}{T^\flat} = O(\eta\mu\sigma_*^2)$, so that as $\eta T \to \infty$,

$$T\mathbb{E}\|v_T\|_2^2 \leq O\left(\frac{\sigma_*^2}{\eta\mu T} + \frac{\eta\mu\sigma_*^2}{\eta^4\mu^4 T^3}\right) \to 0.$$

Therefore, $\sqrt{T}v_T \xrightarrow{p} 0$ holds.

Now to put together the pieces, note that $\frac{1}{T}\sum_{s=1}^{T}\varepsilon_s(\theta^*)$ is the average of i.i.d. random vectors of finite second moment. By standard CLT, we have

$$\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*) \xrightarrow{d} \mathcal{N}(0,\Sigma^*).$$

Consequently, replacing T by $n - T^\flat B$ we can apply Slutsky's rule of weak convergence and obtain the desired weak convergence: as $\eta \to 0$, $n \to \infty$ such that $\eta(n - T^\flat B) \to \infty$

$$\sqrt{T}\nabla F(\theta_{T-1}^{(\eta)}) = \sqrt{T}v_T - \sqrt{T}z_T$$
$$= \sqrt{T}v_T - \left(\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*)\right) - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*) \xrightarrow{d} \mathcal{N}(0,\Sigma^*).$$

Due to our additional scaling condition, we can further replace $T = n - T^\flat B$ by n, concluding Theorem 4.4.

Step 2:

We proceed to prove Eq. (D.29) with the extra initialization condition $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$. By Eqs. (D.3) and (D.4), we have for $T \geq T_0$ there exist constants $a_1, a_2, a_3 > 0$ independent of η, T but depends on the problem parameters $(\mu, L, \ell_\Xi, \sigma_*, \theta_0, \alpha)$, such that

$$\mathbb{E}\|z_T\|_2^2 \leq \frac{2a_2}{T},$$

and consequently, we have from Eq. (D.3) that

$$\mathbb{E}\|v_T\|_2^2 \leq \frac{752\sigma_*^2}{\eta\mu T^2} + \frac{69175}{\eta^4\mu^4T^4}\|\nabla F(\theta_0)\|_2^2 \leq \frac{a_1}{T}\left(\frac{1}{\eta T} + \frac{\eta}{\eta^4T^3}\right) \leq \frac{2a_1}{\eta T^2},$$

and hence

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq 2\left(\mathbb{E}\|v_T\|_2^2 + \mathbb{E}\|z_T\|_2^2\right) \leq \frac{4a_1}{\eta T^2} + \frac{4a_2}{T} \leq \frac{4a_3}{T}.$$

Note from the definition in Eq. (4.29)

$$tz_t = \varepsilon_t(\theta_{t-1}) + (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})).$$

By setting $A_t = (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)$, the process $Tz_T - \sum_{s=1}^{T}\varepsilon_s(\theta^*) = \sum_{s=1}^{T}A_s$ is a martingale. To conclude the bound (D.29), we only need to show the following relation as $T \to \infty$ and $\eta \to 0$:

$$\mathbb{E}\left\|\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*)\right\|_2^2 = \frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\|A_s\|^2 \to 0. \tag{D.30}$$

Since we have

$$\mathbb{E}\left\|\sum_{s=T_0+1}^{T}(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}))\right\|_2^2 = \sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\|_2^2$$
$$\leq \ell_\Xi^2\sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\|\theta_{s-1} - \theta_{s-2}\|_2^2 = \eta^2\ell_\Xi^2\sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\|v_{s-1}\|_2^2$$

$$\le \eta^2 \ell_\Xi^2 \sum_{s=T_0+1}^{T} (s-1)^2 \frac{2a_1}{\eta^4 (s-1)^4} \le \frac{2a_1 \ell_\Xi^2}{\eta^2 T_0}.$$

We note that

$$\begin{aligned}
\mathbb{E}\left\| \sum_{s=T_0+1}^{T} (\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)) \right\|_2^2 &= \sum_{s=T_0+1}^{T} \mathbb{E}\left\| \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*) \right\|_2^2 \\
&\le \ell_\Xi^2 \sum_{s=T_0+1}^{T} \mathbb{E}\left\| \theta_{s-1} - \theta^* \right\|_2^2 \le \frac{\ell_\Xi^2}{\mu^2} \cdot 4a_3 \log\left(\frac{T}{T_0}\right).
\end{aligned}$$

Therefore, combining this with $\mathbb{E}\|A_t\|_2^2 = \mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\right\|_2^2 \le 2\ell_\Xi^2 \eta^2 (t-1)^2 \mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2} \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2$ we have as $T \to \infty$, $\eta \to 0$:

$$\begin{aligned}
\frac{1}{T} \sum_{t=T_0+1}^{T} \mathbb{E}\|A_t\|_2^2 &\le \frac{1}{T} \sum_{t=T_0+1}^{T} \mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\right\|_2^2 \\
&\le 2\ell_\Xi^2 \eta^2 \cdot \frac{1}{T} \sum_{t=T_0+1}^{T} (t-1)^2 \mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2} \cdot \frac{1}{T} \sum_{t=T_0+1}^{T} \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 \\
&\le 2\ell_\Xi^2 \eta^2 \cdot \frac{1}{T} \sum_{t=T_0+1}^{T} (t-1)^2 \frac{2a_1}{\eta (t-1)^2} + \frac{2\ell_\Xi^2}{\mu^2} \cdot \frac{1}{T} \sum_{t=T_0+1}^{T} \frac{4a_3}{t} \\
&= 4a_1 \ell_\Xi^2 \eta + \frac{2\ell_\Xi^2}{\mu^2} \cdot \frac{4a_3 \log\left(\frac{T}{T_0}\right)}{T},
\end{aligned}$$

i.e. the limit (D.30) holds, which implies $\sqrt{T} z_T - \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \varepsilon_s(\theta^*) \xrightarrow{p} 0$, completing our proof of Eq. (D.29).

D.2 Asymptotic results for single-epoch fixed-step-size ROOT-SGD

In this section, we complement Theorem 4.4 in §4.3.3 and establish an additional asymptotic normality result for ROOT-SGD with large step-size. Notably, the covariance of such asymptotic distribution is the sum of the optimal Gaussian limit and a correction term depending on the step-size, which exactly corresponds to existing results on fine-grained CLT for linear stochastic approximation with fixed step-size [134].

First, in order to obtain asymptotic results for single-epoch constant-step-size ROOT-SGD, we impose the following slightly stronger assumptions on the smoothness of stochastic gradients and Hessians:

(CLT.A) For any $\theta \in \mathbb{R}^d$ we have

$$\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}\left\|(\nabla^2 f(\theta;\xi) - \nabla^2 f(\theta^*;\xi))v\right\|_2^2 \le \beta^2 \|\theta - \theta^*\|_2^2. \tag{D.31a}$$

(CLT.B) The fourth moments of the stochastic gradient vectors at θ^* exist, and in particular we have

$$\mathbb{E}\|\nabla f(\theta^*;\xi)\|_2^4 < \infty, \qquad \text{and} \quad \ell'_{\Xi} := \sup_{v \in \mathbb{S}^{d-1}} \left(\mathbb{E}\left\|\nabla^2 f(\theta^*;\xi)v\right\|_2^4\right)^{1/4} < \infty. \tag{D.31b}$$

Note that both conditions are imposed solely at the optimal point θ^*; we do not impose globally uniform bounds in $\mathbb{R}^d$.

Defining the random matrix $\Xi(\theta) := \nabla^2 f(\theta;\xi) - \nabla^2 F(\theta)$ for any $\theta \in \mathbb{R}^d$, we consider the following matrix equation (a.k.a. *modified Lyapunov equation*):

$$\Lambda H^* + H^*\Lambda - \eta\mathbb{E}\left[\Xi(\theta^*)\Lambda\Xi(\theta^*)\right] - \eta H^*\Lambda H^* = \eta\Sigma^*. \tag{D.32}$$

in the symmetric matrix Λ. It can be shown that under the given assumptions, this equation has a unique solution—denoted Λ_η—which plays a key role in the following theorem.

Proposition 2 (Asymptotic efficiency, single-epoch ROOT-SGD) *Suppose that Assumptions 1, 2, and 3 are satisfied, as well as* (CLT.A) *and* (CLT.B). *Then there exist constants c_1, c_2, given the step-size $\eta \in \left(0, c_1\left(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L} \wedge \frac{\mu^{1/3}}{\ell_\Xi'^{4/3}}\right)\right)$, and burn-in time $T_0 = \frac{c_2}{\mu\eta}$, we have*

$$\sqrt{T}(\theta_T - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\left(\Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right)(H^*)^{-1}\right). \tag{D.33}$$

See §D.2.1 for the proof of this theorem.

A few remarks are in order. First, we observe that the asymptotic covariance in Eq. (D.33) is the sum of the matrix $(H^*)^{-1}\Sigma^*(H^*)^{-1}$ and an additional correction term defined in Eq. (D.32). The asymptotic covariance of $(H^*)^{-1}\Sigma^*(H^*)^{-1}$ matches the standard Cramér-Rao lower bound in the asymptotic statistics literature [175, 174] and matches the optimal rates achieved in the theory of stochastic approximation [101, 150, 155]. The correction term is of the same form as that of the constant-step-size linear stochastic approximation of the Polyak-Ruppert-Juditsky averaging prcedure as derived in [134], while our Proposition 2 is applicable to more general nonlinear stochastic problems. For instance in our setting, the correction terms tends to zero as the (constant) step-size decreases to zero, which along with a trace bound leads to the following asymptotics as $T \to \infty$ (see [134]):

$$T\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \sim \mathrm{Tr}\left(\Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right) \le \left(1 + \frac{\ell_\Xi^2\eta}{\mu}\right)\sigma_*^2.$$

The message conveyed by the last display is consistent with the leading two terms in our earlier nonasymptotic bound Eq. (D.1) in Proposition 1, and thanks to our additional smoothness assumptions (CLT.A) and (CLT.B) we are able to characterize this correction term in a more fine-grained fashion as in the asymptotic covariance of Eq. (D.33). Second, we note that Proposition 2 has an additional requirement on the step-size, needing it to be upper bounded by $\frac{\mu^{1/3}}{\ell_{\Xi}'^{4/3}}$. This is a mild requirement on the step-size. In particular, for applications where the noises are light-tailed, ℓ_{Ξ}' and ℓ_{Ξ} are of the same order, and the additional requirement $\eta < \frac{c\mu^{1/3}}{\ell_{\Xi}'^{4/3}}$ is usually weaker than the condition $\eta < \frac{c\mu}{\ell_{\Xi}^2}$ needed in the previous section.

D.2.1 Proof of Proposition 2

Denote $H_t(\theta) := \nabla^2 f(\theta;\xi_t)$ and $\Xi_t(\theta) := H_t(\theta) - \nabla^2 F(\theta)$. Intuitively, since the sequence θ_t is converging to θ^* at a $1/\sqrt{t}$ rate, replacing θ_{s-1} with θ^* will only lead to a small change in the sum. For the martingale Ψ_t, each term can be written as:

$$t(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) = t\int_0^1 \Xi_t\left(\rho\theta_{t-2} + (1-\rho)\theta_{t-1}\right)(\theta_{t-1} - \theta_{t-2})d\rho.$$

By Assumption (CLT.A), this quantity should approach $\eta\Xi_t(\theta^*)\cdot(tv_{t-1})$. If we can show the convergence of the sequence $\{tv_t\}_{t\geq T_0}$ to a stationary distribution, then the asymptotic result follows from the Birkhoff ergodic theorem and a martingale CLT. While the process $\{tv_t\}_{t\geq T_0}$ is not Markovian, we show that it can be well-approximated by a time-homogeneous Markov process that we construct in the proof.

In particular, consider the auxiliary process $\{y_t\}_{t\geq T_0}$, initialized as $y_{T_0} = T_0 v_{T_0}$ and updated as

$$y_t = y_{t-1} - \eta H_t(\theta^*)y_{t-1} + \varepsilon_t(\theta^*), \quad \text{for all } t \geq T_0 + 1. \tag{D.34}$$

Note that $\{y_t\}_{t\geq T_0}$ is a time-homogeneous Markov process that is coupled to $\{(\theta_t, v_t, z_t)\}_{t\geq T_0}$. We have the following coupling estimate:

Lemma D.9. *Supposing that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold, then for any iteration $t \geq T_0$ and any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{8\ell_{\Xi}^2})$, we have*

$$\mathbb{E}\,\|tv_t - y_t\|_2^2 \leq \frac{c_0}{\sqrt{t}},$$

for a constant c_0 depending on the smoothness and strong convexity parameters $L, \ell_{\Xi}, \mu, \beta$ and the step-size η, but independent of t.

See §D.4.1 for the proof of this lemma.

We also need the following lemma, which provides a convenient bound on the difference $H_t(\theta) - H_t(\theta^*)$ for a vector θ chosen in the data-dependent way.

Lemma D.10. *Suppose that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq T_0$, any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{8\ell_{\Xi}^2})$ and for any random vector $\widetilde{\theta}_{t-1} \in \mathscr{F}_{t-1}$, we have*

$$\mathbb{E}\left\| \left[H_t(\widetilde{\theta}_{t-1}) - H_t(\theta^*)\right] y_{t-1} \right\|_2^2 \leq c_1 \sqrt{\mathbb{E}\left\| \widetilde{\theta}_{t-1} - \theta^* \right\|_2^2},$$

where c_1 is a constant independent of t and the choice of $\widetilde{\theta}_{t-1}$.

See §D.4.2 for the proof of this lemma.

Finally, the following lemma characterizes the behavior of the process $\{y_t\}_{t \geq T_0}$ defined in Eq. (D.34):

Lemma D.11. *Suppose that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq T_0$ and any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{16\ell_{\Xi}^2} \wedge \frac{\mu^{1/3}}{6\ell_{\Xi}^{4/3}})$, we have*

$$\mathbb{E}(y_t) = 0 \quad \text{for all } t \geq T_0, \qquad \text{and} \quad \sup_{t \geq T_0} \mathbb{E}\, \|y_t\|_2^4 < a',$$

for a constant $a' > 0$, which is independent of t. Furthermore, the process $\{y_t\}_{t \geq 0}$ has a stationary distribution with finite second moment, and a stationary covariance Q_η that satisfies the equation

$$H^* Q_\eta + Q_\eta H^* - \eta\,[H^* Q_\eta H^* + \mathbb{E}(\Xi(\theta^*) Q_\eta \Xi(\theta^*))] = \frac{1}{\eta}\Sigma^*.$$

See §D.4.3 for the proof of this lemma.

Taking these three lemmas as given, we now proceed with the proof of Proposition 2. We first define two auxiliary processes:

$$N_T := \sum_{t=T_0+1}^{T} \varepsilon_t(\theta^*), \qquad \Upsilon_T := \eta \sum_{t=T_0+1}^{T} \Xi_t(\theta^*) y_{t-1}.$$

Observe that both N_T and Υ_T are martingales adapted to $(\mathscr{F}_t)_{t \geq T_0}$. In the following, we first bound the differences $\|M_T - N_T\|_2$ and $\|\Psi_T - \Upsilon_T\|_2$, respectively, and then show the limiting distribution results for $N_T + \Upsilon_T$.

By Theorem 4.1, define $a_0 := \frac{28\sigma_*^2}{\mu^2} + \frac{2700}{\eta^2 \mu^4 T_0} \|\nabla F(\theta_0)\|_2^2$, we have

$$\mathbb{E}\|\theta_t-\theta^*\|_2^2 \le \frac{1}{\mu^2}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \le \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^4(t+1)^2}+\frac{28\ \sigma_*^2}{\mu^2(t+1)} \le \frac{a_0}{t+1}, \quad \text{for all } t \ge T_0. \tag{D.35}$$

Applying the bound (D.35) with Assumption 3, we have

$$\mathbb{E}\|M_T-N_T\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\|_2^2 \le \ell_\Xi^2\sum_{t=T_0+1}^{T}\mathbb{E}\|\theta_{t-1}-\theta^*\|_2^2 \le a_0\ell_\Xi^2\log T. \tag{D.36}$$

For the process Υ_T, by the Cauchy-Schwartz inequality, we have

$$\mathbb{E}\|\Psi_T-\Upsilon_T\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\|\eta\,\Xi_t(\theta^*)y_{t-1}-(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))\|_2^2$$
$$\le \eta^2\sum_{t=T_0+1}^{T}\mathbb{E}\int_0^1\|\Xi_t(\theta^*)y_{t-1}-\Xi_t(\rho\theta_{t-1}+(1-\rho)\theta_{t-2};\xi_t)(t-1)v_{t-1}\|_2^2\,d\rho \le I_I+I_2,$$

where we define

$$I_1 := 2\eta^2\sum_{t=T_0+1}^{T}\mathbb{E}\int_0^1\|(\Xi_t(\theta^*)-\Xi_t(\rho\theta_{t-1}+(1-\rho)\theta_{t-2}))y_{t-1}\|_2^2\,d\rho, \quad \text{and}$$
$$I_2 := 2\eta^2\sum_{t=T_0+1}^{T}\mathbb{E}\int_0^1\|\Xi_t(\rho\theta_{t-1}+(1-\rho)\theta_{t-2})(y_{t-1}-(t-1)v_{t-1})\|_2^2\,d\rho.$$

We bound each of these two terms in succession.

Bound on I_1:

In order to bound the term I_1, we apply Lemma D.10 with the choice

$$\widetilde{\theta}_{t-1} = \rho\theta_{t-1}+(1-\rho)\theta_{t-2} \in \mathscr{F}_{t-1},$$

so as to obtain

$$\mathbb{E}\left\|\left(H_t(\widetilde{\theta}_{t-1})-H_t(\theta^*)\right)y_{t-1}\right\|_2^2 \le c_1\sqrt{\mathbb{E}\left\|\widetilde{\theta}_t-\theta^*\right\|_2^2}.$$

Applying the Cauchy-Schwartz inequality yields

$$\mathbb{E}\left\|\left(\nabla^2F(\widetilde{\theta}_{t-1})-\nabla^2F(\theta^*)\right)y_{t-1}\right\|_2^2 \le \mathbb{E}\left\|\left(H_t(\widetilde{\theta}_{t-1})-H_t(\theta^*)\right)y_{t-1}\right\|_2^2 \le c_1\sqrt{\mathbb{E}\left\|\widetilde{\theta}_t-\theta^*\right\|_2^2}.$$

Putting the two bounds together, we obtain:

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\Xi_t(\widetilde{\theta}_{t-1})-\Xi_t(\theta^*)\right)y_{t-1}\right\|_2^2\\
&\le 2\mathbb{E}\left\|\left(H_t(\widetilde{\theta}_{t-1})-H_t(\theta^*)\right)y_{t-1}\right\|_2^2+2\mathbb{E}\left\|\left(\nabla^2F(\widetilde{\theta}_{t-1})-\nabla^2F(\theta^*)\right)y_{t-1}\right\|_2^2\\
&\le 4c_1\sqrt{\mathbb{E}\left\|\widetilde{\theta}_t-\theta^*\right\|_2^2}.
\end{aligned}
$$

Thus, we find that

$$
\begin{aligned}
&\mathbb{E}\,\|(\Xi_t(\theta^*)-\Xi_t(\rho\theta_{t-1}+(1-\rho)\theta_{t-2}))y_{t-1}\|_2^2\le 4c_1\sqrt{\mathbb{E}\,\|\rho\theta_{t-1}+(1-\rho)\theta_{t-2}-\theta^*\|_2^2}\\
&\le 4c_1\left(\sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta^*\|_2^2}+\sqrt{\mathbb{E}\,\|\theta_{t-2}-\theta^*\|_2^2}\right)\le 4c_1\sqrt{a_0}\left(\frac{1}{\sqrt{t-1}}+\frac{1}{\sqrt{t-2}}\right)\le\frac{16c_1\sqrt{a_0}}{\sqrt{t}},
\end{aligned}
$$

where in the last step we used the inequality (D.35). Summing over t from T_0+1 to T yields the bound

$$
I_1\le 2\eta^2\sum_{t=T_0+1}^{T}\frac{16c_1\sqrt{a_0}}{\sqrt{t}}\le 64\eta^2c_1\sqrt{a_0T}.
$$

Bound on I_2:

Turning to the term I_2, by Assumption 3 and Lemma D.9, we note that:

$$
I_2\le 2\eta^2\sum_{t=T_0+1}^{T}\ell_\Xi^2\mathbb{E}\,\|y_{t-1}-(t-1)v_{t-1}\|_2^2\le 2\eta^2\ell_\Xi^2\sum_{t=T_0+1}^{T}\frac{c_0}{\sqrt{t}}\le 4\eta^2\ell_\Xi^2c_0\sqrt{T}.
$$

Putting these inequalities together, we conclude that:

$$
\mathbb{E}\,\|\Psi_T-\Upsilon_T\|_2^2\le(64\eta^2c_1\sqrt{a_0}+4\eta^2\ell_\Xi^2c_0)\sqrt{T}. \tag{D.37}
$$

Now we have the estimates for the quantities $\|\Psi_T-\Upsilon_T\|_2$ and $\|M_T-N_T\|_2$. In the following, we first prove the CLT for $N_T+\Upsilon_T$, and then use the error bounds to establish CLT for $M_T+\Psi_T$, which ultimately implies the desired limiting result for $\sqrt{T}(\theta_T-\theta^*)$

Define $\nu_t:=\varepsilon_t(\theta^*)+\eta\,\Xi_t(\theta^*)y_{t-1}$. By definition, $N_T+\Upsilon_T=\sum_{t=T_0}^{T}\nu_t$, and we have

$$
\begin{aligned}
\mathbb{E}(\nu_t\nu_t^\top)=\mathbb{E}(\varepsilon_t(\theta^*)\varepsilon_t(\theta^*)^\top)+\mathbb{E}\left(\Xi_t(\theta^*)y_{t-1}y_{t-1}^\top\Xi_t(\theta^*)^\top\right)\\
+\mathbb{E}\left(\varepsilon_t(\theta^*)y_{t-1}^\top\Xi_t(\theta^*)^\top\right)+\mathbb{E}\left((\Xi_t(\theta^*)y_{t-1}\varepsilon_t(\theta^*)^\top\right).
\end{aligned}
$$

For the first term, we have $\mathbb{E}(\varepsilon_t(\theta^*)\varepsilon_t(\theta^*)^\top)=\Sigma^*$ by definition.

For the second term, according to Lemma D.11, we note that the time-homogeneous Markov process $\{y_t\}_{t\ge T_0}$ converges asymptotically to a stationary distribution with covariance Q_η. Invoking the Birkhoff ergodic theorem, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\left(\Xi_t(\theta^*)y_{t-1}y_{t-1}^\top\Xi_t(\theta^*)^\top \mid \mathscr{F}_{t-1}\right) &= \mathbb{E}(\Xi(\theta^*)\otimes\Xi(\theta^*))\left[\frac{1}{T}\sum_{t=T_0+1}^{T}y_{t-1}y_{t-1}^\top\right]\\ &\xrightarrow{p}\mathbb{E}\left(\Xi(\theta^*)Q_\eta\Xi(\theta^*)^\top\right).\end{aligned}$$

For the cross terms, we note that:

$$\mathbb{E}\left(\varepsilon_t(\theta^*)y_{t-1}^\top\Xi_t(\theta^*)^\top \mid \mathscr{F}_{t-1}\right) = \mathbb{E}(\varepsilon(\theta^*)\otimes\Xi(\theta^*))[y_{t-1}].$$

Note that by Lemma D.11, we have $\mathbb{E}(y_t)=0$ for any $t\geq T_0$. By the weak law of large numbers, we have $\frac{1}{T}\sum_{t=T_0+1}^{T}y_t\xrightarrow{p}0$. Putting together these inequalities, we find that

$$\begin{aligned}\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\left(v_tv_t^\top \mid \mathscr{F}_{t-1}\right) = \frac{1}{T}\sum_{t=T_0+1}^{T}\Big(&\Sigma^*+\eta^2\mathbb{E}(\Xi(\theta^*)\otimes\Xi(\theta^*))[y_ty_t^\top]\\ &+\eta\mathbb{E}(\varepsilon(\theta^*)\otimes\Xi(\theta^*))[y_{t-1}]+\eta\mathbb{E}(\Xi(\theta^*)\otimes\varepsilon(\theta^*))[y_{t-1}]\Big),\end{aligned}$$

and hence the random matrix $\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\left(v_tv_t^\top \mid \mathscr{F}_{t-1}\right)$ converges in probability to the matrix

$$\Sigma^*+\mathbb{E}\left(\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)^\top\right).$$

To prove the limiting distribution result, we use standard martingale CLT (c.f. Corollary 3.1 in [82]). It remains to verify the conditional Lindeberg condition. Indeed, for any $\varepsilon>0$, a straightforward calculation yields:

$$\begin{aligned}&R_T(\varepsilon):=\sum_{t=T_0+1}^{T}\mathbb{E}\left(\left\|\frac{v_t}{\sqrt{T}}\right\|_2^2\mathbf{1}_{\left\|\frac{v_t}{\sqrt{T}}\right\|_2>\varepsilon}\middle|\mathscr{F}_{t-1}\right)\\ &\overset{(i)}{\leq}\frac{1}{T}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\left(\|v_t\|_2^4\mid\mathscr{F}_{t-1}\right)}\cdot\sqrt{\mathbb{P}\left(\|v_t\|_2>\varepsilon\sqrt{T}\mid\mathscr{F}_{t-1}\right)}\overset{(ii)}{\leq}\frac{1}{T}\sum_{t=T_0+1}^{T}\frac{1}{(\varepsilon\sqrt{T})^2}\mathbb{E}\left(\|v_t\|_2^4\mid\mathscr{F}_{t-1}\right).\end{aligned}$$

In step (i), we use the Cauchy-Schwartz inequality, and in step (ii), we use the Markov inequality to bound the conditional probability.

Using the condition (CLT.B) and Young's inequality, we note that:

$$\mathbb{E}\left(\|v_t\|_2^4\mid\mathscr{F}_{t-1}\right)\leq 8\mathbb{E}\|\varepsilon(\theta^*)\|_2^4+8\ell_\Xi^4\|y_{t-1}\|_2^4.$$

Plugging back to the upper bound for $R_T(\varepsilon)$, and applying Lemma D.11, as $T\to\infty$, we have

$$\mathbb{E}[R_T(\varepsilon)] \leq \frac{8}{T\varepsilon^2}\mathbb{E}\,\|\varepsilon(\theta^*)\|_2^4 + \frac{8\ell_\Xi^4}{T^2\varepsilon^2}\sum_{t=T_0+1}^{T}\mathbb{E}\,\|y_{t-1}\|_2^4 \leq \frac{8}{T\varepsilon^2}\mathbb{E}\,\|\varepsilon(\theta^*)\|_2^4 + \frac{8\ell_\Xi^4}{T\varepsilon^2}a' \to 0.$$

Note that $R_T(\varepsilon) \geq 0$ by definition. The limit statement implies that $R_T(\varepsilon) \xrightarrow{p} 0$, for any $\varepsilon > 0$. Therefore, the conditional Lindeberg condition holds true, and we have the CLT:

$$\frac{N_T + \Upsilon_T}{\sqrt{T}} \xrightarrow{d} \mathcal{N}\left(0, \Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)\right]\right).$$

By the second-moment estimates (D.36) and (D.37), we have

$$\frac{\|\Upsilon_T - \Psi_T\|_2}{\sqrt{T}} \xrightarrow{p} 0, \qquad \frac{\|M_T - N_T\|_2}{\sqrt{T}} \xrightarrow{p} 0.$$

With the burn-in time T_0 fixed, we also have $\frac{T_0}{T} z_{T_0} \xrightarrow{p} 0$. By Slutsky's theorem, we have

$$\sqrt{T} z_T \xrightarrow{d} \mathcal{N}\left(0, \Sigma^* + \mathbb{E}\left(\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)^\top\right)\right).$$

Note that $\nabla F(\theta_{t-1}) = v_t - z_t$. By Lemma D.9 and Lemma D.11, we have

$$\mathbb{E}\,\|v_t\|_2^2 \leq \frac{2}{t^2}\mathbb{E}\,\|tv_t - y_t\|_2^2 + \frac{2}{t^2}\mathbb{E}\,\|y_t\|_2^2 \leq \frac{2}{t^2}\left(\sqrt{a'} + \frac{c_0}{\sqrt{t}}\right),$$

which implies that $\sqrt{t} v_t \xrightarrow{p} 0$. Recall that $z_t = v_t - \nabla F(\theta_{t-1})$. By Slutsky's theorem, we obtain:

$$\sqrt{T} \cdot \nabla F(\theta_T) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)\right]\right).$$

Finally, we note that for $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}\|\nabla F(\theta) - H^*(\theta - \theta^*)\|_2 &= \left\|\int_0^1 \nabla^2 F(\theta^* + \rho(\theta - \theta^*))(\theta - \theta^*)d\rho - H^*(\theta - \theta^*)\right\|_2 \\ &\leq \int_0^1 \left|\!\left|\!\left|\nabla^2 F(\theta^* + \rho(\theta - \theta^*)) - H^*\right|\!\right|\!\right|_{\mathrm{op}} \cdot \|\theta - \theta^*\|_2\, d\rho \\ &\leq \|\theta - \theta^*\|_2 \cdot \sup_{\|\theta' - \theta^*\|_2 \leq \|\theta - \theta^*\|_2} \left|\!\left|\!\left|\nabla^2 F(\theta') - H^*\right|\!\right|\!\right|_{\mathrm{op}}.\end{aligned}$$

By Assumption (CLT.A), we have

$$\forall v \in \mathbb{S}^{d-1},\ \theta \in \mathbb{R}^d \quad \left\|(\nabla^2 F(\theta) - \nabla^2 F(\theta^*))v\right\|_2^2 \leq \mathbb{E}\left\|(\nabla^2 f(\theta;\xi) - \nabla^2 f(\theta^*;\xi))v\right\|_2^2 \leq \beta^2\,\|\theta - \theta^*\|_2^2.$$

Consequently, we have the bound:

$$\|\nabla F(\theta) - H^*(\theta - \theta^*)\|_2 \leq \|\theta - \theta^*\|_2 \cdot \sup_{\|\theta' - \theta^*\|_2 \leq \|\theta - \theta^*\|_2} \sup_{v \in \mathbb{S}^{d-1}} \left\|(\nabla^2 F(\theta') - H^*)v\right\|_2 \leq \beta \|\theta - \theta^*\|_2^2 .$$

By Eq. (D.35), we have $\sqrt{T}\|\nabla F(\theta_T) - H^*(\theta_T - \theta^*)\|_2 \xrightarrow{p} 0$. Invoking Slutsky's theorem, this leads to $\sqrt{T}H^*(\theta_T - \theta^*) \xrightarrow{d} \mathscr{N}\left(0, \Sigma^* + \mathbb{E}\left(\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)^\top\right)\right)$, and consequently,

$$\sqrt{T}(\theta_T - \theta^*) \xrightarrow{d} \mathscr{N}\left(0, (H^*)^{-1}\left(\Sigma^* + \mathbb{E}[\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)^\top]\right)(H^*)^{-1}\right),$$

which finishes the proof.

D.3 Proofs of auxiliary lemmas in §D.1

D.3.1 Proof of Lemma D.5

Recall that we have the recursive update rule of z_t as

$$tz_t = (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}).$$

Taking fourth moments on both sides, we have

$$\begin{aligned}\mathbb{E}\|tz_t\|_2^4 &= \mathbb{E}\|(t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\ &= \mathbb{E}\|(t-1)z_{t-1}\|_2^4 + \mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\ &\quad + 4\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^3 \|(t-1)z_{t-1}\|_2 \\ &\quad + 6\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^2 \|(t-1)z_{t-1}\|_2^2, \end{aligned} \quad \text{(D.38)}$$

where one of the terms is zeroed out. By Hölder's inequality and Young's inequality, we bound the third term and the fourth term of the RHS as

$$\begin{aligned}&\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^3 \|(t-1)z_{t-1}\|_2 \\ &\leq \left(\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{3/4} \left(\mathbb{E}\|(t-1)z_{t-1}\|_2^4\right)^{1/4} \\ &\leq \frac{1}{2}\left(\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{1/2} \left(\mathbb{E}\|(t-1)z_{t-1}\|_2^4\right)^{1/2} \\ &\quad + \frac{1}{2}\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^2 \|(t-1)z_{t-1}\|_2^2 \\ &\leq \left(\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{1/2} \left(\mathbb{E}\|(t-1)z_{t-1}\|_2^4\right).\end{aligned}$$

Thus Eq. (D.38) continues as

$$
\begin{aligned}
\mathbb{E}\|tz_t\|_2^4 &\le \mathbb{E}\|(t-1)z_{t-1}+(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\|_2^4\\
&= \mathbb{E}\|(t-1)z_{t-1}\|_2^4+3\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\|_2^4\\
&\quad +8\left(\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\|_2^4\right)^{1/2}\left(\mathbb{E}\|(t-1)z_{t-1}\|_2^4\right)\\
&\le \left(\sqrt{\mathbb{E}\|(t-1)z_{t-1}\|_2^4}+4\sqrt{\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\|_2^4}\right)^2,
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\|_2^4\\
&\le 27(t-1)^4\mathbb{E}\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\|_2^4+27\mathbb{E}\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\|_2^4+27\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^4\\
&\le 27\ell_\Xi^4\eta^4(t-1)^4\mathbb{E}\|v_{t-1}\|_2^4+\frac{27\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4+27\widetilde{\sigma_*}^4.
\end{aligned}
$$

Then

$$
t^2\sqrt{\mathbb{E}\|z_t\|_2^4}\le\sqrt{\mathbb{E}\|(t-1)z_{t-1}\|_2^4}+12\sqrt{3}\ell_\Xi^2\eta^2\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4}+\frac{12\sqrt{3}\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+12\sqrt{3}\widetilde{\sigma_*}^2.
$$

Combining this with Eq. (D.45) in Lemma D.12 that

$$
\sqrt{\mathbb{E}\|tv_t\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4}+\frac{5}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+14\widetilde{\sigma_*}^2.
$$

By the choice of η satisfying $\eta\le\frac{1}{56L}\wedge\frac{\mu}{64\ell_\Xi^2}$, we have $\frac{\ell_\Xi^2}{\mu^2}\le\frac{1}{64\eta\mu}$ and

$$
t^2\sqrt{\mathbb{E}\|z_t\|_2^4}+t^2\sqrt{\mathbb{E}\|v_t\|_2^4}\le\sqrt{\mathbb{E}\|(t-1)z_{t-1}\|_2^4}+\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4}+\frac{6}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+35\widetilde{\sigma_*}^2.
$$

Recursively applying the above inequality and by observing that $\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}\le 2\sqrt{\mathbb{E}\|z_t\|_2^4}+2\sqrt{\mathbb{E}\|v_t\|_2^4}$, we have

$$
\begin{aligned}
&T^2\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4}\le 2T^2\sqrt{\mathbb{E}\|z_T\|_2^4}+2T^2\sqrt{\mathbb{E}\|v_T\|_2^4}\\
&\le 2\sqrt{\mathbb{E}\left\|T_0z_{T_0}\right\|_2^4}+2\sqrt{\mathbb{E}\left\|T_0v_{T_0}\right\|_2^4}+\frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+70(T-T_0)\widetilde{\sigma_*}^2.
\end{aligned}
\tag{D.39}
$$

Further for $T_0z_{T_0}$ and $T_0v_{T_0}$ we note that by applying Khintchine's inequality as well as Young's inequality we have

$$\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^4 = \mathbb{E}\left\|\sum_{t=1}^{T_0}\varepsilon_t(\theta_0)\right\|_2^4 \le \mathbb{E}\left(\sum_{t=1}^{T_0}\|\varepsilon_t(\theta_0)\|_2^2\right)^2 \le T_0\mathbb{E}\sum_{t=1}^{T_0}\|\varepsilon_t(\theta_0)\|_2^4 \le 8{T_0}^2\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_0)\|_2^4 + \widetilde{\sigma_*}^4\right), \tag{D.40}$$

and

$$\begin{aligned}\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4 &= \mathbb{E}\left\|T_0 z_{T_0}\right\|_2^4 + \mathbb{E}\|T_0\nabla F(\theta_0)\|_2^4 + 4\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^3\|T_0\nabla F(\theta_0)\|_2 + 6\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^2\|T_0\nabla F(\theta_0)\|_2^2 \\ &\le 7\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4 + 5\mathbb{E}\|T_0\nabla F(\theta_0)\|_2^4 \le 56{T_0}^2\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_0)\|_2^4 + \widetilde{\sigma_*}^4\right) + 5{T_0}^4\mathbb{E}\|\nabla F(\theta_0)\|_2^4.\end{aligned} \tag{D.41}$$

Taking squared root on Eq. (D.40) and (D.41) and recalling that $\eta \le \frac{\mu}{64\ell_\Xi^2}$, we have

$$\sqrt{\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^4} \le 2\sqrt{2}T_0\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + \widetilde{\sigma_*}^2\right), \tag{D.42}$$

and

$$\sqrt{\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4} \le (\sqrt{5} + 1/8){T_0}^2\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + 8T_0\widetilde{\sigma_*}^2. \tag{D.43}$$

Bringing Eq. (D.42) and (D.43) into Eq. (D.39), we arrive at the following:

$$\begin{aligned}T^2\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4} &\le 4\sqrt{2}T_0\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + \widetilde{\sigma_*}^2\right) + (2\sqrt{5} + 1/4){T_0}^2\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} \\ &\quad + 16T_0\widetilde{\sigma_*}^2 + \frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 70(T - T_0)\widetilde{\sigma_*}^2 \\ &\le 5{T_0}^2\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + \frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 70T\widetilde{\sigma_*}^2.\end{aligned} \tag{D.44}$$

Dividing both sides by T^2, summing up Eq. (D.44) from $T = T_0 + 1$ to $T^* \ge T_0 + 1$ and using the fact that $\eta \le \frac{\mu}{64\ell_\Xi^2}, T_0 \ge 2$, we have

$$\sum_{T=T_0+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4} \le 5T_0\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + \frac{12}{\eta\mu T_0}\sum_{t=T_0+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 70\widetilde{\sigma_*}^2\log\left(\frac{T^*}{T_0}\right).$$

Taking $T_0 = \left\lceil\frac{24}{\eta\mu}\right\rceil$, we have

$$\sum_{T=T_0+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4} \le 10T_0\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4} + 140\widetilde{\sigma_*}^2\log\left(\frac{T^*}{T_0}\right).$$

Again by Eq. (D.44), we have

$$
\begin{aligned}
&T^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^4}\\
&\le 5{T_0}^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+\frac{12}{\eta\mu}\left(10T_0\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+140\widetilde{\sigma_*}^2\log\left(\frac{T}{T_0}\right)\right)+70T\widetilde{\sigma_*}^2\\
&\le 10{T_0}^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+70T_0\widetilde{\sigma_*}^2\log\left(\frac{T}{T_0}\right)+70T\widetilde{\sigma_*}^2.
\end{aligned}
$$

Dividing both sides by T^2 we conclude that

$$
\begin{aligned}
\sqrt{\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^4}&\le\frac{10{T_0}^2}{T^2}\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+70\left(1+\frac{T_0}{T}\log\left(\frac{T}{T_0}\right)\right)\frac{\widetilde{\sigma_*}^2}{T}\\
&\le\frac{10{T_0}^2}{T^2}\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+\frac{140\widetilde{\sigma_*}^2}{T}.
\end{aligned}
$$

which finishes our proof of Lemma D.5.

D.3.2 Proof of Lemma D.6

Our main technical tools is the following lemma, which bound the fourth moment of the v_t recursion.

Lemma D.12. *Under the setting of Proposition 1, when $\eta \le \frac{1}{56L}\wedge\frac{\mu}{64\ell_\Xi^2}$, we have the following bound for $t \ge T_0+1$*

$$
\sqrt{\mathbb{E}\,\|tv_t\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4}+\frac{5}{\eta\mu}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4}+14\widetilde{\sigma_*}^2. \tag{D.45}
$$

The detailed proof is relegated to §D.3.3.1. We are ready for the proof of Lemma D.6. Indeed, from (D.11) and (D.45)

$$
\begin{aligned}
t^2\sqrt{\mathbb{E}\,\|v_t\|_2^4}&\le\left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{5}{\eta\mu}\left[\frac{60\;\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2}+\frac{140\;\widetilde{\sigma_*}}{t}\right]+14\widetilde{\sigma_*}^2\\
&\le\left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{310\;\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3t^2}+714\widetilde{\sigma_*}^2.
\end{aligned}
\tag{D.46}
$$

We have from (D.46)

$$
t^4\sqrt{\mathbb{E}\,\|v_t\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)t^2(t-1)^2\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{310\;\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3}+714\widetilde{\sigma_*}^2t^2
$$

$$\leq \left(1-\frac{\eta\mu}{6}\right)(t-1)^4\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{310\ \|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3}+714\widetilde{\sigma_*}^2t^2,$$

since the following holds $\frac{t^2}{(t-1)^2}\leq\frac{1-\frac{\eta\mu}{6}}{(1-\frac{\eta\mu}{6})^3}\leq\frac{1-\frac{\eta\mu}{6}}{1-\frac{\eta\mu}{2}}$ This gives, by solving the recursion,

$$\begin{aligned}
T^4\sqrt{\mathbb{E}\,\|v_T\|_2^4} &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}{T_0}^4\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\left[\frac{310\ \|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3}+714\widetilde{\sigma_*}^2t^2\right]\\
&\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}{T_0}^4\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\frac{310\ \|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3}+\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}714\widetilde{\sigma_*}^2t^2\\
&\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}{T_0}^4\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\frac{6}{\eta\mu}\cdot\frac{310\ \|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3}+\frac{6}{\eta\mu}T^2\cdot 714\widetilde{\sigma_*}^2\\
&\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}{T_0}^4\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\frac{187500\ \|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu}T^2.
\end{aligned} \tag{D.47}$$

where the summand is increasing so

$$\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}t^2\leq\frac{6}{\eta\mu}T^2.$$

All in all, this concludes

$$\begin{aligned}
\sqrt{\mathbb{E}\,\|v_T\|_2^4} &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}\frac{{T_0}^4}{T^4}\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\frac{187500\ \|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}\\
&\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}\frac{{T_0}^4}{T^4}\sqrt{\mathbb{E}\,\left\|v_{T_0}\right\|_2^4}+\frac{187500\ \|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}.
\end{aligned}$$

Bringing the burn-in upper bounds (D.43), we arrive at our final result for bounding $\sqrt{\mathbb{E}\,\|v_T\|_2^4}$:

$$\begin{aligned}
\sqrt{\mathbb{E}\,\|v_T\|_2^4} &\leq \left(\frac{3{T_0}^4}{T^4}\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4}+\frac{8{T_0}^3}{T^4}\widetilde{\sigma_*}^2\right)+\frac{187500\ \|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}\\
&\leq \frac{1359375\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{4484\widetilde{\sigma_*}^2}{\eta\mu T^2}.
\end{aligned}$$

D.3.3 Proofs of recursive bounds on v_t

In this section, we prove Lemmas D.12 and D.4, the two recursive bounds for $\{v_t\}_{t\geq T_0}$ used in the proof of main theorems.

D.3.3.1 Proof of Lemma D.12

By definition, we note that:

$$tv_t = (t-1)\left(v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right) + \nabla f(\theta_{t-1};\xi_t).$$

Subtracting off a $\nabla F(\theta_{t-1})$ term from both sides we have

$$tv_t - \nabla F(\theta_{t-1}) = (t-1)\left(v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right) + \underbrace{\nabla f(\theta_{t-1};\xi_t) - \nabla F(\theta_{t-1})}_{=\varepsilon_t(\theta_{t-1})}.$$

Taking the fourth moments on both sides, we have

$$\begin{aligned}
&\mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^4 \\
&= \mathbb{E}\,\|(t-1)v_{t-1} + (t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&= (t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 + \underbrace{4\mathbb{E}\left[\|(t-1)v_{t-1}\|_2^2\left\langle (t-1)v_{t-1},\,(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\rangle\right]}_{=:T_2} \\
&\quad + \underbrace{6\mathbb{E}\left[\|(t-1)v_{t-1}\|_2^2\,\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^2\right]}_{=:T_1} \\
&\quad + \underbrace{4\mathbb{E}\left[\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^3\,\|(t-1)v_{t-1}\|_2\right]}_{=:T_3} \\
&\quad + \mathbb{E}\left[\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4\right].
\end{aligned} \tag{D.48}$$

To bound term T_1, we apply the Hölder's inequality and have

$$T_1 \leq 6\left(\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4\right)^{1/2}\left(\mathbb{E}\,\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{1/2}. \tag{D.49}$$

To bound term T_3, we again apply the Hölder's inequality:

$$\begin{aligned}
T_3 &\leq 4\left(\mathbb{E}\,\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{3/4}\left(\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4\right)^{1/4} \\
&\leq 2\left(\mathbb{E}\,\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4\right)^{1/2}\left(\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4\right)^{1/2}
\end{aligned}$$

$$+2\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4. \tag{D.50}$$

To bound term T_2, we first take expectation with respect to ξ_t and have

$$T_2=4\mathbb{E}\left[\|(t-1)v_{t-1}\|_2^2\left\langle(t-1)v_{t-1},\,(t-1)\left(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right)\right\rangle\right],$$

where

$$\left\langle v_{t-1},\,\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle\le-\frac{1}{\eta L}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2$$

and

$$\left\langle v_{t-1},\,\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\rangle\le-\frac{\mu}{\eta}\left\|\theta_{t-1}-\theta_{t-2}\right\|_2^2$$

holds true for any μ-strongly convex and L-smooth F. Then we have

$$\begin{aligned}
T_2&\le-(t-1)^4\mathbb{E}\left[\|v_{t-1}\|_2^2\left(\frac{1}{\eta L}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2+3\eta\mu\left\|v_{t-1}\right\|_2^2\right)\right]\\
&=-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\mathbb{E}\left\|v_{t-1}\right\|_2^2\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2\\
&\le-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}.
\end{aligned} \tag{D.51}$$

Combining Eqs. (D.49), (D.50) and (D.51) into Eq. (D.48) we have

$$\begin{aligned}
&\mathbb{E}\left\|tv_t-\nabla F(\theta_{t-1})\right\|_2^4\\
&\le(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4+3\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\
&\quad+8\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2}\\
&\quad-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}.
\end{aligned} \tag{D.52}$$

We now turn to bound the term $\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4$ by the following decomposition scheme:

$$\begin{aligned}
&\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\
&\le\mathbb{E}\left\|(t-1)(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}))+(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)+\varepsilon_t(\theta^*)\right\|_2^4\\
&\le8(t-1)^4\underbrace{\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4}_{=:I_1}+8\underbrace{\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)+\varepsilon_t(\theta^*)\right\|_2^4}_{=:I_2}.
\end{aligned} \tag{D.53}$$

We claim that

$$I_1 \leq 5\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 7\ell_\Xi^4 \eta^4 \mathbb{E}\,\|v_{t-1}\|_2^4, \tag{D.54}$$

and

$$I_2 \leq 8\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^4 + 8\mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^4 \leq \frac{8\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + 8\widetilde{\sigma_*}^4. \tag{D.55}$$

Combining Eqs. (D.53), (D.54) and (D.55) we have the bound

$$\begin{aligned}
&\mathbb{E}\,\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&\leq 40(t-1)^4\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 56\ell_\Xi^4\eta^4(t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 + \frac{64\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + 64\widetilde{\sigma_*}^4.
\end{aligned} \tag{D.56}$$

Then, we bring Eq. (D.56) into Eq. (D.52) and have

$$\begin{aligned}
&\mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^4 \\
&\leq (t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 + 120(t-1)^4\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 168\ell_\Xi^4\eta^4(t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 \\
&\quad + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + 192\widetilde{\sigma_*}^4 + 8\sqrt{40}(t-1)^4\left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad + 64\ell_\Xi^2\eta^2(t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad - 3\eta\mu(t-1)^4\mathbb{E}\,\|v_{t-1}\|_2^4 - \frac{(t-1)^4}{\eta L}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2} \\
&\leq \left(1 - 3\eta\mu + 64\ell_\Xi^2\eta^2 + 168\ell_\Xi^4\eta^4\right)\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4 \\
&\quad + \left(8\sqrt{40} - \frac{1}{\eta L} + 120L^2\eta^2\right)(t-1)^4\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2} \\
&\quad + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + 192\widetilde{\sigma_*}^4 \\
&\leq (1-\eta\mu)^2\mathbb{E}\,\|(t-1)v_{t-1}\|_2^4 + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 + 192\widetilde{\sigma_*}^4,
\end{aligned}$$

where the last inequality is due to the choice of $\eta \leq \frac{\mu}{64\ell_\Xi^2}$ and $\eta \leq \frac{1}{56L}$ such that

$$168\ell_\Xi^4\eta^4 \leq \eta^2\mu^2, \ell_\Xi^2\eta^2 \leq \eta\mu \qquad \text{and} \quad 8\sqrt{40} - \frac{1}{\eta L} + 120L^2\eta^2 \leq 0.$$

Taking squared root on both sides, we have

$$\sqrt{\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4} \le (1-\eta\mu)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4} + 32\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + \widetilde{\sigma_*}^2\right). \tag{D.57}$$

Furthermore, Young's inequality gives[1]

$$\begin{aligned}
&\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4 \\
&= t^4\mathbb{E}\|v_t\|_2^4 + \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + 6\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 - 4\mathbb{E}\|tv_t\|_2^3\|\nabla F(\theta_{t-1})\|_2 - 4\mathbb{E}\|tv_t\|_2\|\nabla F(\theta_{t-1})\|_2^3 \\
&\ge t^4\mathbb{E}\|v_t\|_2^4 + \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + 6\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 - 2\mathbb{E}\left[\frac{2}{\eta\mu}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 + \frac{\eta\mu}{2}\|tv_t\|_2^4\right] \\
&\quad - 2\mathbb{E}\left[\frac{\eta\mu}{2}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2}{\eta\mu}\|\nabla F(\theta_{t-1})\|_2^4\right] \\
&\ge (1-\eta\mu)\mathbb{E}\|tv_t\|_2^4 + \left(1-\frac{4}{\eta\mu}\right)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \left(6-\frac{4}{\eta\mu}-\eta\mu\right)\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 \\
&\ge (1-\eta\mu)\mathbb{E}\|tv_t\|_2^4 - \left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 - (1-\eta\mu)\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2.
\end{aligned}$$

Combining this we have

$$\begin{aligned}
&(1-\eta\mu)\mathbb{E}\|tv_t\|_2^4 - \left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 - (1-\eta\mu)\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 \\
&\le (1-\eta\mu)^2\mathbb{E}\|(t-1)v_{t-1}\|_2^4 + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + 192\widetilde{\sigma_*}^4.
\end{aligned}$$

Now we multiply both sides by $(1-\eta\mu)^{-1}$, noting that $(1-\eta\mu)^{-1} \le (1-\eta L)^{-1} \le \frac{56}{55}$, rearranging, and have

$$\begin{aligned}
\mathbb{E}\|tv_t\|_2^4 &\le (1-\eta\mu)\mathbb{E}\|(t-1)v_{t-1}\|_2^4 + \frac{(56)(64)}{(55)}\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad + \frac{\frac{(56)(192)}{(55)}\ell_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \frac{(56)(192)}{(55)}\widetilde{\sigma_*}^4 \\
&\quad + \frac{\frac{(4)(56)}{(55)}}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \frac{4}{\eta\mu}\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 \\
&\le \left(1-\frac{\eta\mu}{2}\right)^2\mathbb{E}\|(t-1)v_{t-1}\|_2^4 + \frac{7}{55\eta^2\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \frac{(56)(192)}{55}\widetilde{\sigma_*}^4
\end{aligned}$$

[1] Here, a different coefficient from the analysis as in the proof of Theorem 4.1 is adopted.

$$+\frac{56}{55}\left(\frac{1}{64\eta^2\mu^2}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+64\widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}+\frac{4}{\eta\mu}\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2.$$

Rearranging and taking squared root on both sides we conclude that

$$\sqrt{\mathbb{E}\left\|tv_t\right\|_2^4}-\frac{2}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+\frac{3}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+14\widetilde{\sigma_*}^2.$$

Further rearranging, we have

$$\sqrt{\mathbb{E}\left\|tv_t\right\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+\frac{5}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+14\widetilde{\sigma_*}^2,$$

which concludes our proof.

Proof of Eq. (D.54)**:**

We use similar decomposition as in the decomposition in Eq. (D.48) and have

$$\begin{aligned}I_1&=\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4+\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4\\&\quad+4\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^3\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2\\&\quad+6\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^2,\end{aligned}$$

where we note that we used the fact that one of the cross terms in the fourth moment decomposition $\mathbb{E}\left[\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2\left\langle\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}),\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\rangle\right]=0$. Further utilizing the Hölder's inequality, we have

$$\begin{aligned}I_1&\le\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4+\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4\\&\quad+4\left(\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4\right)^{3/4}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2\right)^{1/4}\\&\quad+6\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^2\right)^{1/2}\\&\le\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4+3\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4\\&\quad+8\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4\right)^{1/2}\\&\le\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4+3\ell_\Xi^4\eta^4\mathbb{E}\left\|v_{t-1}\right\|_2^4\\&\quad+8\ell_\Xi^2\eta^2\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\\&\le5\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4+7\ell_\Xi^4\eta^4\mathbb{E}\left\|v_{t-1}\right\|_2^4.\end{aligned}$$

This completes the proof of Eq. (D.54).

D.3.3.2 Proof of Lemma D.4

By definition, we note that:

$$v_t = \left(1 - \frac{1}{t}\right)(v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \frac{1}{t}\nabla f(\theta_{t-1};\xi_t).$$

Taking the second moments for both sides, we have:

$$\begin{aligned}\mathbb{E}\|v_t\|_2^2 = &\left(1-\frac{1}{t}\right)^2 \underbrace{\mathbb{E}\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2}_{I_1} + \frac{1}{t^2}\underbrace{\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2}_{I_2} \\ &+ 2\frac{t-1}{t^2}\underbrace{\mathbb{E}\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \nabla f(\theta_{t-1};\xi_t)\rangle}_{I_3}.\end{aligned}$$

For the first term, using the fact that $\theta_{t-1} - \theta_{t-2} = -\eta v_{t-1}$, we start with the following decomposition:

$$\begin{aligned}&\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right) \\ &= \|v_{t-1}\|_2^2 + 2\mathbb{E}\left(\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right) \\ &= \|v_{t-1}\|_2^2 - \frac{2}{\eta}\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right).\end{aligned}$$

Since F is μ-strongly convex and L-smooth, we have the following standard inequality:

$$\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle \geq \frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}.$$

Hence, when the step-size satisfies the bound $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{2\ell_{\Xi}^2}$, there is the bound:

$$\begin{aligned}I_1 &\leq \mathbb{E}\|v_{t-1}\|_2^2 - \frac{2}{\eta}\mathbb{E}\left(\frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}\right) + 2\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\ &\qquad + 2\mathbb{E}\left(\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta_{t-2};\xi_t)\|_2^2\right) \\ &\leq (1 - \eta\mu + 2\eta^2\ell_{\Xi}^2)\mathbb{E}\|v_{t-1}\|_2^2 + 2\left(1 - \frac{1}{\eta(\mu + L)}\right)\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\ &\leq (1 - 3\eta\mu/4)\mathbb{E}\|v_{t-1}\|_2^2.\end{aligned}$$

Now we study the second term I_2, note that

$$\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2 \leq \mathbb{E}\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)\|_2^2 + 4\mathbb{E}\|\nabla f(\theta^*;\xi_t)\|_2^2$$

$$
\begin{aligned}
&\le 2\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 2\mathbb{E}\left\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta^*;\xi_t)\right\|_2^2 + 4\mathbb{E}\left\|\nabla f(\theta^*;\xi_t)\right\|_2^2 \\
&\le 2\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 2\ell_\Xi^2\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^2 + 4\sigma_*^2 \\
&\le 2\left(1 + \frac{\ell_\Xi^2}{\mu^2}\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 4\sigma_*^2.
\end{aligned}
$$

For the cross term I_3, we note that:

$$
\begin{aligned}
&\mathbb{E}\left(\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \nabla f(\theta_{t-1};\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) \\
&= \mathbb{E}\left(\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\langle \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \nabla F(\theta_{t-1})\rangle \mid \mathscr{F}_{t-1}\right) \\
&\quad + \mathbb{E}\left(\langle \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \varepsilon_t(\theta_{t-1})\rangle \mid \mathscr{F}_{t-1}\right) \\
&= \underbrace{\langle v_{t-1}, \nabla F(\theta_{t-1})\rangle + \langle \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}), \nabla F(\theta_{t-1})\rangle}_{:=T_1} \\
&\quad + \underbrace{\mathbb{E}\left(\langle \varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2};\xi_t), \varepsilon(\theta_{t-1},\xi_t)\rangle \mid \mathscr{F}_{t-1}\right)}_{:=T_2}.
\end{aligned}
$$

For the term T_1, we note that:

$$
T_1 \le \left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2 + \left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2 \le (1+\eta L)\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2.
$$

For the term T_2, we have:

$$
\begin{aligned}
T_2 &\le \mathbb{E}\left(\left\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta_{t-2};\xi_t)\right\|_2 \cdot \left\|\varepsilon(\theta_{t-1};\xi_t)\right\|_2 \mid \mathscr{F}_{t-1}\right) \\
&\le \sqrt{\mathbb{E}(\left\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta_{t-2};\xi_t)\right\|_2^2 \mid \mathscr{F}_{t-1}) \cdot \mathbb{E}(\left\|\varepsilon(\theta_{t-1},\xi_t)\right\|_2^2 \mid \mathscr{F}_{t-1})} \\
&\le \ell_\Xi^2\eta\left\|v_{t-1}\right\|_2 \cdot \left\|\theta_{t-1} - \theta^*\right\|_2 \\
&\le \frac{\ell_\Xi^2}{\mu}\eta\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2.
\end{aligned}
$$

So we have:

$$
\begin{aligned}
I_3 &\le \frac{3}{2}\mathbb{E}\left(\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2\right) \le \frac{3}{2}\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2 \cdot \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2} \\
&\le \frac{t\eta\mu}{8}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{9}{2t\mu\eta}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2.
\end{aligned}
$$

Putting above estimates together, we obtain:

$$
\begin{aligned}
\mathbb{E}\left\|v_t\right\|_2^2 &\le \left(1 - \frac{1}{t}\right)^2(1 - 3\eta\mu/4)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{1}{t^2}\left(4\sigma_*^2 + 2\big(1 + \frac{\ell_\Xi^2}{\mu^2}\big)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2\right) \\
&\qquad + \frac{(t-1)\eta\mu}{4t}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{9}{t^2\mu\eta}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 \\
&\le \left(1 - \frac{1}{t}\right)^2\left(1 - \frac{\eta\mu}{2}\right)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{10}{t^2\mu\eta}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + \frac{4\sigma_*^2}{t^2},
\end{aligned}
$$

which completes the proof of this lemma.

D.3.4 Proof of Lemmas D.7 and D.8

In this section, we present the proofs of lemma D.7 and D.8, the two technical lemmas involving a test matrix $G \in \mathbb{R}^{d\times d}$.

D.3.4.1 Proof of Lemma D.7

The proof is similar to that of Lemma D.2, and we follow the notation in such lemma throughout. Indeed, we note the following telescope result:

$$T^2\mathbb{E}\,\|Gz_T\|_2^2 - T_0{}^2\mathbb{E}\,\left\|Gz_{T_0}\right\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\,\|G\varepsilon_t(\theta^*)\|_2^2 + \sum_{t=T_0+1}^{T}\mathbb{E}\,\|G\zeta_t\|_2^2 + 2\sum_{t=T_0+1}^{T}\mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\zeta_t\rangle.$$

Clearly, for each t, we have the following identity:

$$\mathbb{E}\,\|G\varepsilon_t(\theta^*)\|_2^2 = \mathrm{Tr}\big(G\Sigma^* G^\top\big).$$

For the additional terms, we note that $\mathbb{E}\,\|G\zeta_t\|_2^2 \le |||G|||_{\mathrm{op}}^2\mathbb{E}\,\|\zeta_t\|_2^2$, and following the derivation in the proof of Lemma D.2, we have the following identity:

$$\begin{aligned}&\sum_{t=T_0+1}^{T}\mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\zeta_t\rangle\\ &= T\cdot\mathbb{E}\langle G\varepsilon_T(\theta^*),\, G\varepsilon_T(\theta_{T-1}) - G\varepsilon_T(\theta^*)\rangle - T_0\cdot\mathbb{E}\langle G\varepsilon_{T_0}(\theta^*),\, G\varepsilon_{T_0}(\theta_{T_0-1}) - G\varepsilon_{T_0}(\theta^*)\rangle.\end{aligned}$$

Applying the Cauchy-Schwartz inequality, we obtain bounds similar to Eq. (D.9), for $t \in \{T_0, T\}$:

$$\begin{aligned}|t\cdot\mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta^*)\rangle| &\le t|||G|||_{\mathrm{op}}^2\cdot\sqrt{\mathbb{E}\,\|\varepsilon_t(\theta^*)\|_2^2}\cdot\sqrt{\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2}\\ &\le |||G|||_{\mathrm{op}}^2\frac{t\sigma_*\ell_\Xi}{\mu}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2}.\end{aligned}$$

For the burn-in period, we have that:

$$T_0{}^2\mathbb{E}\,\left\|Gz_{T_0}\right\|_2^2 \le 2T_0\mathbb{E}\,\left\|G\big(\varepsilon_1(\theta_0) - \varepsilon_1(\theta^*)\big)\right\|_2^2 + 2T_0\mathbb{E}\,\|G\varepsilon_1(\theta^*)\|_2^2 \le \frac{2T_0\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2 + 2T_0\mathrm{Tr}\big(G\Sigma^* G\big).$$

Putting them together, and following the derivation in Lemma D.2, we obtain the conclusion of this lemma.

D.3.4.2 Proof of Lemma D.8

The proof is similar to that of Lemma D.3. Following the notation in Lemma D.3, we have the decomposition:

$$|\mathbb{E}\langle tGz_t, Gv_t\rangle| \leq (t-\widetilde{T}^*)\left|\mathbb{E}\langle Gz_{t-\widetilde{T}^*}, Gv_t\rangle\right| + \left|\mathbb{E}\langle G(tz_t-(t-\widetilde{T}^*)z_{t-\widetilde{T}^*}), Gv_t\rangle\right|.$$

Noting that

$$\left|\mathbb{E}\langle G(tz_t-(t-\widetilde{T}^*)z_{t-\widetilde{T}^*}), Gv_t\rangle\right| \leq |||G|||_{\text{op}}^2\sqrt{\mathbb{E}\left\|tz_t-(t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|v_t\right\|_2^2},$$

and that

$$\left|\mathbb{E}\langle Gz_{t-\widetilde{T}^*}, Gv_t\rangle\right| \leq |||G|||_{\text{op}}^2\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}[v_t \mid \mathscr{F}_{t-\widetilde{T}^*}]\right\|_2^2}.$$

The rest of the proof simply follows that of Lemma D.3, with an additional factor of $|||G|||_{\text{op}}^2$ in each term.

D.4 Proofs of auxiliary lemmas in §D.2

In this section, we prove the three auxiliary lemmas used in the proof of Proposition 2. Note that the proofs of the lemmas have inter-dependencies. In the following, we first prove Lemma D.9 assuming Lemma D.10, and then prove Lemma D.10 assuming Lemma D.11. Finally, we give a self-contained proof for Lemma D.11.

D.4.1 Proof of Lemma D.9

We begin by making note of the identities

$$\begin{aligned} tv_t &= (t-1)(v_{t-1}+\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\nabla f(\theta_{t-1};\xi_t), \quad \text{and}\\ y_t &= y_{t-1}-\eta\nabla^2 f(\theta^*;\xi_t)y_{t-1}+\nabla f(\theta^*;\xi_t). \end{aligned}$$

Defining the quantity $e_t := tv_t - y_t$, we see that the two identities above imply that

$$\begin{aligned} e_t &= e_{t-1}+\big((t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))-\eta\nabla^2 f(\theta^*;\xi_t)y_{t-1}\big)+(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta^*;\xi_t))\\ &= Q_1(t)+Q_2(t)+Q_3(t), \end{aligned}$$

where we define

$$Q_1(t) := e_{t-1}-\eta\int_0^1 \nabla^2 f(\rho\theta_{t-1}+(1-\rho)\theta_{t-2};\xi_t)e_{t-1}d\rho, \qquad Q_2(t) := (\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta^*;\xi_t)),$$

$$Q_3(t) := \eta \int_0^1 \left(\nabla^2 f(\rho\theta_{t-1} + (1-\rho)\theta_{t-2};\xi_t) - \nabla^2 f(\theta^*;\xi_t)\right) y_{t-1} d\rho.$$

By the triangle inequality, we have

$$\mathbb{E}\,\|e_t\|_2^2 \le \left(\sqrt{\mathbb{E}\,\|Q_1(t)\|_2^2} + \sqrt{\mathbb{E}\,\|Q_2(t)\|_2^2} + \sqrt{\mathbb{E}\,\|Q_3(t)\|_2^2}\right)^2.$$

In the following, we bound each term $\mathbb{E}\,\|Q_i(t)\|_2^2$ in succession.

Upper bound on $\mathbb{E}\,\|Q_1(t)\|_2^2$:

Assumption 1 and Assumption 3 together imply that

$$\begin{aligned}
&\mathbb{E}\,\|Q_1(t)\|_2^2 \\
&= \mathbb{E}\,\|e_{t-1}\|_2^2 - 2\eta\mathbb{E}\int_0^1 e_{t-1}^\top \nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) e_{t-1} d\rho \\
&\quad + \eta^2 \int_0^1 \mathbb{E}\left\|\nabla^2 f(\rho\theta_{t-1} + (1-\rho)\theta_{t-2};\xi_t) e_{t-1}\right\|_2^2 d\rho \\
&= \mathbb{E}\,\|e_{t-1}\|_2^2 - \mathbb{E}\int_0^1 e_{t-1}^\top \left(2\eta\nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) - \eta^2(\nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}))^2\right) e_{t-1} d\rho \\
&\quad + \eta^2 \int_0^1 \mathbb{E}\,\|\Xi_t(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) e_{t-1}\|_2^2 d\rho \\
&\overset{(i)}{\le} \mathbb{E}\,\|e_{t-1}\|_2^2 - (2\eta - \eta^2 L)\int_0^1 e_{t-1}^\top \nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) e_{t-1} d\rho + \eta^2 \ell_\Xi^2 \int_0^1 \|e_{t-1}\|_2^2 d\rho \\
&\overset{(ii)}{\le} \mathbb{E}\,\|e_{t-1}\|_2^2 - \mu\left(2\eta - \eta^2 L\right)\mathbb{E}\,\|e_{t-1}\|_2^2 + \ell_\Xi^2 \eta^2 \mathbb{E}\,\|e_{t-1}\|_2^2.
\end{aligned}$$

In step (i), we are using the fact that $0 \preceq \nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) \preceq LI_d$, and in step (ii), we use the strong convexity of F.

For $\eta < \frac{1}{2L} \wedge \frac{\mu}{2\ell_\Xi^2}$, we have $\mathbb{E}\,\|Q_1(t)\|_2^2 \le (1-\mu\eta)\mathbb{E}\,\|e_{t-1}\|_2^2$.

Upper bound on $\mathbb{E}\,\|Q_2(t)\|_2^2$:

By Assumption 3 and Eq. (D.35), we have

$$\mathbb{E}\,\|Q_2(t)\|_2^2 \le \ell_\Xi^2 \mathbb{E}\,\|\theta_{t-1} - \theta^*\|_2^2 \le \frac{a_0 \ell_\Xi^2}{t},$$

where the last inequality follows from Theorem 4.1.

Upper bound on $\mathbb{E}\|Q_3(t)\|_2^2$:

Applying Lemma D.10 with $\widetilde{\theta}_{t-1} := \rho\theta_{t-1} + (1-\rho)\theta_{t-2} \in \mathscr{F}_{t-1}$, we have

$$\mathbb{E}\|(H_t(\rho\theta_{t-1}+(1-\rho)\theta_{t-2}) - H_t(\theta^*))y_{t-1}\|_2^2 \le c_1\sqrt{\mathbb{E}\|\rho\theta_{t-1}+(1-\rho)\theta_{t-2}) - \theta^*\|_2^2}$$
$$\le c_1\left(\sqrt{\mathbb{E}\|\theta_{t-1}-\theta^*\|_2^2} + \sqrt{\mathbb{E}\|\theta_{t-2}-\theta^*\|_2^2}\right) \le c_1\sqrt{a_0}\left(\frac{1}{\sqrt{t-1}} + \frac{1}{\sqrt{t-2}}\right) \le \frac{16c_1\sqrt{a_0}}{\sqrt{t}}.$$

Putting the bounds for (Q_1, Q_2, Q_3) together, we obtain:

$$\sqrt{\mathbb{E}\|e_t\|_2^2} \le \left(1 - \frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\|e_{t-1}\|_2^2} + \frac{4c_1^{1/2}a_0^{1/4}}{t^{1/4}} + \frac{\ell_\Xi\sqrt{a_0}}{\sqrt{t}}.$$

Solving the recursion, we have

$$\sqrt{\mathbb{E}\|e_T\|_2^2} \le (4c_1^{1/2}a_0^{1/4} + \ell_\Xi\sqrt{a_0})\sum_{s=T_0+1}^{t} s^{-\frac{1}{4}}\exp\left(-\frac{\mu\eta}{2}(T-s)\right) + e^{-\frac{\mu\eta(T-T_0)}{2}}\sqrt{\mathbb{E}\|e_{T_0}\|_2^2}.$$

For the first term, we note that:

$$\sum_{s=T_0+1}^{T} s^{-\frac{1}{4}}\exp\left(-\frac{\mu\eta}{2}(T-s)\right) \le \sum_{s=1}^{T/2}\exp\left(-\frac{\mu\eta}{2}T\right) + \frac{1}{(T/2)^{1/4}}\sum_{s=T/2}^{T} e^{-\frac{\mu\eta(T-s)}{2}}$$
$$\le \frac{T}{2}e^{-\frac{\mu\eta T}{2}} + \frac{4}{\mu\eta T^{1/4}}.$$

For T large enough, the exponentially decaying term is dominated by the $T^{-1/4}$ term. So there exists a constant $c_0 > 0$, depending on the constants $(a_0, c_1, a', \eta, \mu, T_0)$ but independent of t, such that

$$\mathbb{E}\|tv_t - y_t\|_2^2 \le \frac{c_0}{\sqrt{t}},$$

which finishes the proof.

D.4.2 Proof of Lemma D.10

Observe that Assumption (CLT.A) guarantees that

$$\mathbb{E}\left(\left\|(H_t(\theta^*) - H_t(\widetilde{\theta}_{t-1}))y_{t-1}\right\|_2^2 \,\middle|\, \mathscr{F}_{t-1}\right) \le \beta^2\left\|\widetilde{\theta}_{t-1} - \theta^*\right\|_2^2 \cdot \|y_{t-1}\|_2^2.$$

On the other hand, by Assumption 3, we have

$$\mathbb{E}\left(\left\|(H_t(\theta^*) - H_t(\widetilde{\theta}_{t-1}))y_{t-1}\right\|_2^2 \,\middle|\, \mathscr{F}_{t-1}\right) \leq 4\ell_\Xi^2 \|y_{t-1}\|_2^2 .$$

Taking a geometric average and applying the tower law yields the bound

$$\begin{aligned}\mathbb{E}\left\|\left(H_t(\widetilde{\theta}_{t-1}) - H_t(\theta^*)\right) y_{t-1}\right\|_2^2 &\leq 2\ell_\Xi \beta \mathbb{E}\left(\left\|\widetilde{\theta}_{t-1} - \theta^*\right\|_2 \cdot \|y_{t-1}\|_2^2\right) \\ &\overset{(i)}{\leq} 2\ell_\Xi \beta \sqrt{\mathbb{E}\left\|\widetilde{\theta}_{t-1} - \theta^*\right\|_2^2} \cdot \sqrt{\mathbb{E}\|y_{t-1}\|_2^4},\end{aligned}$$

where step (i) follows from the Cauchy-Schwarz inequality. Applying Lemma D.11, we are guaranteed the existence of a constant $a' > 0$ such that

$$\sup_{t \geq T_0} \mathbb{E}\|y_t\|_2^4 \leq a' < \infty.$$

Setting $c_1 = 2\ell_\Xi \beta \sqrt{a'}$ completes the proof of the claim.

D.4.3 Proof of Lemma D.11

Throughout this section, we adopt the shorthand notation $H_t := H_t(\theta^*)$ and $\Xi_t := \Xi_t(\theta^*)$. We also use Ξ to denote a generic random variable have the same law as Ξ_1. Beginning with the proof of the first claim, we take expectations on both sides of Eq. (D.34), thereby finding that

$$\mathbb{E}(y_t) = \mathbb{E}\left(y_{t-1} - \eta H_t(\theta^*) y_{t-1} + \varepsilon_t(\theta^*)\right) = (I - \eta H^*)\mathbb{E}(y_{t-1}) = (I - \eta H^*)^{t-T_0}\mathbb{E}(y_{T_0}) = 0.$$

Our next step is to control the fourth moment. For $\eta \leq \frac{1}{2L} < \frac{1}{2\mu}$, we observe that:

$$\begin{aligned}\mathbb{E}\|y_t\|_2^4 &= \mathbb{E}\|y_{t-1} - \eta H_t(\theta^*)y_{t-1} + \varepsilon_t\|_2^4 \\ &\leq \mathbb{E}\|(I - \eta H_t)y_{t-1}\|_2^4 + 4\mathbb{E}(\|(I - \eta H_t)y_{t-1}\|_2^3 \cdot \|\varepsilon_t\|_2) + 6\mathbb{E}(\|(I - \eta H_t)y_{t-1}\|_2^2 \cdot \|\varepsilon_t\|_2^2) \\ &\qquad + 4\mathbb{E}(\|\varepsilon_t\|_2^3 \cdot \|(I - \eta H_t)y_{t-2}\|_2) + \mathbb{E}\|\varepsilon_t\|_2^4 \\ &\overset{(i)}{\leq} \left(1 + \frac{\eta\mu}{2}\right)\mathbb{E}\|(I - \eta H_t)y_{t-1}\|_2^4 + \frac{24}{(\eta\mu)^3}\mathbb{E}\|\varepsilon_t\|_2^4 + \frac{216}{(\eta\mu)^2}\mathbb{E}\|\varepsilon_t\|_2^4 + \frac{24}{(\eta\mu)}\mathbb{E}\|\varepsilon_t\|_2^4 + \mathbb{E}\|\varepsilon_t\|_2^4 \\ &\leq \left(1 + \frac{\eta\mu}{2}\right)\mathbb{E}\|(I - \eta H_t)y_{t-1}\|_2^4 + \frac{157}{(\mu\eta)^3}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4,\end{aligned}$$

where in step (i), we use Young's inequality for the last four terms.

Now we study the term $\mathbb{E}\|(I - \eta H_t)y_{t-1}\|_2^4$. For $\eta < \frac{1}{L}$, straightforward calculation yields:

$$\mathbb{E}\left(\|(I - \eta H_t)y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right)$$

$$\begin{aligned}
&\leq \|(I-\eta H^*)y_{t-1}\|_2^4 + 4\mathbb{E}\left(\langle \eta \Xi_t y_{t-1}, (I-\eta H^*)y_{t-1}\rangle \|(I-\eta H^*)y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\|\eta \Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right)\\
&\quad + 6\mathbb{E}\left(\|(I-\eta H^*)y_{t-1}\|_2^2 \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) + 4\mathbb{E}\left(\langle \eta \Xi_t y_{t-1}, (I-\eta H^*)y_{t-1}\rangle \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right)\\
&\leq \|(I-\eta H^*)y_{t-1}\|_2^4 + \eta^4 \mathbb{E}\left(\|\Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) + 6\eta^2 \ell_\Xi^2 \|y_{t-1}\|_2^4\\
&\quad + 2\mathbb{E}\left(\|\eta \Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) + 2\mathbb{E}\left(\|(I-\eta H^*)y_{t-1}\|_2^2 \cdot \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right)\\
&\leq (1-3\eta\mu)\|y_{t-1}\|_2^4 + 8\eta^2 \ell_\Xi^2 \|y_{t-1}\|_2^4 + 3\eta^4 \ell_\Xi'^4 \|y_{t-1}\|_2^4 .
\end{aligned}$$

For a step-size $\eta < \frac{1}{4L} \wedge \frac{\mu}{16\ell_\Xi^2} \wedge \frac{\mu^{1/3}}{6\ell_\Xi'^{4/3}}$, we have $\mathbb{E}\left(\|(I-\eta H_t)y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) \leq (1-2\mu\eta)\|y_{t-1}\|_2^4$. Putting together these bounds, we find that

$$\mathbb{E}\|y_t\|_2^4 \leq (1-\mu\eta)\mathbb{E}\|y_{t-1}\|_2^4 + \frac{157}{(\mu\eta)^3}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4,$$

with the initial condition $\mathbb{E}\left\|y_{T_0}\right\|_2^4 = 0$. Solving this recursion leads to the bound

$$\sup_{t\geq T_0} \mathbb{E}\|y_t\|_2^4 \leq \frac{157}{(\mu\eta)^4}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 .$$

Let $a' = \frac{157}{(\eta\mu)^4}$, we prove the second claim.

Finally we study the stationary covariance of the process $\{y_t\}_{t\geq T_0}$. The existence and uniqueness of the stationary distribution was established in [134]. Let π_η denote the stationary distribution of $(y_t)_{t\geq T_0}$, and let $Q_\eta := \mathbb{E}_{Y\sim\pi_\eta}(YY^\top)$. From the first part of this lemma, we can see that $\mathbb{E}_{Y\sim\pi_\eta}(Y) = 0$. For $y_t \sim \pi_\eta$, we have $y_{t+1} \sim \pi_\eta$, and consequently,

$$\begin{aligned}
Q_\eta &= \mathbb{E}(y_{t+1}y_{t+1}^\top)\\
&= \mathbb{E}\left((I-\eta H_{t+1})y_t y_t^\top (I-\eta H_{t+1}^\top) + \varepsilon_{t+1}\varepsilon_{t+1}^\top\right) + \mathbb{E}\left(\varepsilon_{t+1}y_t^\top (I-\eta H_{t+1}^\top) + (I-\eta H_{t+1})y_t \varepsilon_{t+1}^\top\right)\\
&= Q_\eta - \eta(H^* Q_\eta + Q_\eta H^*) + \eta^2 (H^* Q_\eta H^* + \mathbb{E}(\Xi Q_\eta \Xi)) + \Sigma^* .
\end{aligned}$$

In the last equation, we use the fact that $\mathbb{E}(y_t) = 0$ and that y_t is independent of $(H_{t+1}, \varepsilon_{t+1})$, which leads to the following equation:

$$\mathbb{E}\left(\varepsilon_{t+1}y_t^\top (I-\eta H_{t+1}^\top)\right) = \mathbb{E}\left(\varepsilon_{t+1}(\theta^*) \otimes (I-\eta H_{t+1}(\theta^*))\right)[\mathbb{E}(y_t)] = 0.$$

Therefore, the matrix Q_η satisfies the equation

$$H^* Q_\eta + Q_\eta H^* - \eta(H^* Q_\eta H^* + \mathbb{E}(\Xi Q_\eta \Xi)) = \frac{\Sigma^*}{\eta},$$

which completes the proof of the last part of the lemma.

Appendix E
Appendix for Chapter 5

E.1 Proof of auxiliary lemmas

For the proofs of auxiliary lemmas, we first describe a simple decomposition result for the process $(z_t)_{t\geq T_0}$ which plays a central role in our analysis.

A key decomposition result

The proof for all the results about ROOT-SGD relies on a decomposition of the difference $z_t := v_t - \nabla F(\theta_{t-1})$ that exposes the underlying martingale structure. In particular, beginning with the definition (5.5) of the updates, for any iterate $t \geq T_0$, we have

$$\begin{aligned} z_t = v_t - \nabla F(\theta_{t-1}) &= \frac{1}{t}\varepsilon_t(\theta_{t-1}) + \left(1-\frac{1}{t}\right)(v_{t-1} - \nabla F(\theta_{t-2})) + \left(1-\frac{1}{t}\right)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) \\ &= \frac{1}{t}\varepsilon_t(\theta_{t-1}) + \left(1-\frac{1}{t}\right)z_{t-1} + \left(1-\frac{1}{t}\right)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) \end{aligned}$$

Unwinding this relation recursively yields

$$z_t = \underbrace{\frac{1}{t}\sum_{s=T_0}^{t}\varepsilon_s(\theta_{s-1})}_{:=M_t} + \frac{T_0}{t}z_{T_0} + \underbrace{\frac{1}{t}\sum_{s=T_0}^{t}(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}))}_{:=\Psi_t} \tag{E.1}$$

It can be seen that both both of the sequences $\{M_t\}_{t\geq T_0}$ and $\{\Psi_t\}_{t\geq T_0}$ are martingales adapted to the filtration $(\mathscr{F}_t)_{t\geq T_0}$. We make use of this martingale decomposition throughout our analysis.

E.1.1 Proof of Lemma 5.6

By definition, we note that:

$$v_t = \left(1 - \frac{1}{t}\right)(v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \frac{1}{t}\nabla f(\theta_{t-1};\xi_t)$$

Taking the second moments for both sides, we have:

$$\begin{aligned}\mathbb{E}\|v_t\|_2^2 = {} & \left(1 - \frac{1}{t}\right)^2 \underbrace{\mathbb{E}\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2}_{I_1} + \frac{1}{t^2}\underbrace{\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2}_{I_2} \\ & + 2\underbrace{\frac{t-1}{t^2}\mathbb{E}\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \nabla f(\theta_{t-1};\xi_t)\rangle}_{I_3}\end{aligned}$$

For the first term, using the fact that $\theta_{t-1} - \theta_{t-2} = -\eta_{t-1}v_{t-1}$, we start with the following decomposition:

$$\begin{aligned}& \mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right) \\ & = \|v_{t-1}\|_2^2 + 2\mathbb{E}(\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle \mid \mathscr{F}_{t-1}) + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right) \\ & = \|v_{t-1}\|_2^2 - \frac{2}{\eta_{t-1}}\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right)\end{aligned}$$

Since F is μ-strongly convex and L-smooth, we have the following standard inequality:

$$\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle \geq \frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}$$

Hence, when the step size satisfies the bound $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{2\ell_\Xi^2}$, there is the bound:

$$\begin{aligned}I_1 \leq {} & \mathbb{E}\|v_{t-1}\|_2^2 - \frac{2}{\eta_{t-1}}\mathbb{E}\left(\frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}\right) \\ & + 2\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + 2\mathbb{E}\left(\|\varepsilon(\theta_{t-1}, \xi_t) - \varepsilon(\theta_{t-2}, \xi_t)\|_2^2\right) \\ \leq {} & (1 - \eta_{t-1}\mu + 2\eta_{t-1}^2\ell_\Xi^2)\mathbb{E}\|v_{t-1}\|_2^2 + 2\left(1 - \frac{1}{\eta_{t-1}(\mu + L)}\right)\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\ \leq {} & \left(1 - \frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\|v_{t-1}\|_2^2\end{aligned}$$

Now we study the second term, note that

$$\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2 \leq 2\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)\|_2^2 + 2\mathbb{E}\|\nabla f(\theta^*;\xi_t)\|_2^2$$

$$
\begin{aligned}
&\leq 4\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 4\mathbb{E}\left\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta^*,\xi_t)\right\|_2^2 + 2\mathbb{E}\left\|\nabla f(\theta^*;\xi_t)\right\|_2^2 \\
&\leq 4\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 4\ell_{\Xi}^2\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^2 + 2\sigma_*^2 \\
&\leq 4\left(1 + \frac{\ell_{\Xi}^2}{\mu^2}\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 2\sigma_*^2
\end{aligned}
$$

For the cross term, we note that:

$$
\begin{aligned}
&\mathbb{E}\left(\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t),\, \nabla f(\theta_{t-1};\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) \\
&= \mathbb{E}\left(\langle v_{t-1},\, \nabla f(\theta_{t-1},\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\langle \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t),\, \nabla F(\theta_{t-1})\rangle \mid \mathscr{F}_{t-1}\right) \\
&\quad + \mathbb{E}\left(\langle \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t),\, \varepsilon_t(\theta_{t-1})\rangle \mid \mathscr{F}_{t-1}\right) \\
&= \underbrace{\langle v_{t-1},\, \nabla F(\theta_{t-1})\rangle + \langle \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}),\, \nabla F(\theta_{t-1})\rangle}_{:=T_1} + \underbrace{\mathbb{E}\left(\langle \varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t),\, \varepsilon(\theta_{t-1},\xi_t)\rangle \mid \mathscr{F}_{t-1}\right)}_{:=T_2}
\end{aligned}
$$

For the term T_1, we note that:

$$
T_1 \leq \left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2 + \left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2 \leq (1 + \eta_{t-1}L)\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2
$$

For the term T_2, we have:

$$
\begin{aligned}
T_2 &\leq \mathbb{E}\left(\left\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)\right\|_2 \cdot \left\|\varepsilon(\theta_{t-1},\xi_t)\right\|_2 \mid \mathscr{F}_{t-1}\right) \\
&\leq \sqrt{\mathbb{E}(\left\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)\right\|_2^2 \mid \mathscr{F}_{t-1}) \cdot \mathbb{E}(\left\|\varepsilon(\theta_{t-1},\xi_t)\right\|_2^2 \mid \mathscr{F}_{t-1})} \\
&\leq \ell_{\Xi}^2\eta_{t-1}\left\|v_{t-1}\right\|_2 \cdot \left\|\theta_{t-1} - \theta^*\right\|_2 \\
&\leq \frac{\ell_{\Xi}^2}{\mu}\eta_{t-1}\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2
\end{aligned}
$$

So we have:

$$
\begin{aligned}
I_3 &\leq 3\mathbb{E}\left(\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2\right) \leq 3\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2 \cdot \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2} \\
&\leq \frac{t\eta_{t-1}\mu}{8}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{18}{t\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2
\end{aligned}
$$

Putting above estimates together, we obtain:

$$
\begin{aligned}
\mathbb{E}\left\|v_t\right\|_2^2 &\leq \left(1 - \frac{1}{t}\right)^2\left(1 - \frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{1}{t^2}\left(2\sigma_*^2 + 4\big(1 + \frac{\ell_{\Xi}^2}{\mu^2}\big)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2\right) \\
&\quad + \frac{(t-1)\eta_{t-1}\mu}{8t}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{18}{t^2\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 \\
&\leq \left(1 - \frac{1}{t}\right)^2\left(1 - \frac{\eta_{t-1}\mu}{4}\right)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{26}{t^2\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + \frac{2\sigma_*^2}{t^2}
\end{aligned}
$$

which finishes the proof.

E.1.2 Proof of Lemma 5.7

Taking the squared norm of z_t in the martingale decomposition (E.1) and applying the triangle inequality yields

$$\mathbb{E}\|z_t\|_2^2 \le \frac{2}{t^2}\mathbb{E}\|M_t\|_2^2 + \frac{{T_0}^2}{t^2}\|z_0\|_2^2 + \frac{2}{t^2}\mathbb{E}\|\Psi_t\|_2^2$$

For the martingale M_t, we have:

$$\mathbb{E}\|M_t\|_2^2 = \sum_{s=1}^{t}\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^2 \le 2t\sigma_*^2 + 2\ell_\Xi^2\sum_{s=1}^{t}\mathbb{E}\|\theta_{s-1}-\theta^*\|_2^2$$

For the martingale Ψ_t, we have:

$$\mathbb{E}\|\Psi_t\|_2^2 = \sum_{s=1}^{t}(s-1)^2\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2 \le \ell_\Xi^2\sum_{s=1}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\|v_{s-1}\|_2^2$$

Combining the pieces yields

$$\mathbb{E}\|z_t\|_2^2 \le \frac{{T_0}^2\|z_0\|_2^2}{t^2} + \frac{4\sigma_*^2}{t} + \frac{4\ell_\Xi^2}{t^2}\sum_{s=1}^{t}\mathbb{E}\|\theta_{s-1}-\theta^*\|_2^2 + \frac{\ell_\Xi^2}{t^2}\sum_{s=1}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\|v_{s-1}\|_2^2$$

Note that the μ-strong convexity condition (cf. Assumption 9) ensures that $\|\theta_{s-1}-\theta^*\|_2 \le \frac{1}{\mu}\|\nabla F(\theta_{t-1})\|_2$. Plugging this bound into the inequality above completes the proof.

E.1.3 Proof of Lemma 5.8

Denote $\ell_t := \sum_{s=T_0}^{t}\eta_s$, which is the aggregated step sizes up to time t.

Recursively applying the inequality (5.19b), and noting that H_t is a non-decreasing sequence and that η_t is non-increasing, we obtain:

$$\begin{aligned}W_T &\le 2\sigma_*^2\sum_{t=T_0}^{T-1}e^{-\mu(\ell_T-\ell_t)} + 2CH_{T-1}\sum_{t=T_0}^{T-1}\frac{e^{-\mu(\ell_T-\ell_t)}}{t\mu\eta_{t-1}} + \sum_{t=T_0}^{T-1}e^{-\mu(\ell_T-\ell_{T_0})}W_{T_0}\\ &\le \frac{2\sigma_*^2}{\eta_T\mu} + \frac{CH_{T-1}}{T(\mu\eta_{T-1})^2} + e^{-\mu(\ell_T-\ell_{T_0})}{T_0}^2\mathbb{E}\left\|v_{T_0}\right\|_2^2\end{aligned}$$

Substituting the bound into Eq (5.19a), we obtain:

$$H_T \le 4\sigma_*^2 + 2\mathbb{E}\left\|z_{T_0}\right\|_2^2 T_0 + C'\ell_\Xi^2\frac{2\sigma_*^2}{\mu}\sup_{T_0\le t\le T}\frac{1}{t}\sum_{s=T_0}^{t-1}\eta_s + 2CC'\ell_\Xi^2 H_T\sup_{T_0\le t\le T}\frac{1}{t}\sum_{s=T_0}^{t-1}\frac{1}{s\mu^2}$$

$$+C'\ell_{\Xi}^2 {T_0}^2 \mathbb{E}\left\|v_{T_0}\right\|_2^2 \cdot \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2$$

For the quantities involving step size sequences in the inequality above, we have:

$$\sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \eta_s \leq \sup_{T_0 \leq t \leq T} \frac{1}{t - T_0 + 1} \sum_{s=T_0}^{t-1} \eta_s \leq \eta_{T_0}$$

$$\sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2 \leq \frac{1}{T_0} \sum_{s=T_0}^{T-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2 \leq \frac{\eta_{T_0}}{T_0 \mu}$$

For $T_0 > \frac{4CC'\ell_{\Xi}^2}{\mu^2}$, we have $2CC'\ell_{\Xi}^2 \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \frac{1}{s\mu^2} \leq \frac{1}{2}$, and consequently:

$$H_T \leq c \left(\sigma_*^2 + \frac{\ell_{\Xi}^2 T_0 \eta_{T_0}}{\mu} W_{T_0} + H_{T_0} \right)$$

for universal constants $c > 0$.

Substituting back into the bound (5.19b), for $T \geq T_0 \geq (\mu \eta_T)^{-1}$, we obtain:

$$W_T = T^2 \mathbb{E}\left\|v_T\right\|_2^2 \leq \frac{c'}{\eta_T \mu} \sigma_*^2 + c' \left(\frac{T_0}{T\mu^2 \eta_{T-1}^2} + e^{-\mu(\ell_T - \ell_{T_0})} {T_0}^2 \right) W_{T_0}$$

E.1.4 Proof of Lemma 5.10

By the martingale decomposition (E.1), for any $t \geq T_0$, we have the identity

$$t^2 \mathbb{E}\left\|G z_t\right\|_2^2 = {T_0}^2 \mathbb{E}\left\|G z_{T_0}\right\|_2^2 + \mathbb{E}\left([GM]_t\right) + \mathbb{E}\left([G\Psi]_t\right) + 2\mathbb{E}\left([GM, G\Psi]_t\right) \qquad \text{(E.2)}$$

For the quadratic variation terms, we note that

$$\begin{aligned}
\mathbb{E}\left([GM]_t\right) &= \sum_{s=T_0+1}^{t} \mathbb{E}\left\|G\varepsilon_s(\theta_{s-1})\right\|_2^2 \\
&\leq \sum_{s=T_0+1}^{t} \left(\sqrt{\mathbb{E}\left\|G\varepsilon_s(\theta^*)\right\|_2^2} + \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} \sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\right\|_2^2} \right)^2 \\
&\leq \sum_{s=T_0+1}^{t} \left(\sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)} + \ell_{\Xi} \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} \sqrt{\mathbb{E}\left\|\theta_{s-1} - \theta^*\right\|_2^2} \right)^2 \\
&\leq (t - T_0)\mathrm{Tr}\left(G\Sigma^* G^\top\right) + 2\sum_{s=T_0+1}^{t} \sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)} \ell_{\Xi} \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} r_\theta(s) + \sum_{s=T_0+1}^{t} \ell_{\Xi}^2 \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}}^2 r_\theta^2(s)
\end{aligned}$$

$$\leq (t-T_0)\mathrm{Tr}\left(G\Sigma^* G^\top\right) + |||G|||_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} \left(2\sigma_* \ell_\Xi r_\theta(s) + \ell_\Xi^2 r_\theta^2(s)\right) \tag{E.3}$$

and

$$\begin{aligned}
\mathbb{E}\left([G\Psi]_t\right) &= \sum_{s=T_0+1}^{t} (s-1)^2 \mathbb{E}\,\|G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\|_2^2 \\
&\leq \ell_\Xi^2 |||G|||_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)^2 \mathbb{E}\,\|\theta_{s-1} - \theta_{s-2}\|_2^2 \\
&\leq \ell_\Xi^2 |||G|||_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)^2 \eta_{s-1}^2 r_v^2(s) \qquad \text{(E.4)}
\end{aligned}$$

We decompose the cross variation term in two parts, and bound them separately.

$$\begin{aligned}
\mathbb{E}\left([GM, G\Psi]_t\right) &= \sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle \\
&= \underbrace{\sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_1(t)} \\
&\quad + \underbrace{\sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_2(t)}
\end{aligned}$$

For the term Q_1, Cauchy–Schwartz inequality leads to the bound:

$$\begin{aligned}
Q_1(t) &\leq \sum_{s=T_0+1}^{t} (s-1)|||G|||_{\mathrm{op}}^2 \sqrt{\mathbb{E}\,\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\|_2^2 \cdot \mathbb{E}\,\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\|_2^2} \\
&\leq \ell_\Xi^2 |||G|||_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)\eta_{s-1} \sqrt{\mathbb{E}\,\|\theta_{s-1} - \theta^*\|_2^2 \cdot \mathbb{E}\,\|v_{s-1}\|_2^2} \\
&\leq \ell_\Xi^2 |||G|||_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)\eta_{s-1} r_v(s) r_\theta(s) \qquad \text{(E.5)}
\end{aligned}$$

For the term Q_2, we note that

$$\begin{aligned}
Q_2(t) &= \sum_{s=T_0+1}^{t} (s-1)\left(\mathbb{E}\langle G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1})\rangle - \mathbb{E}\langle G\varepsilon_{s-1}(\theta^*), G\varepsilon_{s-1}(\theta_{s-2})\rangle\right) \\
&\overset{(i)}{=} (T_0-1)\mathbb{E}\langle G\varepsilon_t(\theta^*), G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{T_0-1})\rangle + \sum_{s=T_0}^{t-1} \mathbb{E}\langle G\varepsilon_t(\theta^*), G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{s-1})\rangle
\end{aligned}$$

$$\overset{(ii)}{\leq} (T_0-1)\sigma_*\ell_\Xi |||G|||_{\mathrm{op}}^2\sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta_{T_0-1}\right\|_2^2}+\sigma_*\ell_\Xi |||G|||_{\mathrm{op}}^2\sum_{s=T_0}^{t-1}\sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta_{s-1}\right\|_2^2}$$

$$\leq 2\sigma_*\ell_\Xi |||G|||_{\mathrm{op}}^2\left(T_0\left\|\theta_0-\theta^*\right\|_2+tr_\theta(t)+\sum_{s=T_0}^{t-1}r_\theta(s)\right) \tag{E.6}$$

In step (i), we apply Abel's summation formula, and in step (ii), we use the Cauchy–Schwartz inequality.

Finally, for the initial condition, we have the bound:

$$\mathbb{E}\left\|Gz_{T_0}\right\|_2^2 \leq |||G|||_{\mathrm{op}}^2\cdot\mathbb{E}\left\|z_{T_0}\right\|_2^2 \leq |||G|||_{\mathrm{op}}^2\cdot\frac{2(\sigma_*^2+\ell_\Xi^2\left\|\theta_0-\theta^*\right\|_2^2)}{T_0} \tag{E.7}$$

Collecting the bounds (E.3)-(E.7) and substituting into the decomposition (E.2), we obtain the inequality:

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \left(1+\frac{T_0}{T}\right)\cdot\frac{\mathrm{Tr}\left(G\Sigma^*G^\top\right)}{T}+c\frac{|||G|||_{\mathrm{op}}^2\sigma_*\ell_\Xi}{T^2}\sum_{s=T_0}^{T}r_\theta(s)$$
$$+c\frac{|||G|||_{\mathrm{op}}^2\ell_\Xi^2}{T^2}\sum_{s=T_0}^{T}\left(r_\theta(s)+(s-1)\eta_{s-1}r_v(s)\right)^2+c\frac{T_0|||G|||_{\mathrm{op}}^2\left(\sigma_*+\ell_\Xi\left\|\theta_0-\theta^*\right\|_2\right)^2}{T^2}$$

for a universal constant $c>0$.

Invoking Proposition 5.1, we note that:

$$r_\theta(t)\leq c\frac{\sigma_*}{\mu\sqrt{t}}+\frac{\sqrt{T_0\log t}}{\mu^2 t}\left(\ell_\Xi+\frac{1}{\eta_t\sqrt{t}}\right)\left\|\nabla F(\theta_0)\right\|_2 \qquad \text{and} \quad r_v(t)\leq c\frac{\sigma_*}{t\sqrt{\mu\eta_t}}+\frac{\sqrt{T_0}}{\mu\eta_t t^{3/2}}\left\|\nabla F(\theta_0)\right\|_2$$

Substituting into above upper bound, we obtain:

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \frac{\mathrm{Tr}\left(G\Sigma^*G^\top\right)}{T}+c|||G|||_{\mathrm{op}}^2\left(\frac{\ell_\Xi}{\mu T^{3/2}}+\frac{\ell_\Xi^2\sum_{s=T_0}^{T}\eta_s}{\mu T^2}+\frac{T_0}{T^2}\right)\sigma_*^2$$
$$+c|||G|||_{\mathrm{op}}^2\frac{\ell_\Xi^2 T_0\log T}{\mu^2T^2}\left(1+\frac{1}{\mu\ell_\Xi}\sum_{s=T_0}^{T}\frac{1}{\eta_s^2 s^{5/2}}\right)\left\|\nabla F(\theta_0)\right\|_2^2$$

For the stepsize choice $\eta_t=\frac{1}{\mu {T_0}^{1-\alpha}t^\alpha}$, we have the bound

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \frac{\mathrm{Tr}\left(G\Sigma^*G^\top\right)}{T}+c|||G|||_{\mathrm{op}}^2\left(\frac{T_0}{T}\right)^{1/2\wedge\alpha}\frac{\sigma_*^2}{T}+c|||G|||_{\mathrm{op}}^2\frac{{T_0}^2\log T}{T^2}\left(1+\frac{T^{2\alpha-3/2}}{{T_0}^{2\alpha-3/2}}\right)\left\|\nabla F(\theta_0)\right\|_2^2$$

which proves this lemma.

E.1.5 Proof of Lemma 5.11

Similar to the proof of Lemma 5.6, we use the decomposition

$$\begin{aligned}
\mathbb{E}\|v_t\|_2^4 \leq \left(1-\frac{1}{t}\right)^4 \mathbb{E}\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4 \\
+\frac{4}{t}\left(1-\frac{1}{t}\right)^3 \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \langle v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t), \nabla f(\theta_{t-1},\xi_t)\rangle\right) \\
+\frac{6}{t^2}\left(1-\frac{1}{t}\right)^2 \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \cdot \|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right) \\
+\frac{4}{t^3}\left(1-\frac{1}{t}\right) \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2 \cdot \|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right) \\
+\frac{1}{t^4}\mathbb{E}\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \quad \text{(E.8)}
\end{aligned}$$

We claim the following bounds on the relevant terms in Eq (E.8), for stepsize choice $\eta_{t-1} \leq \frac{1}{8}\left(\frac{1}{L} \wedge \frac{\mu}{\ell_\Xi^2}\right)$

$$\mathbb{E}\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4 \leq (1-\mu\eta_{t-1})\mathbb{E}\|v_{t-1}\|_2^4 \qquad \text{(E.9a)}$$

and

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \langle v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t), \nabla f(\theta_{t-1},\xi_t)\rangle\right) \\
&\leq \frac{t\mu\eta_{t-1}}{3}\mathbb{E}\|v_{t-1}\|_2^4 + \frac{c}{t}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2} \\
&\qquad \text{(E.9b)}
\end{aligned}$$

Recall that Eq (5.28) implies the bound

$$\mathbb{E}\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \leq 27\widetilde{\sigma_*}^4 + \frac{27}{(\mu\eta_{t-1})^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 \qquad \text{(E.9c)}$$

Taking these two bounds as given, we now bound the fourth moment $\mathbb{E}\|v_t\|_2^4$. First, by Hölder's inequality and Young's inequality, we have the following bounds:

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \cdot \|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right) \\
&\leq \left(\mathbb{E}\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/2}\cdot\left(\mathbb{E}\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{1/2} \\
&\leq c\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right)\\
&\leq\left(\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/4}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{3/4}\\
&\leq ct\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)+\frac{c}{t}\left(\widetilde{\sigma_*}^4+\frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)
\end{aligned}$$

Collecting above bounds, we arrive at the conclusion

$$\begin{aligned}
\mathbb{E}\,\|v_t\|_2^4\leq&\left(1-\frac{1}{t}\right)^4(1-\mu\eta_{t-1})\mathbb{E}\,\|v_{t-1}\|_2^4+\frac{c_1}{t^2}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\\
&+\frac{c_2}{t^4}\left(\widetilde{\sigma_*}^4+\frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)\\
\leq&\left[\left(1-\frac{1}{t}\right)^2\left(1-\frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{c'}{t^2}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4}\right)\right]^2
\end{aligned}$$

for universal constants $c_1,c_2,c'>0$. This completes the proof of this lemma.

Proof of Eq (E.9a)**:**

We note the following expansion:

$$\begin{aligned}
&\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\\
&\leq\mathbb{E}\,\|v_{t-1}\|_2^4+4\mathbb{E}\left(\|v_{t-1}\|_2^2\,\langle v_{t-1},\,\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}\rangle\right)+6\mathbb{E}\left(\|v_{t-1}\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2\right)\\
&\qquad+4\mathbb{E}\left(\|v_{t-1}\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^3\right)+\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\\
&\leq\left(1-4\eta_{t-1}\frac{\mu L}{\mu+L}+6\eta_{t-1}^2\ell_\Xi^2\right)\mathbb{E}\,\|v_{t-1}\|_2^4+\left(8-\frac{4}{(\mu+L)\eta_{t-1}}\right)\mathbb{E}\left(\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}\|_2^2\cdot\|v_{t-1}\|_2^2\right)\\
&\qquad+3\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4
\end{aligned}$$

For the last term, we note that

$$\begin{aligned}
&\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\\
&\leq 8\mathbb{E}\,\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\|_2^4+8\mathbb{E}\,\|\varepsilon(\theta_{t-1},\xi_t)-\varepsilon(\theta_{t-2},\xi_t)\|_2^4\\
&\leq 8L^2\eta_{t-1}^2\mathbb{E}\left(\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\|_2^2\cdot\|v_{t-1}\|_2^2\right)+8\ell_\Xi^4\eta_{t-1}^4\mathbb{E}\,\|v_{t-1}\|_2^4
\end{aligned}$$

Putting them together, for $\eta_{t-1}\leq\frac{1}{8}\left(\frac{1}{L}\wedge\frac{\mu}{\ell_\Xi^2}\right)$, we arrive at the contraction bound

$$\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4$$

$$
\begin{aligned}
&\le \left(1 - 4\eta_{t-1}\frac{\mu L}{\mu + L} + 6\eta_{t-1}^2 \ell_\Xi^2 + 24\eta_{t-1}^4 \ell_\Xi^4\right)\mathbb{E}\,\|v_{t-1}\|_2^4 \\
&\qquad + \left(8 - \frac{4}{(L+\mu)\eta_{t-1}} + 24L^2\eta_{t-1}^2\right)\mathbb{E}\left(\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}\|_2^2 \cdot \|v_{t-1}\|_2^2\right) \\
&\le (1 - \mu\eta_{t-1})\mathbb{E}\,\|v_{t-1}\|_2^4
\end{aligned}
$$

which proves this bound.

Proof of Eq (E.9b)**:**

Denote the following random variables for notational convenience

$$
\lambda_{t-1} := v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) \qquad \text{and} \qquad \zeta_t := \varepsilon(\theta_{t-1}, \xi_t) - \varepsilon(\theta_{t-2}, \xi_t)
$$

For $\eta_{t-1} \le \frac{1}{2L}$, it is easy to see the bound $\|\lambda_{t-1}\|_2 \le \|v_{t-1}\|_2$ almost surely. And we note by Assumption $10'$ that

$$
\mathbb{E}\left(\|\zeta_t\|_2^4 \mid \mathscr{F}_{t-1}\right) \le \ell_\Xi^4 \,\|\theta_{t-1} - \theta_{t-2}\|_2^4 = \ell_\Xi^4 \eta_{t-1}^4 \,\|v_{t-1}\|_2^4
$$

We note the decomposition

$$
\begin{aligned}
&\mathbb{E}\left(\|\lambda_{t-1} + \zeta_t\|_2^2 \langle \lambda_{t-1} + \zeta_t, \nabla f(\theta_{t-1}, \xi_t)\rangle\right) \\
\le\, &\mathbb{E}\left(\|\lambda_{t-1}\|_2^2 \langle \lambda_{t-1}, \nabla F(\theta_{t-1})\rangle\right) + 6\mathbb{E}\left(\|\zeta_t\|_2 \cdot \left(\|\lambda_{t-1}\|_2^2 + \|\zeta_t\|_2^2\right) \cdot \|\nabla f(\theta_{t-1}, \xi_t)\|_2\right)
\end{aligned}
$$

Applying Eq (5.28) accompanied with Hölder's inequality, we can bound the above terms as follows

$$
\mathbb{E}\left(\|\lambda_{t-1}\|_2^2 \langle \lambda_{t-1}, \nabla F(\theta_{t-1})\rangle\right) \le \left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4} \cdot \left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}
$$

$$
\begin{aligned}
\mathbb{E}\left(\|\zeta_t\|_2 \,\|\lambda_{t-1}\|_2^2 \,\|\nabla f(\theta_{t-1}, \xi_t)\|_2\right) &\le 3\ell_\Xi \eta_{t-1}\mathbb{E}\left(\|v_{t-1}\|_2^3 \cdot \left(\widetilde{\sigma_*} + \frac{\ell_\Xi}{\mu}\,\|\nabla F(\theta_{t-1})\|_2\right)\right) \\
&\le 3\ell_\Xi \eta_{t-1}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4} \cdot \left(\widetilde{\sigma_*} + \frac{\ell_\Xi}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left(\|\zeta_t\|_2^3 \cdot \|\nabla f(\theta_{t-1}, \xi_t)\|_2\right) &\le \left(\mathbb{E}\,\|\zeta_t\|_2^4\right)^{3/4} \cdot \left(\mathbb{E}\,\|\nabla f(\theta_{t-1}, \xi_t)\|_2^4\right)^{1/4} \\
&\le 3\ell_\Xi^3 \eta_{t-1}^3 \left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4} \cdot \left(\widetilde{\sigma_*} + \frac{\ell_\Xi}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)
\end{aligned}
$$

Collecting the three terms, and noting that $\eta_{t-1} \leq \left(\frac{1}{L} \wedge \frac{\mu}{\ell_\Xi^2}\right) \leq \frac{1}{\ell_\Xi}\sqrt{\frac{\mu}{L}} \leq \frac{1}{\ell_\Xi}$, we have

$$\begin{aligned}
&\mathbb{E}\left(\|\lambda_{t-1}+\zeta_t\|_2^2 \langle \lambda_{t-1}+\zeta_t, \nabla f(\theta_{t-1},\xi_t)\rangle\right)\\
&\leq c\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{3/4} \cdot \left(\left(\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4} + \ell_\Xi \eta_{t-1}\widetilde{\sigma_*}\right)\\
&\leq \frac{t\mu\eta_{t-1}}{3}\mathbb{E}\|v_{t-1}\|_2^4 + \frac{c}{t}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2}
\end{aligned}$$

which proves this inequality.

E.1.6 Proof of Lemma 5.12

By Eq (E.1) and Minkowski's inequality, we have the bound

$$\mathbb{E}\|z_t\|_2^4 \leq \frac{{T_0}^4}{t^4}\mathbb{E}\left\|z_{T_0}\right\|_2^4 + \frac{8}{t^4}\mathbb{E}\|M_t\|_2^4 + \frac{8}{t^4}\mathbb{E}\|\Psi_t\|_2^4$$

Invoking the BDG inequality for Hilbert-space-valued martingales, we have the moment bound

$$\begin{aligned}
\mathbb{E}\|M_t\|_2^4 &\leq c\mathbb{E}\left([M]_t^2\right) = c\cdot\mathbb{E}\left(\sum_{s=T_0+1}^{t}\|\varepsilon_s(\theta_{s-1})\|_2^2\right)^2 \qquad \text{and}\\
\mathbb{E}\|\Psi_t\|_2^4 &\leq c\mathbb{E}\left([\Psi]_t^2\right) \leq c\cdot\mathbb{E}\left(\sum_{s=T_0+1}^{t}(s-1)^2\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2\right)^2
\end{aligned}$$

Invoking Cauchy–Schwartz inequality, we note that

$$\begin{aligned}
\mathbb{E}\|M_t\|_2^4 &\leq c\sum_{s=T_0+1}^{t}\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4 + 2c\sum_{T_0+1\leq s\leq u\leq t}\mathbb{E}\left(\|\varepsilon_s(\theta_{s-1})\|_2^2\cdot\mathbb{E}\|\varepsilon_u(\theta_{u-1})\|_2^4\right)\\
&\leq c\sum_{s=T_0+1}^{t}\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4 + 2c\sum_{T_0+1\leq s\leq u\leq t}\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4}\cdot\sqrt{\mathbb{E}\|\varepsilon_u(\theta_{u-1})\|_2^4}\\
&= c\left(\sum_{s=T_0+1}^{t}\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4}\right)^2
\end{aligned}$$

Similarly, for the martingale $(\Psi_t)_{t\geq T_0}$, we have the bound

$$\sqrt{\mathbb{E}\|\Psi_t\|_2^4} \leq c\sum_{s=T_0+1}^{t}(s-1)^2\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^4}$$

By Eq (5.28), we have the bound

$$\sqrt{\mathbb{E}\,\|\varepsilon_s(\theta_{s-1})\|_2^4} \leq c\left(\widetilde{\sigma_*}^2 + \frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{s-1})\|_2^4}\right)$$

By Assumption 10′, we note that

$$\sqrt{\mathbb{E}\,\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\|_2^4} \leq \ell_\Xi^2 \sqrt{\mathbb{E}\,\|\theta_{s-1} - \theta_{s-2}\|_2^4} = \ell_\Xi^2 \eta_{t-1}^2 \sqrt{\mathbb{E}\,\|v_{s-1}\|_2^4}$$

Collecting the terms above, we arrive at the conclusion.

E.1.7 Proof of Lemma 5.13

We first note the following decomposition, which holds true for any $\widetilde{T} \in [0, t - T_0]$

$$|\mathbb{E}\langle tGz_t, Gv_t\rangle| \leq \underbrace{(t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}}, Gv_t\rangle\right|}_{:=Q_3\left(t,\widetilde{T}\right)} + \underbrace{\left|\mathbb{E}\langle G\big(tz_t - (t-\widetilde{T})z_{t-\widetilde{T}}\big), Gv_t\rangle\right|}_{:=Q_4\left(t,\widetilde{T}\right)}.$$

We claim the following upper bounds for the terms $Q_3(t,\widetilde{T})$ and $Q_4(t,\widetilde{T})$, for $\widetilde{T} \in \left[c{T_0}^{1-\alpha} t^\alpha \log t, t/2\right]$:

$$Q_3(t,\widetilde{T}) \leq c\frac{|||G|||_{\text{op}}^2 L_2}{\mu^2} t^{\frac{1+\alpha}{2}} {T_0}^{\frac{1-\alpha}{2}} \left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2} \log^2 t\, \|\nabla F(\theta_0)\|_2^3\right) \tag{E.10a}$$

$$Q_4(t,\widetilde{T}) \leq c|||G|||_{\text{op}}^2 \sqrt{\widetilde{T}{T_0}^{1-\alpha} t^\alpha} \cdot \left(\frac{\sigma_*^2}{t} + (\frac{T_0}{t})^{2-\alpha} \|\nabla F(\theta_0)\|_2^2\right) \tag{E.10b}$$

Taking these two bounds as given, we choose the time-lag parameter $\widetilde{T} := c{T_0}^{1-\alpha} t^\alpha \log t$, and arrive at the bound:

$$\begin{aligned}|\mathbb{E}\langle Gz_t, Gv_t\rangle| \leq\ & c|||G|||_{\text{op}}^2 \left(\frac{T_0}{t}\right)^{1-\alpha} \left(\frac{\sigma_*^2}{t} + (\frac{T_0}{t})^{2-\alpha} \|\nabla F(\theta_0)\|_2^2\right) \log t \\ & + c\frac{|||G|||_{\text{op}}^2 L_2}{\mu^2} \left(\frac{T_0}{t}\right)^{\frac{1-\alpha}{2}} \left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2} \log^2 t\, \|\nabla F(\theta_0)\|_2^3\right)\end{aligned}$$

which completes the proof of this lemma.

Proof of the bound (E.10a)**:**

To bound the term Q_3, we use the following lemma

Lemma E.1. *For $t > T_0$ and $s > 0$, the following bound holds true*

$$\mathbb{E}\left\|\mathbb{E}\left(v_{t+s} \mid \mathscr{F}_t\right)\right\|_2^2 \le c\widetilde{r}_v^2(t)e^{-\mu\sum_{k=1}^{s-1}\eta_k} + c\frac{L_2^2}{\mu^2}\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t)$$

See §E.1.8 for the proof of this lemma.

Taking Lemma E.1 as given, the bound for the term $Q_3(t,\widetilde{T})$ directly follows from Cauchy–Schwartz inequality.

$$Q_3(t,\widetilde{T}) = (t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}}, G\mathbb{E}\left(v_t \mid \mathscr{F}_{t-\widetilde{T}}\right)\rangle\right| \le t|||G|||_{\mathrm{op}}^2\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathscr{F}_{t-\widetilde{T}}\right)\right\|_2^2}$$

For the time-lag $\widetilde{T} \le \frac{t}{2}$, Proposition 5.1 yields the bound:

$$t\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2} \le c\sigma_*\sqrt{t} + T_0\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2 \tag{E.11a}$$

By Lemma E.1, for a non-increasing stepsize sequence, when the time-lag $\widetilde{T}$ satisfies $\mu\widetilde{T}\eta_t \ge c\log t$, we have the bound $e^{-\mu\sum_{k=1}^{s-1}\eta_k} \le \frac{1}{t^3}$. Therefore, given the stepsize choice $\eta_t = \frac{1}{\mu {T_0}^{1-\alpha}t^\alpha}$, we have the following bound holding true for $\widetilde{T} \ge c{T_0}^{1-\alpha}t^\alpha\log t$:

$$\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathscr{F}_{t-\widetilde{T}}\right)\right\|_2^2 \le cL_2\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t) \tag{E.11b}$$

Combining the bounds (E.11a) and (E.11b), we have the following bound holds true for the time-lag taking values in the interval $\widetilde{T} \in \left[c{T_0}^{1-\alpha}t^\alpha\log t, t/2\right]$ (the interval is non-empty for any $t \ge cT_0\log T_0$):

$$Q_3(t,\widetilde{T}) \le c\frac{L_2|||G|||_{\mathrm{op}}^2}{\mu}\left(\sigma_*\sqrt{t} + T_0\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2\right)\widetilde{r}_v(t)\cdot\widetilde{r}_\theta(t)$$

Noting that $\sigma_* \le \widetilde{\sigma_*}$ and $\ell_\Xi \le \ell_\Xi$, above bounds lead to the inequality:

$$Q_3(t,\widetilde{T}) \le c\frac{L_2|||G|||_{\mathrm{op}}^2}{\mu^2}t^{\frac{1+\alpha}{2}}{T_0}^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right)$$

which proves the desired result.

Proof of the bound (E.10b)**:**

For the term Q_4, we also apply Cauchy-Schwartz inequality, and obtain the following bound:

$$Q_4(t,\widetilde{T}) \leq |||G|||_{\text{op}}^2 \sqrt{2\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + 2\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2} \cdot \sqrt{\mathbb{E}\left\|v_t\right\|_2^2}$$

The mean-squared norms of martingales are just their expected quadratic variation:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([M]_t - [M]_{t-\widetilde{T}}\right) \leq 2\widetilde{T}\sigma_*^2 + 2\sum_{s=t-\widetilde{T}+1}^{t} \ell_\Xi^2 r_\theta^2(s)$$

$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([\Psi]_t - [\Psi]_{t-\widetilde{T}}\right) \leq \ell_\Xi^2 \sum_{s=t-\widetilde{T}+1}^{t} (s-1)^2 \eta_{s-1}^2 r_v^2(s)$$

Substituting with the rates in Proposition 5.1, we have the bounds:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \log t}{\mu^4 t^2}(\ell_\Xi^2 + \frac{1}{\eta_t^2 t})\left\|\nabla F(\theta_0)\right\|_2^2\right) \qquad \text{and} \tag{E.12a}$$

$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\frac{\ell_\Xi^2 \sigma_*^2 \eta_t}{\mu} + \frac{{T_0}^2}{t^3}\left\|\nabla F(\theta_0)\right\|_2^2\right) \tag{E.12b}$$

For the stepsize choice $\eta_t = \frac{1}{\mu {T_0}^{1-\alpha} t^\alpha}$, we have the bound:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + \mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\sigma_*^2 + T_0\left(\frac{T_0}{t}\right)^{\min(2,3-2\alpha)}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$

Invoking Proposition 5.1, we can bound the moment of v_t as:

$$\mathbb{E}\left\|v_t\right\|_2^2 \leq c\frac{\sigma_*^2 {T_0}^{1-\alpha}}{t^{2-\alpha}} + \left(\frac{T_0}{t}\right)^{3-2\alpha}\left\|\nabla F(\theta_0)\right\|_2^2$$

Combining above bounds, we conclude that

$$Q_4(t,\widetilde{T}) \leq c|||G|||_{\text{op}}^2 \sqrt{\widetilde{T}{T_0}^{1-\alpha} t^\alpha} \cdot \left(\frac{\sigma_*^2}{t} + (\frac{T_0}{t})^{2-\alpha}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$

E.1.8 Proof of Lemma E.1

Given $t > T_0$ fixed, denote $\Delta_s := \mathbb{E}\left(v_{t+s} \mid \mathscr{F}_t\right)$ for any $s > 0$.

Taking conditional expectations on both sides of Eq (5.5a), for $s > 0$, we have that

$$\mathbb{E}\left[v_{t+s}\mid\mathscr{F}_t\right]=\frac{t+s-1}{t+s}\mathbb{E}\left[v_{t+s-1}+\nabla F(\theta_{t+s-1})-\nabla F(\theta_{t+s-2})\mid\mathscr{F}_t\right]+\frac{1}{t+s}\mathbb{E}\left[\nabla F(\theta_{t+s-1})\mid\mathscr{F}_t\right] \tag{E.13}$$

By the decomposition $\nabla F(\theta_{t+s-1})=v_{t+s}-z_{t+s}$ and the fact that $(z_t)_{t\geq T_0}$ is a martingale, we note that

$$\mathbb{E}\left[\nabla F(\theta_{t+s-1})\mid\mathscr{F}_t\right]=\mathbb{E}\left[v_{t+s}\mid\mathscr{F}_t\right]$$

By the one-point Hessian Lipschitz condition, we note that

$$\begin{aligned}
&\|\nabla F(\theta_{t+s-1})-\nabla F(\theta_{t+s-2})+\eta_{t+s-1}H^*v_{t+s-1}\|_2\\
&=\eta_{t+s-1}\left\|\int_0^1\left(\nabla^2F\left(\gamma\theta_{t+s-1}+(1-\gamma)\theta_{t+s-2}\right)-\nabla^2F(\theta^*)\right)v_{t+s-1}\,d\gamma\right\|_2\\
&\leq\eta_{t+s-1}L_2\|v_{t+s-1}\|_2\cdot\int_0^1\|\gamma\theta_{t+s-1}+(1-\gamma)\theta_{t+s-2}-\theta^*\|_2\,d\gamma\\
&\leq\eta_{t+s-1}L_2\|v_{t+s-1}\|_2\cdot(\|\theta_{t+s-1}-\theta^*\|_2+\|\theta_{t+s-2}-\theta^*\|_2)
\end{aligned}$$

Substituting into the identity (E.13), we obtain the following inequality, which holds true almost surely for any $s>0$:

$$\begin{aligned}
\|\Delta_s\|_2&\leq\|(I-\eta_{t+s-1}H^*)\Delta_{s-1}\|_2+\eta_{t+s-1}L_2\mathbb{E}\left[\|v_{t+s-1}\|_2\cdot\left(\|\theta_{t+s-1}-\theta^*\|_2+\|\theta_{t+s-2}-\theta^*\|_2\right)\mid\mathscr{F}_t\right]\\
&\leq(1-\eta_{t+s-1}\mu)\|\Delta_{s-1}\|_2+\eta_{t+s-1}L_2\mathbb{E}\left[\|v_{t+s-1}\|_2\cdot\left(\|\theta_{t+s-1}-\theta^*\|_2+\|\theta_{t+s-2}-\theta^*\|_2\right)\mid\mathscr{F}_t\right]
\end{aligned}$$

Taking the second moment and applying Cauchy-Schwartz inequality, we arrive at the bound

$$\begin{aligned}
&\sqrt{\mathbb{E}\|\Delta_s\|_2^2}\\
&\leq(1-\eta_{t+s-1}\mu)\sqrt{\mathbb{E}\|\Delta_{s-1}\|_2^2}+2\eta_{t+s-1}L_2\left(\mathbb{E}\|v_{t+s-1}\|_2^4\cdot\left(\mathbb{E}\|\theta_{t+s-1}-\theta^*\|_2^4+\mathbb{E}\|\theta_{t+s-2}-\theta^*\|_2^4\right)\right)^{1/4}\\
&\leq(1-\eta_{t+s-1}\mu)\sqrt{\mathbb{E}\|\Delta_{s-1}\|_2^2}+2\eta_{t+s-1}L_2\widetilde{r}_v(t)\widetilde{r}_\theta(t)
\end{aligned}$$

Solving the recursion, we obtain the bound

$$\mathbb{E}\|\Delta_s\|_2^2\leq c\widetilde{r}_v^2(t)e^{-\mu\sum_{k=1}^{s-1}\eta_k}+c\frac{L_2^2}{\mu^2}\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t)$$

which finishes the entire proof.