Chris Junchi Li

Stochastic Recursive Approximation

Enhancing Stochastic Optimization for Statistical Efficiency

Technical Report, UC Berkeley College of Engineering, 2024^{+}

Current Draft: August 17, 2024

Springer

Contents

Chapter 1
The Mean-Squared Smoothness Setting

The theory and practice of stochastic optimization has focused on stochastic gradient descent (SGD) in recent years, retaining the basic first-order stochastic nature of SGD while aiming to improve it via mechanisms such as averaging, momentum, and variance reduction. Motivated from a general statistical point of M-estimators, we revisit the classical problem of solving for strongly-convex and smooth stochastic optimization problem algorithms. The resulting algorithm, which we refer to as *Recursive One-Over-T SGD* (ROOT-SGD), connects the idea of Cesàro averaging with a modern variance-reduced gradient method that efficiently reuses the historical stochastic gradients, matches the state-of-the-art convergence rate among online variance-reduced stochastic approximation methods. Under two different sets of continuity and convexity assumptions on the stochastic gradients, we first provided a unified sharp nonasymptotic convergence rate upper bound that matches the optimal leading term that exactly matches the optimal asymptotic risk $O(N^{-1})$ up to a constant prefactor in squared gradient norm, where N is the number of samples the algorithm has processed. Under the assumption of Lipschitz stochastic noise, we establish a non-asymptotic gradient norm upper bound with the near-unity prefactor and, when a multi-epoch design is employed, we achieve the resulting upper bound measured in gradient norm is of unity prefactor (arbitrarily close to 1) of the optimal asymptotic risk plus a higher-order term that scales as $O(N^{-4/3})$. Finally, we establish asymptotic efficiency of our algorithm with the asymptotic covariance matches the optimal statistical risk exactly, also the first among the provided setting.

1.1 Introduction

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

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{for a function of the form } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big]. \tag{1.1}$$

Here the expectation is taken over a random vector $\xi \in \Xi$ with distribution $\mathbb{P}$. In a statistical setting, the expectation may be an expectation over a population or it may arise as an empirical average over a random sample from a population. In either case, there is often a significant computational cost associated with computing such expectations, and the hope is that the optimization problem (3.1) can be solved efficiently based on an oracle that supplies only *stochastic gradients* of the form $\nabla f(\theta;\xi)$, for $\xi \sim \mathbb{P}$ and any θ. We assume in particular that we have access to independent and identically distributed draws, $\xi_t \sim \mathbb{P}$, for $t = 1,2,\ldots$

The simplest example of such a solution strategy 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. While such a strategy has been surprisingly successful in modern large-scale statistical machine learning problems (Nemirovski et al., 2009; Bottou et al., 2018), it can be improved on, 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*. Although the specific algorithmic variants that have been studied have similar forms, involving various weightings of past stochastic gradients, the arguments that have been employed to derive these algorithms, and to motivate theoretical analyses, have been quite different, as suggested by the range of terminology that has been employed, including "momentum," "averaging," "acceleration," and "variance reduction." Roughly speaking, these ideas reflect two main underlying goals—that of proceeding quickly to a minimum and that of arriving at a final state that provides calibrated measures of uncertainty.

Unfortunately, these two goals are in tension, and the literature has not yet arrived at a single algorithmic framework that achieves both goals. Consider in particular two seminal lines of research:

(i) The Ruppert-Polyak-Juditsky (PRJ) procedure (Polyak & Juditsky, 1992; Ruppert, 1988) incorporates slowly diminishing step sizes into SGD, thereby achieving asymptotic normality with an optimal covariance (the prefactor is unity). This meets the goal of calibrated uncertainty. However, the PRJ procedure is not optimal from a nonasymptotic point of view—it is characterized by large high-order nonasymptotic terms and does not achieve the optimal sample complexity in general (Bach & Moulines, 2011).

(ii) 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* (Le Roux et al., 2012; Shalev-Shwartz & Zhang, 2013; Johnson & Zhang, 2013; Lei & Jordan, 2017; Defazio et al., 2014). While the latter can be controlled to obtain nonasymptotic rates that are optimal, the statistical rates are nonoptimal, yielding an asymptotic efficiency that has a constant prefactor that is strictly greater than unity.

In the current chapter we present an algorithm that achieves both goals—the nonasymptotic goal of a fast finite-time convergence rate and the asymptotic goal of achieving limiting normality with a near-optimal covariance. The algorithm is closely related to existing first-order algorithms, but differs critically in its choice

of step sizes. 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 *estimator*, $\mathsf{Estimator}_t$, at iteration t, of $\nabla F(\theta_{t-1})$. We then perform a gradient step based on this estimator: $\theta_t = \theta_{t-1} - \eta_t \cdot \mathsf{Estimator}_t$.

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

$$\mathsf{Estimator}_t - \nabla F(\theta_{t-1}) = \frac{1}{t}\sum_{s=1}^{t}(\nabla f(\theta_{s-1};\xi_s) - \nabla F(\theta_{s-1})). \tag{1.2}$$

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. Our algorithm, which we refer to as *Recursive One-Over-T SGD* (ROOT-SGD), is based critically on this particular choice of weights. The recursive aspect of the algorithm arises as follows. We set $\mathsf{Estimator}_1 = \nabla f(\theta_0;\xi_1)$ and express (1.2) 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

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

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. Letting v_t denote $\mathsf{Estimator}_t$ we obtain the following recursive update:

$$\begin{aligned} v_t &= \frac{1}{t}\nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}\left(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right) + \frac{t-1}{t}v_{t-1} \\ &= \underbrace{\nabla f(\theta_{t-1};\xi_t)}_{\text{stochastic gradient}} + \underbrace{\frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2};\xi_t)\right)}_{\text{correction term}}, \end{aligned} \tag{1.3}$$

consisting of both a stochastic gradient and a correction term.

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

$$v_t = \nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2};\xi_t)\right), \tag{1.4}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t, \tag{1.5}$$

for a suitably chosen sequence $(\eta_t)_{t\geq 1}$ of positive step sizes. Note that v_t defined in (1.3) maintains a recursive estimator of $\nabla F(\theta_{t-1})$ that is unbiased *unconditionally* in the sense that $\mathbb{E}[v_t] = \mathbb{E}[\nabla F(\theta_{t-1})]$, so that the update of θ_t is an approximate gradient-descent step that moves along the negated direction $-v_t$.[1]

We initialize $\theta_0 \in \mathbb{R}^d$, and, to avoid ambiguity we define the update (3.5) at $t = 1$ to use only $v_1 = \nabla f(\theta_0; \xi_1)$. Overall, given the initialization $[\theta_0^\top, v_0, \theta_{-1}] = [\theta_0^\top, 0, \text{arbitrary}]$, at each step $t \geq 1$ we take as input $\xi_t \sim P$, and perform an update of $[\theta_t^\top, v_t^\top, \theta_{t-1}^\top]$. This update depends only on $[\theta_{t-1}^\top, v_{t-1}^\top, \theta_{t-2}^\top]$ and ξ_t, and is first order and Markovian. Moreover, the updates in ROOT-SGD algorithm have a two-timescale design: the θ_t sequence in (1.5) has prescribed step sizes η_t that are asymptotically a constant, serving as a *fast* process compared to the v_t sequence.

1.1.1 Bridging Cesàro averaging with recursive variance-reduced method

In this subsection, we discuss the connections to both Ruppert-Polyak-Juditsky averaging and extrapolation methods, and recursive variance-reduced methods.

First of all, ROOT-SGD is highly connected to the stochastic gradient descent with Ruppert-Polyak-Juditsky (PRJ) averaging (Ruppert, 1988; Polyak, 1990; Polyak & Juditsky, 1992). The latter algorithm consists of the pair of updates

$$\theta_t = \theta_{t-1} - \eta_t \nabla f(\theta_{t-1}; \xi_t), \tag{1.6}$$

$$\bar{\theta}_t = \frac{1}{t+1}\theta_t + \frac{t}{t+1}\bar{\theta}_{t-1}, \tag{1.7}$$

with $\bar{\theta}_0 = \theta_0$ as the initialization. The PRJ-averaged form of SGD $\bar{\theta}_t$ stabilizes oscillations in the SGD step and is robust to the choice of step sizes η_t. Its two-timescale design plays a key role in achieving the near-unity nonasymptotic and asymptotic convergence rate (Bach & Moulines, 2011, 2013; Polyak & Juditsky, 1992). Moreover, after the initial completion of our previous version of arXiv report, we became aware of a concurrent work on an extrapolation-smoothing method called IGT (Arnold et al., 2019), which is highly related to our ROOT-SGD algorithm with a $\frac{1}{t}$ design with an estimator that takes value at an extrapolation point.

Secondly, our ROOT-SGD in (3.5) is also closely connected to *STOchastic Recursive Momentum* (STORM) (Cutkosky & Orabona, 2019), and *Hybrid Stochastic Gradient Descent* (HSGD) (Tran-Dinh et al., 2021), with the main difference between the choice of step size, reflecting the different perspectives underlying STORM and HSGD compared to ROOT-SGD. Born in May 2019, both STORM and HSGD improves over the state-of-the-art stochastic recursive variance-reduced algorithm to a single-epoch update via a moving-average method (Nguyen et al.,

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

2017; Fang et al., 2018; Zhou et al., 2020; Wang et al., 2019). Indeed, for ROOT-SGD, a time-varying step size $\frac{1}{t}$ is used for the v_t update while an asymptotically larger value η_t is adopted for the θ_t update. As we will discuss, this critical difference is what allows ROOT-SGD to output an estimator that achieves asymptotic efficiency as the number of online samples $N \to \infty$.

From algorithmic design viewpoint, our ROOT-SGD algorithm takes advantage of both ideas. Instead of averaging the iteration trajectory, ROOT-SGD algorithm heavily utilizes idea of $\frac{1}{t}$ to average the noises of the gradients incurred at the whole iteration trajectory to maintain an asymptotic efficient gradient estimator. With a gradient transportation mechanism that is driven by recursive momentum instead of heavy extrapolation, it avoids the potentially hazardous landscape geometry and allows the gradient estimator to be globally unbiasedness, which renders its superior theoretical performance. We defer more discussions on connections to aforementioned algorithms in §2.4 of the appendix.

1.1.2 Summary of main results

We summarize our main results here. We present a single-loop first-order algorithm, ROOT-SGD, and show that the algorithm achieves desirable convergence rates both nonasymptotically and asymptotically. For generic $\theta_0 \in \mathbb{R}^d$ sampled from certain initial distribution, Our convergence rate result are classified in either the *Individually Smooth and Convex* case (smoothness and convexity assumption on the individual function, shortened as **ISC**) or the *Lipschitz Stochastic Noise* case (stochastic Lipschitz continuity assumption on the stochastic gradient noise, shortened as **LSN**). In particular, we show the following:

(i) Under either the **ISC** or the **LSN** cases, we show that ROOT-SGD achieves a nonasymptotic upper bound that for some appropriately chosen η_{max} and burn-in period $\mathscr{B}$, whenever $\eta \leq \eta_{max}$ and $T \geq \mathscr{B}$, $\mathbb{E}\|\nabla F(\theta_T)\|_2^2$ is bounded by, up to an absolute constant factor,

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \lesssim \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{\sigma_*^2}{T},$$

where $\sigma_*^2 := \mathbb{E}\|\nabla f(\theta^*;\xi_1)\|_2^2$ is the variance of the stochastic gradient at the minimizer [Theorem 1.1.A]. In terms of expected gradient norm squared, such a convergence rate bound improves upon state-of-the-art results for smooth and strongly convex objective functions F by at least a logarithmic factor in its leading term [Theorem 1.1.A]. When augmented with a multi-epoch design, we achieve a convergence rate guarantee that improves over the state-of-the-art nonasymptotic rate in Nguyen et al. (2021) by a logarithmic factor in $\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}$ for the **ISC** case, and also Allen-Zhu (2018) by a polylogarithmic factor in condition number for the **LSN** case.

(ii) Under the **LSN** case, we establish a non-asymptotic gradient norm upper bound with the leading term that matches the optimal asymptotic risk $\frac{\sigma_*^2}{N}$ of an asymptotic unity prefactor (arbitrarily close to 1) [Theorem 1.2.A]. Furthermore when an appropriate multi-epoch design is employed, the resulting upper bound on expected squared gradient norm exactly matches the optimal asymptotic risk, plus a higher-order term that scales as $O\left(\frac{1}{N^{4/3}}\right)$ [Theorem 1.2.B].[2] Moreover, under a Lipschitz smooth stochastic gradient assumption, via a refined analysis we are able to show that the ROOT-SGD algorithm with constant step-size converges non-asymptotically in gradient norm to the asymptotically optimal risk with near-unity pre-factor with the additional higher-order term being explicitly characterized that scales as $O(N^{-4/3})$. To our best knowledge, this is the first near-unity result under that set of assumptions.

(iii) For asymptotic efficiency, we show essentially in Theorem 3.1 that ROOT-SGD with a constant step size that scales down with N at rate $N^{-\alpha}$, $\alpha \in (0,1)$ converges asymptotically to a zero-mean multivariate normal distribution with optimal covariance as the number of samples $N \to \infty$:

$$\sqrt{T}\nabla F(\theta_N^{\text{final},(\eta)}) \xrightarrow{d} \mathcal{N}(0, \Sigma^*),$$

where $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top\right]$ is the covariance matrix of the stochastic gradient at the minimizer. Notably, this result only requires strong convexity, (first-order) smoothness, and a set of noise moment assumptions standard in asymptotic statistics. To our knowledge, this provides a first result for a stochastic first-order optimization algorithm that enjoys asymptotic optimality without additional continuity conditions on the Hessians.

1.1.3 Related works

Our algorithm ROOT-SGD belongs to the family of stochastic first-order algorithms, represented methods are the gradient descent (GD) and stochastic gradient descent (SGD) methods. A family that dates back to the era of Cauchy (1847) and Robbins et al. (1951), GD and SGD have recently gained unprecedented popularity in scalable machine learning (Bubeck, 2015; Bottou et al., 2018) featured by the tremendous development of deep learning applications (Goodfellow et al., 2016), primarily due to its exceptional performance when handling large-scale data samples. For expository treatments from a theoretical viewpoint, see Nemirovskii & Yudin (1983); Bertsekas & Tsitsiklis (1989); Benveniste et al. (2012); Kushner & Yin (2003); Nesterov (2004, 2018); Borkar (2008); Shalev-Shwartz (2012); Sra et al. (2012); Bubeck (2015); for applications to incremental methods see Nedic

[2] In fact, as will be discussed immediately after Theorem 3.1, this upper bound matches the Cramér-Rao asymptotic covariance in an exact fashion with the help of a mild one-point Hessian continuity assumption.

& Bertsekas (2001); Nemirovski et al. (2009); Bertsekas (2011); Ghadimi & Lan (2012, 2013a); Devolder et al. (2014); and for applications to machine learning see Zhang (2004); Rakhlin et al. (2012); Shamir & Zhang (2013); Juditsky & Nesterov (2014); Hazan & Kale (2014); Ghadimi & Lan (2013b, 2016). A recent state-of-the-art result for this class of methods asserts that stochastic gradient descent and its accelerated variant achieve an $O(1/N)$ convergence rate in expected objective gap that matches the corresponding minimax information-theoretic lower bounds in the objective gap (Agarwal et al., 2012; Woodworth & Srebro, 2016).

Stemming out of both theoretical advances and real-world practice, many variants of the (stochastic) gradient methods have been proposed, including the variance-reduced methods (Le Roux et al., 2012; Johnson & Zhang, 2013; Defazio et al., 2014), momentum-accelerated methods (Nesterov, 1983; Beck & Teboulle, 2009), second-order methods (Dennis & Moré, 1974; Nesterov & Polyak, 2006), adaptive gradient methods (Duchi et al., 2011; Kingma & Ba, 2015), iteration averaging (Ruppert, 1988; Polyak, 1990; Polyak & Juditsky, 1992), coordinate descent (Wright, 2015), and many more. In particular, the Ruppert-Polyak-Juditsky iteration averaging method (Polyak & Juditsky, 1992; Ruppert, 1988) has provably improves the robustness with respect to the step-size and achieves asymptotic normality with optimal covariance that matches the local minimax optimality (Zhu et al., 2016; Duchi & Ruan, 2021). Recent progresses have been made to study the nonasymptotic behavior of the stochastic gradient descent with iteration averaging (Bach & Moulines, 2011, 2013; Bach, 2014; Flammarion & Bach, 2015; Gadat & Panloup, 2017; Dieuleveut et al., 2017, 2020). In the context of linear regression, Jain et al. (2017, 2018) analyze the so-called *tail-averaging* that achieves exponential forgetting and optimal statistical risk simultaneously. In the context of linear stochastic approximation, Lakshminarayanan & Szepesvari (2018) studies the Ruppert-Polyak averaging method for general linear stochastic approximation that is not necessarily an optimization algorithm and includes many interesting applications in minimax game and reinforcement learning. Under stronger noise conditions, Mou et al. (2020) establish Gaussian limit and concentration inequalities for constant step-size algorithms.

Organization: The remainder of this chapter is organized as follows. §1.2 presents the statements of our main theoretical results that includes both nonasymptotic and asymptotic results, the latter specializing in asymptotic normality. §1.3 presents the proofs of our nonasymptotic and asymptotic convergence rate results. §1.4 compares our result with some concurrent work. We conclude the chapter with some future directions in §1.5.

Notation: 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 $|||M|||_{\text{op}} := \sup_{\|v\|_2=1} \|Mv\|_2$. For scalars $a, b \in \mathbb{R}$, we sometimes adopt the shorthand notation $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For sequences, $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, of positive real numbers, the notation $a_n \lesssim b_n$ means that there is a global constant C such that $a_n \leq Cb_n$ for all n, and similarly, we write $a_n \gtrsim b_n$ to mean that there is a constant $C > 0$ such that $a_n \geq Cb_n$. Throughout the chapter, we use the σ-fields $\mathscr{F}_t := \sigma(\xi_1, \xi_2, \cdots, \xi_t)$ for any $t \geq 0$. We also denote

$\lceil x \rceil$ as the least integer n that is no less than x.[3] Finally, we write $a_n \asymp b_n$ when the conditions $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ both hold.

1.2 Main results

This section is devoted to the statements and discussion of our main results, separately in the nonasymptotic or asymptotic guarantees in the strongly convex case. We build the assumptions and settings in §1.2.1. In §1.2.2, we formally introduce a single-loop ROOT-SGD and its multi-epoch variant, each of which enjoys a nonasymptotic error bound that achieves the state-of-the-art in its corresponding category. For the first time to our best knowledge, both of these error bounds enjoy an optimal leading term up to an absolute constant factor.[4] In §1.2.3 we study the improved nonasymptotic results in the **LSN** case for both single-epoch and multi-epoch ROOT-SGD that enjoys a unity prefactor of the statistical error. §1.2.4 is devoted to the asymptotic analysis of ROOT-SGD. We show that under mild distributional assumptions when the number of processed samples grows, it has an asymptotic normality behavior with a limiting covariance that nearly matches that of the optimal asymptotic risk in statistical lower bound.

1.2.1 Settings and assumptions

In this work, we consider the constant step size version of our ROOT-SGD in (3.5), initialized with a burn-in phase of length $\mathscr{B} \geq 1$ in which only the v variable is updated while the θ variable is held fixed. The purpose of the burn-in phase is to stabilize the iterates. Given an initial vector $\theta_0 \in \mathbb{R}^d$, we set $\theta_t = \theta_0$ for all $t = 1, \ldots, \mathscr{B} - 1$, and compute

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

Equivalently, we can view the step sizes in the update for θ_t as being scheduled as follows:

$$\eta_t = \begin{cases} \eta, & \text{for } t \geq \mathscr{B}, \\ 0, & \text{for } t = 1, \ldots, \mathscr{B} - 1, \end{cases} \tag{1.8}$$

[3] In our derivations, we sometimes forgo rounding up to integer to avoid unnecessary complications.

[4] Our bound is in terms of an expectation over the squared gradient norm, which imply guarantees on objective gap and squared iteration distance to optimality.

Algorithm 1 ROOT-SGD

Input: initialization θ_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: θ_T

Algorithm 2 ROOT-SGD, multi-epoch version

Input: initialization θ_0; step-size sequence $(\eta_t)_{t\geq 1}$; burn-in time $\mathscr{B}$; short epochs length $T_b \geq \mathscr{B}$; short epochs number $\mathscr{E}$
Set initialization for first epoch $\theta_0^{(1)} = \theta_0$
for $b = 1, 2, \cdots, \mathscr{E}$ **do**
  Run ROOT-SGD (Algorithm 5) with burn-in time $\mathscr{B}$ (meaning $\eta_t = 0$ for $t \leq \mathscr{B} - 1$) for T_b iterates
  Set the initialization $\theta_0^{(b+1)} := \theta_{T_b}^{(b)}$ for the next epoch
end for
Run ROOT-SGD (Algorithm 5) for $T := N - T_b\mathscr{E}$ iterates with burn-in time $\mathscr{B}$
Output: The final iterate estimator $\theta_N^{\text{final}} := \theta_T^{(\mathscr{E}+1)}$

briefed as $\eta_t = \eta \cdot 1[t \geq \mathscr{B}]$, and, accordingly, the update rule from equations (3.5) and (1.5) splits into two phases:

$$v_t = \begin{cases} \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right) & \text{for } t \geq \mathscr{B}+1, \\ \frac{1}{t}\sum_{s=1}^{t} \nabla f(\theta_0; \xi_s) & \text{for } t = 1, \ldots, \mathscr{B}, \end{cases} \tag{1.9}$$

$$\theta_t = \begin{cases} \theta_{t-1} - \eta v_t & \text{for } t \geq \mathscr{B}, \\ \theta_0 & \text{for } t = 1, \ldots, \mathscr{B}-1. \end{cases} \tag{1.10}$$

Comparing to SVRG and SARAH (Johnson & Zhang, 2013; Nguyen et al., 2017), the length of the burn-in period for our algorithm is identical to the number of processed samples, so the iteration number is identical to the sample complexity. Our ROOT-SGD is formally presented in Algorithm 5 (in the rest of this work, when referring to ROOT-SGD we mean Algorithm 5 unless otherwise specified).

For each $\theta \in \mathbb{R}^d$ let $\varepsilon_t(\theta)$ denote the noise term

$$\varepsilon_t(\theta) := \nabla f(\theta; \xi_t) - \nabla F(\theta). \tag{1.11}$$

We impose the following assumptions on the stochastic function $f(\cdot; \xi)$ and the objective function F as its expectation. First, we assume the strong convexity and smoothness of 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 $0 < \mu \leq L < \infty$:*

$$\left\|\nabla F(\theta)-\nabla F(\theta')\right\|_2 \le L\left\|\theta-\theta'\right\|_2,$$

$$\left\langle \nabla F(\theta)-\nabla F(\theta'),\theta-\theta'\right\rangle \ge \mu\left\|\theta-\theta'\right\|_2,$$

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

Second, we pose regularity conditions on the covariance matrix at the global minimizer θ^*:

Assumption 2 (Finite variance at optimality) *At the minimizer* θ^**, the stochastic gradient* $\nabla f(\theta^*;\xi)$ *has a positive definite covariance matrix* $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)(\nabla f(\theta^*;\xi))^\top\right]$*, and its trace* $\sigma_*^2 := \mathbb{E}\left\|\nabla f(\theta^*;\xi)\right\|_2^2$ *is 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 Nguyen et al. (2019); Lei & Jordan (2020) for a detailed discussion of this noise assumption.
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*

$$\mathbb{E}\left\|\varepsilon(\theta;\xi)-\varepsilon(\theta';\xi)\right\|_2^2 \le \ell_\Xi^2\left\|\theta-\theta'\right\|_2^2 \qquad \text{for all pairs } \theta,\theta' \in \mathbb{R}^d. \tag{1.12}$$

Observe that Assumptions 5 and 7 imply a mean-squared Lipschitz condition on the stochastic gradient function:

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

where the final step uses the L-Lipschitz condition on the population function F. Rather than imposing this condition directly on the stochastic gradient, it is clarifying to separate the contributions due to the noise from those of the population objective function.

In part of our analysis, as an alternative to Assumption 7, we impose the following *individual convexity and smoothness* condition (Le Roux et al., 2012; Johnson & Zhang, 2013; Defazio et al., 2014; Nguyen et al., 2017):

Assumption 4 (Individual convexity/smoothness) *Almost surely, the (random) function* $\theta \mapsto f(\theta;\xi)$ *is convex, twice continuously differentiable and satisfies the Lipschitz condition*

$$\left\|\nabla f(\theta;\xi)-\nabla f(\theta';\xi)\right\|_2 \le \ell_{max}\left\|\theta-\theta'\right\|_2 \text{ a.s.}, \quad \text{for all pairs } \theta,\theta' \in \mathbb{R}^d. \tag{1.13}$$

We remark that all Assumptions 5 and 14 along with either Assumption 7 or 4, are standard in the stochastic optimization literature (cf. Nguyen et al. (2019); Asi & Duchi (2019); Lei & Jordan (2020)). Note that Assumption 4 implies Assumption 7

with constant ℓ_{max}; in many applications, the quantity ℓ_{max} can be significantly larger than $\sqrt{L^2+\ell_\Xi^2}$ in magnitude.[5]

1.2.2 Unified nonasymptotic results

With the assumptions and settings in §1.2.1 in place, let us formalize the two cases in which we analyze the ROOT-SGD algorithm. We refer to these cases as the *Individually Smooth and Convex* case (or **ISC** for short), and the *Lipschitz Stochastic Noise* case (or **LSN** for short).

LSN case: Suppose that Assumptions 5, 14 and 7 hold, and define

$$\eta_{max} := \frac{1}{4L} \wedge \frac{\mu}{8\ell_\Xi^2}. \tag{1.14a}$$

ISC case: Suppose that Assumptions 5, 14 and 4 hold, and define

$$\eta_{max} := \frac{1}{4\ell_{max}}. \tag{1.14b}$$

The settings and assumptions in the **LSN** case is commonly seen in optimization and statistics literatures; for instance, variants of them are satisfied by a broad class of statistical models and estimators. In control and optimization literatures, the **ISC** case also sees its prevalence.

We first provide our unified nonasymptotic results for both cases above for single-epoch ROOT-SGD, as follows:

Theorem 1.1.A (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* (1.8) *for some $\eta \in (0, \eta_{max}]$, and assume that we use the following burn-in time:*

$$\mathscr{B} := \frac{24}{\eta\mu}. \tag{1.15}$$

Then, for any iteration $T \geq 1$, the iterate output θ_T from Algorithm 5 satisfies the bound

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \frac{28\,\sigma_*^2}{T} + \frac{2700\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2}. \tag{1.16}$$

We provide the proof of Theorem 1.1.A in §1.3.1. Recall that we are in either the **LSN** or **ISC** case in Theorems 1.1.A and 1.1.B. In accordance with the discussion in §1.1, our nonasymptotic convergence rate upper bound (1.16) for the expected

[5] With that said, we can treat **ISC** as a special case of **LSN** with ℓ_{max} holding in the place of ℓ_Ξ.

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 $\frac{\|\nabla F(\theta_0)\|_2^2 \mathscr{B}^2}{T^2}$, corresponds to the *optimization error* that indicates the polynomial forgetting from the initialization. Theorem 1.1.A copes with a wide range of step sizes η: fixing the number of online samples T, (1.16) 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 (1.16), we can achieve a tight upper bound on the sample complexity to achieve a statistical estimator of θ^* with gradient norm bounded by $O(\varepsilon)$:[6]

$$\begin{aligned} C_{1.1.A}(\varepsilon) &= \max\left\{\frac{74}{\eta_{max}\mu}\cdot\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{56\sigma_*^2}{\varepsilon^2}\right\} \\ &\asymp \begin{cases} \max\left\{\left(\frac{L}{\mu}+\frac{\ell_\Sigma^2}{\mu^2}\right)\cdot\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the } \textbf{LSN} \text{ case}, \\ \max\left\{\frac{\ell_{max}}{\mu}\cdot\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the } \textbf{ISC} \text{ case}. \end{cases} \end{aligned} \tag{1.17}$$

In above, the step size $\eta = \eta_{max}$ is optimized as in (1.14a) for the **LSN** and (1.14b) 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 $C_{1.1.A}(\varepsilon)$ in either case is $\asymp \frac{\sigma_*^2}{\varepsilon^2}$ which matches the optimal statistical error up to universal constants. To our best knowledge, this is achieved for the first time by single-loop stochastic first-order algorithms in the setting where only first-order smoothness condition holds, i.e. *no* continuity assumption on the Hessians are posed.

We now turn to further improve the finite-sample rate via a multi-epoch approach. On the other hand, the relatively slow $O(1/T^2)$ decay from an arbitrary initialization can be in turn mitigated by a multi-epoch approach, i.e. one periodically restarting the process for a few epochs. The algorithm consists of $\mathscr{E}$ short epochs and connected by one long epoch. Each short epoch only consumes a shared number of data points of small scales, while the long epoch uses the remaining data points. We present a multi-epoch version of the ROOT-SGD algorithm formally in Algorithm 6. Our analysis makes clear how the multi-epoch algorithm improves the lower-order term in our nonasymptotic convergence rate for stochastic gradient algorithms.

Theorem 1.1.B (Unified nonasymptotic results, multi-epoch ROOT-SGD). *Suppose that the conditions of Theorem 1.1.A hold, the step sizes are chosen according to Eq.* (1.8) *for some* $\eta \in (0, \eta_{max}]$*, and the burn-in time* $\mathscr{B}$ *is chosen according to Eq.* (1.15)*. Let the number of short epochs* $\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$*. Then the*

[6] Indeed, we choose T in Eq. (1.16) to be sufficiently large such that it satisfies the inequalities $T \geq \mathscr{B} = \lceil \frac{24}{\eta\mu} \rceil$, as well as $\frac{2700\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 T^2} \leq \frac{\varepsilon^2}{2}$ 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 (1.17) serves as a tight sample complexity upper bound.

multi-epoch estimator produced by Algorithm 6 with burn-in time $\mathscr{B} = \frac{24}{\eta\mu}$ *and small epoch length* $T_b(\eta) = \frac{142}{\eta\mu}$ *satisfies the following bound: for* $N \geq T_b(\eta)\mathscr{E} + 1$

$$\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 \leq \frac{132\sigma_*^2}{N - T_b(\eta)\mathscr{E}}. \tag{1.18}$$

By further choosing the step-size as $\eta = \eta_{max}$*, we obtain for* $N \geq 2T_b(\eta_{max})\mathscr{E} + 1$

$$\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 \leq \frac{264\sigma_*^2}{N}. \tag{1.19}$$

We provide the proof of Theorem 1.1.B in §1.3.2. As can be seen in the proof of Theorem 1.1.B in §1.3.2, when taking the optimal choice of $\eta = \eta_{max}$, the multi-epoch ROOT-SGD estimator produced by Algorithm 6 takes a constant number $\mu^{-1}{\eta_{max}}^{-1}$ of iterates for a logarithmic number of short epochs and achieves a complexity that achieves $\frac{\sigma_*^2}{\varepsilon^2}$ up to a constant prefactor (264). Moreover, as in a similar fashion as the discussions after Theorem 1.1.A, it is straightforward to verify that the bound (1.19) indicates the sample complexity of the multi-epoch ROOT-SGD to achieve $\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \varepsilon^2$:

$$\begin{aligned} C_{1.1.B}(\varepsilon) &= \max\left\{\frac{284}{\eta_{max}\mu}\left\lceil\frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta_{max}\mu\sigma_*^2}\right)\right\rceil, \frac{264\sigma_*^2}{\varepsilon^2}\right\} \\ &\asymp \begin{cases} \max\left\{\left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right)\log\left(\left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right)\cdot\frac{\|\nabla F(\theta_0)\|_2}{\sigma_*^2}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the } \mathbf{LSN} \text{ case}, \\ \max\left\{\frac{\ell_{max}}{\mu}\log\left(\frac{\ell_{max}}{\mu}\cdot\frac{\|\nabla F(\theta_0)\|_2^2}{\sigma_*^2}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the } \mathbf{ISC} \text{ case}. \end{cases} \end{aligned} \tag{1.20}$$

which has the lower-order term substantially improved to a logarithmic one, matching our discussions in §1.1.2. Such a convergence rate improves the state-of-the-art one in Nguyen et al. (2021) among variance-reduced stochastic first-order methods, and also Allen-Zhu (2018) among SGD methods, by at least a logarithmic factor in $1/\varepsilon$ [Theorem 1.1.B].[7] For the **LSN** case the complexity seems new to the community; it is strictly better than the state-of-the-art complexity result as in (Nguyen et al., 2019) (the "individually nonconvex" case) in its dominating term after a metric conversion (gradient norm to iteration distance).

It is worth mentioning that our method has a complexity that is better than the prevailing multi-epoch methodology that *restarts whenever halved* by a logarithmic factor, given σ_* is bounded away from zero.[8] The disadvantage is that the system

[7] To the best of our knowledge, the state-of-the-art rate in squared gradient norm is achieved by the multi-epoch Inexact SARAH algorithm (Nguyen et al., 2021), which is superior to SGD (Nguyen et al., 2019) and SCSG (Lei & Jordan, 2017) after a gradient-iterate bound conversion.

[8] To briefly describe this halving method, the sequence $G_k^2 := G_{k-1}^2/2$, for each $k \geq 1$, so that $G_{k-1}^2 := \|\nabla F(\theta_0)\|_2^2 2^{1-k}$ where $\|\nabla F(\theta_0)\|_2^2 := \|\nabla F(\theta_0)\|_2^2$. Thus, the second moment bound G_k^2 on $\mathbb{E}\|\nabla F(\theta_{\Delta_k})\|_2^2$ is chosen as half of the corresponding bound in the previous loop. We further pin down the number of loops for ROOT-SGD to achieve an upper bound of ε^2 for the second

need explicit knowledge of the total number N of samples in advance to tune the parameters for epoch lengths (to minimize the complexity).

1.2.3 Improved nonasymptotic results in the LSN case

We recall that in Theorem 1.1.A, an upper bound on the gradient norm is proved consisting of a nonimprovable $\frac{\sigma_*^2}{T}$ term and a polynomially decaying mixing term. Theorem 1.1.A serves as a starting point of our improved analysis.

Throughout this section and the next asymptotic section, we continue to make the previous Assumptions 5, 14 and 7, i.e. we are in the case of **LSN**. The strong convexity and smoothness Assumption 5 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 7, 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^2 that scales inversely proportional to the batch-size. This case is called "Lipschitz stochastic noise" (**LSN**), which often requires weaker conditions than the "individual smooth and convex" (**ISC**) case in §1.2.2.

We focus on the **LSN** case, in which case the unified upper bound (1.16) has a non-unity pre-factor in the leading term $\frac{\sigma_*^2}{T}$. We will nevertheless use this bound as a starting point towards the sharp inequalities with the constant being unity. Now we are ready to present our first main result, where more careful estimates yield the following improved upper bound of the ROOT-SGD algorithm.

Theorem 1.2.A (Improved nonasymptotic result, single-epoch ROOT-SGD). *For F satisfying Assumptions 5, 14 and 7, the estimator produced by Algorithm 5 satisfies the following convergence rate upper bound:*

$$\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 - \frac{\sigma_*^2}{T} \le \left(\frac{947}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{636\log\left(\frac{eT}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{84}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2. \tag{1.21}$$

Here in (1.21), the leading-order term is the optimal statistical risk $\frac{\sigma_*^2}{T}$ for a constant step-size that only decays slowly with T. For multi-epoch algorithm, we con-

moment as follows: $K := \left\lceil \log_2\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\varepsilon^2}\right)\right\rceil$. We further define the timestamps Δ_k as $\Delta_0 := 0$, $\Delta_k - \Delta_{k-1} := \max\left\{\frac{105}{\eta\mu}, \frac{112\sigma_*^2}{G_{k-1}^2}\right\}$. We can now appeal repeatedly to Theorem 1.1.A to obtain the complexity of ROOT-SGD with restarting $C(\varepsilon) \asymp \frac{1}{\eta_{max}\mu}\log\left(\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}\right) \vee \frac{\sigma_*^2}{\varepsilon^2}$, which is worse than our complexity (1.20) by a logarithmic factor when ε is smaller in magnitude than $\eta_{max}\mu\sigma_*^2$.

tinue to use the short/long epoch length idea and improve the lower-order term to a sharp polynomially decaying one and conclude the following result:

Theorem 1.2.B (Improved nonasymptotic result, multi-epoch ROOT-SGD). *For F satisfying Assumptions 5, 14 and 7, and let the number of short epochs $\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$, the burn-in time $\mathscr{B} = \frac{24}{\eta\mu}$, and the small epoch length $T_b(\eta) = \frac{7340}{\eta\mu}$. Then the multi-epoch estimator produced by Algorithm 6 satisfies the following bound: for $N \geq T_b(\eta)\mathscr{E} + 1$*

$$\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 - \frac{\sigma_*^2}{T} \leq \left(\frac{1192}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T}, \tag{1.22}$$

where $T := N - T_b(\eta)\mathscr{E}$. By further choosing the step-size as $\eta = \eta(N) = \frac{0.49\mu^{1/3}}{\ell_\Xi^{4/3}N^{1/3}} \wedge \frac{1}{4L}$, we obtain

$$\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 - \frac{\sigma_*^2}{N} \leq \left(\frac{\left(7352\sqrt{\mathscr{E}} + \log\left(\frac{N}{\mathscr{B}}\right)\right)\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \frac{(9879\sqrt{\mathscr{E}})L^{1/2}}{\mu^{1/2}N^{1/2}}\right)\frac{\sigma_*^2}{N}. \tag{1.23}$$

See §1.3.3 and §1.3.4 for the proof of the above two theorems, separately.

Despite that the pre-factor in the leading-order rate of $\frac{\sigma_*^2}{N}$ is asymptotically unity, the bound in Theorem 1.2.A has an additional second-order term decaying polynomially at $\tilde{O}\left(\frac{\ell_\Xi^{2/3}}{N^{4/3}} + \frac{L^{1/2}}{N^{3/2}}\right)$, which is the first result that achieves this additional lower-order term in the **LSN** case. Furthermore, this final rate achieves the optimal statistical risk with a prefactor that is asymptotically unity, while enjoying a logarithmic dependency on the initialization $\|\nabla F(\theta_0)\|_2^2$, which is a desirable property for stochastic optimization algorithm.

1.2.4 Asymptotic efficiency

In this subsection we study the asymptotic efficiency of our ROOT-SGD algorithm. We continue to study the multi-epoch specifications in Theorem 1.2.B where the **LSN** case is in force. In this case, Assumptions 5, 14 and 7 are the standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see e.g. van der Vaart (2000, Theorem 5.21)). Under this setup, we are ready to state our weak convergence asymptotic normality result in the triangular-array fashion, as in the following theorem, whose proof is provided in §1.3.5:

Theorem 1.3 (Asymptotic normality). *Under Assumptions 5, 14 and 7, for any $\alpha \in (0,1)$, the multi-epoch estimator produced by Algorithm 6 with burn-in time*

$\mathscr{B} = \frac{24}{\eta\mu}$, short-epoch length $T_b = \frac{7340}{\eta\mu}$ and number of short epoches $\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$. Then as $N - T_b\mathscr{E} \to \infty$, $\eta \to 0$ such that $\eta(N - T_b\mathscr{E}) \to \infty$ the following weak convergence holds:[9]

$$\sqrt{T}\nabla F(\theta_N^{final,(\eta)}) \xrightarrow{d} \mathscr{N}(0, \Sigma^*), \tag{1.24}$$

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.

Due to the exchangeability of the expectation and trace operators, we observe $\mathrm{Tr}(\Sigma^*) = \sigma_*^2$ and hence Theorem 3.1 establishes asymptotic normality with covariance matrix that matches the optimal statistical risk as presented in its nonasymptotic counterpart, Theorem 1.2.B. Specially, this result does not require any Hessian continuity assumption at θ^* which has been another standard condition for asymptotic normality. To our best knowledge, this provides the first known optimal asymptotic efficiency result for stochastic first-order algorithms without posing any continuity assumptions to Hessians. We remark that in Theorem 3.1, we are adopting the multi-epoch ROOT-SGD with fixed the constant step-size that scales down with N, where the scaling condition is essentially $\eta \to 0, N \to \infty$ with $\eta N - \log(\eta^{-1}) \to \infty$, which is always satisfied when $\eta \asymp \frac{1}{N^\alpha}$ with $\alpha \in (0,1)$. We also observe that the asymptotic covariance of Σ^* in terms of the gradient matches the optimal asymptotic risk but carries significantly more information. As a direct corollary to Theorem 3.1 (proof omitted), if we add upon an additional one-point Hessian continuity condition that at the minimizer θ^* of

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

then the above Theorem 3.1 implies the following weak convergence result:

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

which is the standard Cramér-Rao asymptotic covariance in standard asymptotic statistics literature (van der Vaart, 2000; van der Vaart & Wellner, 1996) and matches the optimal rates achieved by stochastic approximation rates (Kushner & Yin, 2003; Polyak & Juditsky, 1992; Ruppert, 1988).

1.3 Proofs of general nonasymptotic and asymptotic results

We provide in this section the convergence rate analysis proofs of our nonasymptotic results, in particular the proofs of our theorems in this section. Theorem 1.1.A and Theorem 1.1.B are discussed in §1.3.1 and §1.3.2, respectively. For improved

[9] Here we explicitly write out the dependency on η in the superscript in (1.24).

nonasymptotic rate for the **LSN** case, §1.3.3 and §1.3.4 study the problem, separately. Finally we prove the asymptotic efficiency in §1.3.5.

1.3.1 Proof of Theorem 1.1.A

This subsection is devoted to the proof of Theorem 1.1.A. It is straightforward to show first (1.16) automatically holds for $T < \mathscr{B}$, since $\mathbb{E}\|\nabla F(\theta_T)\|_2^2 = \mathbb{E}\|\nabla F(\theta_0)\|_2^2$, so we only need to prove the result for $T \geq \mathscr{B}$.

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

$$\omega_{max} := \begin{cases} \frac{2\ell_\Xi^2}{\mu^2}, & \text{for } \textbf{LSN} \text{ case}, \\ \frac{2\ell_{max}}{\mu}, & \text{for } \textbf{ISC} \text{ case}. \end{cases} \tag{1.25}$$

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

$$z_t := v_t - \nabla F(\theta_{t-1}), \qquad \text{for } t \geq \mathscr{B}. \tag{1.26}$$

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 (3.5), 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

$$\begin{aligned} tz_t = t\left(v_t - \nabla F(\theta_{t-1})\right) &= \varepsilon_t(\theta_{t-1}) + (t-1)(v_{t-1} - \nabla F(\theta_{t-2})) + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) \\ &= \varepsilon_t(\theta_{t-1}) + (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})). \end{aligned} \tag{1.27}$$

Unwinding this relation recursively yields the decomposition

$$tz_t - \mathscr{B}z_{\mathscr{B}} = \sum_{s=\mathscr{B}+1}^{t} \varepsilon_s(\theta_{s-1}) + \sum_{s=\mathscr{B}+1}^{t} (s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})). \tag{1.28}$$

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 1.4 (Recursion involving z_t). *Under the conditions of Theorem 1.1.A, for all $t \geq \mathscr{B}+1$, we have*

$$t^2\mathbb{E}\|z_t\|_2^2 \leq (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + 2(t-1)^2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2. \tag{1.29a}$$

On the other hand, for $t = \mathscr{B}$, we have

$$\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 - \mathscr{B}^2\mathbb{E}\|\nabla F(\theta_0)\|_2^2 = \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 = \mathscr{B}\mathbb{E}\|\varepsilon_{\mathscr{B}}(\theta_0)\|_2^2. \tag{1.29b}$$

See §1.7.1 for the proof of this claim. Note we have $z_{\mathscr{B}} = v_{\mathscr{B}} - \nabla F(\theta_0)$ which is simply the arithmetic average of $\mathscr{B}$ i.i.d. noise terms at θ_0, $\varepsilon_1(\theta_0), \ldots, \varepsilon_{\mathscr{B}}(\theta_0)$.
Our next auxiliary lemma characterizes the evolution of the sequence $(v_t : t \geq \mathscr{B})$ in terms of the quantity $\mathbb{E}\|v_t\|_2^2$.

Lemma 1.5 (Evolution of v_t). *Under the settings of Theorem 1.1.A, for any $\eta \in (0, \eta_{max}]$, we have*

$$t^2\mathbb{E}\|v_t\|_2^2 - 2t\mathbb{E}\langle v_t, \nabla F(\theta_{t-1})\rangle + \mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^2, \quad (1.30a)$$

and

$$\mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^2 \leq (1-\eta\mu)\cdot(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 \\ - 2(t-1)^2\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2, \quad (1.30b)$$

for all $t \geq \mathscr{B}+1$.

See §1.7.2 for the proof of this claim.

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

Lemma 1.6 (Second moment of pointwise stochastic noise). *Under the conditions of Theorem 1.1.A, we have*

$$\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 \leq \omega_{max}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2, \qquad \text{for all } t \geq \mathscr{B}+1. \quad (1.31)$$

See §1.7.3 for the proof of this claim.

Equipped with these three auxiliary results, we are now ready to prove Theorem 1.1.A.

Proof (Proof of Theorem 1.1.A).

(i) 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

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

Taking the expectation of both sides and applying the bound (1.30a) from Lemma 1.5 yields

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

Moreover, since we have $\eta \leq \eta_{max} \leq \frac{1}{4\mu}$ under either condition (1.14a) or condition (1.14b), we can multiply both sides by $(1-\eta\mu)^{-1}$, which lies in $[1, \frac{3}{2}]$. Doing so yields the bound

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

Combining with the bound (1.29a) from Lemma 1.4 gives

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

By telescoping this inequality from $\mathscr{B}+1$ to T, we find that

$$\begin{aligned} T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2 - \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 - \mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 \\ \leq \sum_{t=\mathscr{B}+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right]. \quad (1.32) \end{aligned}$$

Next, applying the result (1.29b) from Lemma 1.4 yields

$$\begin{aligned} &\frac{T^2}{2}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2 \\ &\leq \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 + \mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 + \sum_{t=\mathscr{B}+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right] \\ &= \mathscr{B}^2\|\nabla F(\theta_0)\|_2^2 + 2\mathscr{B}\mathbb{E}\|\varepsilon_{\mathscr{B}}(\theta_0)\|_2^2 + 5\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2. \end{aligned}$$

Following some algebra, we find that

$$\begin{aligned} \mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq \frac{2\mathscr{B}^2\|\nabla F(\theta_0)\|_2^2 + 4\mathscr{B}\mathbb{E}\|\varepsilon_{\mathscr{B}}(\theta_0)\|_2^2}{T^2} \\ + \frac{10}{T^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{2}{\eta\mu T^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2. \quad (1.33) \end{aligned}$$

Combining inequality (1.33) with the bound (1.31) from Lemma 1.6 gives

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

$$\leq \frac{(4\omega_{max}+2\mathscr{B})\mathscr{B}\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2}+\frac{10\omega_{max}+2\mu^{-1}\eta^{-1}}{T^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+\frac{20\sigma_*^2}{T},$$

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

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq \frac{1}{T^2}\left\{\alpha_1\mathbb{E}\|\nabla F(\theta_0)\|_2^2+\alpha_2\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right\}+\frac{20\sigma_*^2}{T}, \tag{1.34}$$

where $\alpha_1 := (4\omega_{max}+2\mathscr{B})\,\mathscr{B}$ and $\alpha_2 := 10\omega_{max}+\frac{2}{\eta\mu}$.

(ii) Based on this lemma, the proof of Theorem 1.1.A relies on a bootstrapping argument in order to remove the dependence of the right-hand side of Eq. (1.34) on the quantity $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$. Let $T^* \geq \mathscr{B}+1$ be arbitrary. Telescoping the bound (1.34) over the iterates $T = \mathscr{B}+1,\ldots,T^*$ yields

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

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

$$\sum_{T=\mathscr{B}+1}^{T^*}\frac{1}{T^2}\overset{(i)}{\leq}\int_{\mathscr{B}}^{T^*}\frac{d\tau}{\tau^2}\leq\frac{1}{\mathscr{B}}, \qquad \text{and} \qquad \sum_{T=\mathscr{B}+1}^{T^*}\frac{1}{T}\overset{(ii)}{\leq}\int_{\mathscr{B}}^{T^*}\frac{d\tau}{\tau}=\log\Big(\frac{T^*}{\mathscr{B}}\Big). \tag{1.35}$$

We clearly have

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

Moreover, by using the fact that $T^* \geq T$, interchanging the order of summation, and then using inequality (1.35)(i) again, we have

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

Finally, turning to the third quantity, we have $Q_3 \leq 20\sigma_*^2\log\big(\frac{T^*}{\mathscr{B}}\big)$, where we have used inequality (1.35)(ii). Putting together the pieces yields the upper bound

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

Eqs. (1.14a) and (1.14b) imply that, for either setting 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 (1.15) that $\mathscr{B} = \left\lceil \frac{24}{\eta\mu} \right\rceil \leq \frac{25}{\eta\mu}$, resulting in

$$4\omega_{max} + 2\mathscr{B} \leq \frac{4}{\eta\mu} + 2\left(\frac{25}{\eta\mu}\right) = \frac{54}{\eta\mu},$$

where we have the choice of burn-in time $\mathscr{B}$ from Eq. (1.15). Similarly, we have $\alpha_2 = 10\omega_{max} + \frac{2}{\eta\mu} \leq \frac{12}{\eta\mu} \leq \frac{\mathscr{B}}{2}$. Putting together the pieces yields

$$\frac{1}{2}\sum_{t=\mathscr{B}+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \leq \frac{54}{\eta\mu}\mathbb{E}\|\nabla F(\theta_0)\|_2^2 + 20\sigma_*^2 \log\left(\frac{T^*}{\mathscr{B}}\right). \tag{1.36}$$

Now substituting the inequality (1.36) back into the earlier bound (1.34) with $T^* = T$ allows us to obtain a bound on $\mathbb{E}\|\nabla F(\theta_{T-1})\|_2$. In particular, for any $T \geq \mathscr{B}+1$, we have

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

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

$$\begin{aligned}
\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 &\leq \frac{2(54)\mathscr{B}}{\eta\mu}\cdot\frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{20\sigma_*^2}{T}\left[1 + \frac{\mathscr{B}}{T}\log\left(\frac{T}{\mathscr{B}}\right)\right] \\
&\leq \frac{108}{\eta\mu}\cdot\frac{25}{\eta\mu}\cdot\frac{\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{T^2} + \frac{20\sigma_*^2}{T}\left[1 + \frac{1}{e}\right] \\
&\leq \frac{2700\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{28\sigma_*^2}{T}.
\end{aligned}$$

Shifting the subscript forward by one yields Theorem 1.1.A.

1.3.2 Proof of Theorem 1.1.B

We present the proof of the unified multi-epoch result, Theorem 1.1.B. Invoking Eq. (1.16) in Theorem 1.1.A, we obtain for $b = 1, 2, \ldots, \mathscr{E}$ the following bound for $T_b(\eta) = \sqrt{2700}\mathscr{A}\mu^{-1}\eta^{-1}$, for some $\mathscr{A} \geq \sqrt{2}$ to be determined:

$$\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2 \leq \frac{28\sigma_*^2}{T_b(\eta)} + \frac{2700\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2}{\eta^2\mu^2 T_b(\eta)^2} \leq \frac{28\sigma_*^2}{T_b(\eta)} + \mathscr{A}^{-2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2.$$

Solving the recursion, we arrive at the bound:

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2 &\leq \frac{28\sigma_*^2}{T_b(\eta)}\sum_{n=1}^{\mathscr{E}}\mathscr{A}^{2-2n} + \mathscr{A}^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2 \\ &\leq \frac{28\sigma_*^2}{\sqrt{2700}(1-\mathscr{A}^{-2})\mathscr{A}\mu^{-1}\eta^{-1}} + \mathscr{A}^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2 \\ &\leq \mathscr{A}^{-1}\eta\mu\sigma_*^2 + \mathscr{A}^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2.\end{aligned}$$

To make the above $\leq 2\mathscr{A}^{-1}\eta\mu\sigma_*^2$ we need $\mathscr{A}^{1-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2 \leq \eta\mu\sigma_*^2$ so we simply set $\mathscr{E} = \left\lceil \frac{\log\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)}{2\log\mathscr{A}} + \frac{1}{2} \right\rceil$. Since $T_b(\eta)$ is proportional to $\mathscr{A}$ it is easy to verify that the choice of $\mathscr{A} = e$ is such that $\min_{\mathscr{A}>1}\frac{\mathscr{A}}{\log\mathscr{A}} = e$, and hence the total number iterates of the small epochs $T_b(\eta)\mathscr{E}$ is *approximately* minimized.[10] Substituting this initial condition into the bound (1.21), we obtain the final bound

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 = \mathbb{E}\|\nabla F(\theta_T^{(\mathscr{E}+1)})\|_2^2 &\leq \frac{28\sigma_*^2}{T} + \frac{2700\|\nabla F(\theta_0^{(\mathscr{E}+1)})\|_2^2}{\eta^2\mu^2T^2} \\ &\leq \frac{28\sigma_*^2}{T} + \frac{2700(2e\eta\mu\sigma_*^2)}{\eta^2\mu^2T^2} \leq \frac{132\sigma_*^2}{T}.\end{aligned}$$

Substituting $T = N - T_b(\eta)\mathscr{E}$, we finally conclude (1.22) and the whole theorem.

1.3.3 Proof of Theorem 1.2.A

In this subsection we prove Theorem 1.2.A. Inherited from §1.3.1 we start from the gradient decomposition (2.14) and provide a refined analysis of the quantities v_t and z_t. From Theorem 1.1.A we can derive *sharp* bounds for v_t and z_t as follows:

[10] Here by *approximately*, we omit the additive constant or the ceiling effect in the epoch number $\mathscr{E}$.

Lemma 1.7 (Sharp bound on v_t). *Under the setting of Theorem 1.2.A we have the following bound for $T \geq \mathscr{B}+1$*

$$\mathbb{E}\,\|v_T\|_2^2 \leq \frac{752\sigma_*^2}{\eta\mu T^2} + \frac{69175}{\eta^4\mu^4 T^4}\,\|\nabla F(\theta_0)\|_2^2. \tag{1.37}$$

We defer the proof of Lemma 1.7 to §1.8.1. Note that the dependency of $\mathbb{E}\|v_T\|_2^2$ on $\|\nabla F(\theta_0)\|_2^2$ decays polynomially at rate $T^{-4}\eta^{-4}$ (we only write out the dependency on η and T), which is critical for our step-size scaling condition in the asymptotic result.

Now we turn to the tracking error z_t. However, owing to the inherent martingale structure in the process z_t, one could extract the main part of the variance and bound the additional part using Theorem 1.1.A. The multiplicative constant in such bounds will only contribute to the high-order terms in the final bound. See Theorem 1.2.A and its proofs for details.

For the process z_t, we have the following lemma which leads to an $(1+o(1))\frac{\sigma_*}{\sqrt{t}}$ bound with the exact unity in the constant prefactor. We highlight that this fine-grained analysis is tight in the $\ell_\Xi \to 0$ regime.

Lemma 1.8 (Sharp bound on z_t). *Under settings of Theorem 1.2.A, the following bounds hold true for $T \geq \mathscr{B}+1$ and η satisfying $\eta \in (0, \eta_{max}]$*

$$\begin{aligned}\mathbb{E}\,\|z_T\|_2^2 - \frac{\sigma_*^2}{T} \leq &\left(\frac{20\ell_\Xi^2\eta}{\mu} + \frac{12\ell_\Xi}{\mu\sqrt{T}} + \frac{504\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} \\ &+ \frac{9\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\,\|\nabla F(\theta_0)\|_2 + \frac{183\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\,\|\nabla F(\theta_0)\|_2^2.\end{aligned} \tag{1.38}$$

Very roughly, we have from (1.38) that[11]

$$\mathbb{E}\,\|z_T\|_2^2 - \frac{\sigma_*^2}{T} \lesssim \frac{\ell_\Xi^2\eta}{\mu}\cdot\frac{\sigma_*^2}{T} + \frac{\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\,\|\nabla F(\theta_0)\|_2^2.$$

Especially the right hand is of $o(1)$ if we let $\ell_\Xi \to 0$. Earlier analysis largely upper-bound $\frac{\ell_\Xi^2\eta}{\mu}$ by 1. Note when applying Young's inequality to the middle term, it cannot be absorbed into the rest terms as it will incur an additional $\frac{L\sigma_*^2}{\mu T^2}$ term.

The proof of Lemma 1.8 is deferred to §1.8.2. With Lemmas 1.7 and 1.8 prepared along with a sharp estimate on the cross term, we turn to prove our sharp upper bound result on the gradient norm, Theorem 1.2.A.

Proof (Proof of Theorem 1.2.A). We first establish the results for the single-loop algorithm, and then use it to prove the results for the multi-epoch one. We start by observing the following decomposition:

[11] Rigorously speaking, the raw bound (1.16) can be better than the current bound in some corner cases, due to its an extra logarithmic factor. Our bounds guarantees a characterization of the lower-order dependency.

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^2 &= \mathbb{E}\left\|z_T\right\|_2^2 + \mathbb{E}\left\|v_T\right\|_2^2 - 2\mathbb{E}\left\langle z_T, v_T\right\rangle \\ &\leq \left(1+\frac{c}{\sqrt{\eta\mu T}}\right)\mathbb{E}\left\|z_T\right\|_2^2 + \left(1+\frac{\sqrt{\eta\mu T}}{c}\right)\mathbb{E}\left\|v_T\right\|_2^2, \end{aligned} \tag{1.39}$$

where the constant $c \geq 1$ and we applied Young's inequality. Here the weights are carefully selected for both v_T and z_T to best reduce the constant in the $O(\sigma_*^2)$-term. For $T \geq \mathscr{B} = \frac{24}{\eta\mu}$, we have $1+\frac{c}{\sqrt{\eta\mu T}} \leq \frac{5c}{4}$ which is $\sqrt{\eta\mu T}+c \leq \frac{5c}{4}\sqrt{\eta\mu T}$, then for z_T, we have

$$\begin{aligned}&\left(1+\frac{c}{\sqrt{\eta\mu T}}\right)\mathbb{E}\left\|z_T\right\|_2^2 \\ \leq &\left(1+\frac{c}{\sqrt{\eta\mu T}}\right)\frac{\sigma_*^2}{T} + \left(\frac{(\frac{5c}{4})(20)\ell_\Xi^2\eta}{\mu} + \frac{(\frac{5c}{4})(12)\ell_\Xi}{\mu\sqrt{T}} + \frac{(\frac{5c}{4})(504)\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} \\ &+ \frac{(\frac{5c}{4})(9)\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2 + \frac{(\frac{5c}{4})(183)\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2 \\ \leq &\left(1+\frac{7c}{\sqrt{\eta\mu T}}\right)\frac{\sigma_*^2}{T} + \left(\frac{(\frac{5c}{4})(20)\ell_\Xi^2\eta}{\mu} + \frac{(\frac{5c}{4})(504)\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} \\ &+ \frac{(\frac{5c}{4})(9)\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2 + \frac{(\frac{5c}{4})(183)\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2,\end{aligned}$$

where the second inequality above comes from

$$\frac{(\frac{5c}{4})(12)\ell_\Xi}{\mu\sqrt{T}} \leq \frac{6c}{\sqrt{\eta\mu T}} \qquad \text{which is } \eta \leq \left(\frac{6}{15\sqrt{\mu T}}\cdot\frac{\mu\sqrt{T}}{\ell_\Xi}\right)^2 = \frac{4\mu}{25\ell_\Xi^2}.$$

On the other hand, for v_T we have

$$\left(\sqrt{\eta\mu T}+c\right)\mathbb{E}\left\|v_T\right\|_2^2 \leq \sqrt{\eta\mu T}\left(\frac{(\frac{5c}{4})(752)\sigma_*^2}{\eta\mu T^2} + \frac{(\frac{5c}{4})(69175)}{\eta^4\mu^4 T^4}\left\|\nabla F(\theta_0)\right\|_2^2\right).$$

Adding up the above two bounds into the decomposition Eq. (1.39) and merging the terms of the same order, we have by taking $c=1$

$$\begin{aligned}&\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^2 \\ \leq &\left(1+\frac{c}{\sqrt{\eta\mu T}}\right)\mathbb{E}\left\|z_T\right\|_2^2 + \left(1+\frac{\sqrt{\eta\mu T}}{c}\right)\mathbb{E}\left\|v_T\right\|_2^2 \\ \leq &\left(1+\frac{7c}{\sqrt{\eta\mu T}}\right)\frac{\sigma_*^2}{T} + \left(\frac{(\frac{5c}{4})(20)\ell_\Xi^2\eta}{\mu} + \frac{(\frac{5c}{4})(504)\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{(\frac{5c}{4})(9)\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2 \\ &+ \frac{(\frac{5c}{4})(183)\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2 + \sqrt{\eta\mu T}\left(\frac{(\frac{5}{4})(752)\sigma_*^2}{\eta\mu T^2} + \frac{(\frac{5}{4})(69175)}{\eta^4\mu^4 T^4}\left\|\nabla F(\theta_0)\right\|_2^2\right)\end{aligned}$$

$$\leq \left(1+\frac{947}{\sqrt{\eta\mu T}}+\frac{(\frac{5}{4})(20)\ell_\Xi^2\eta}{\mu}+\frac{(\frac{5}{4})(504)\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T}$$
$$+\frac{12\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2+\frac{60}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2.$$

Finally, we conclude

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2-\frac{\sigma_*^2}{T} &\leq \left(\frac{947}{\sqrt{\eta\mu T}}+\frac{25\ell_\Xi^2\eta}{\mu}+\frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T}\\ &\quad+\frac{12\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2+\frac{60}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2\\ &\leq \left(\frac{947}{\sqrt{\eta\mu T}}+\frac{25\ell_\Xi^2\eta}{\mu}+\frac{636\log\left(\frac{eT}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T}+\frac{84}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2,\end{aligned} \tag{1.40}$$

where we absorb the cross term in its left and right neighborhood terms so that $\frac{2\ell_\Xi\sigma_*}{\mu T}\cdot\frac{\mathscr{B}}{T}\left\|\nabla F(\theta_0)\right\|_2\leq\frac{6\ell_\Xi^2\sigma_*^2}{\mu^2T^2}+\frac{\mathscr{B}^2}{6T^2}\left\|\nabla F(\theta_0)\right\|_2^2$, which proves the first claim (1.21) by shifting the subscript by 1. Theorem 1.2.A is hence concluded.

1.3.4 Proof of Theorem 1.2.B

Now we turn to the proof of the improved multi-epoch result, Theorem 1.2.B.

(i) Invoking Eq. (1.16) in Theorem 1.1.A, we obtain for $b=1,2,\ldots,\mathscr{E}$ the following bound for $T_b(\eta)=\sqrt{2700}\mathscr{A}\mu^{-1}\eta^{-1}$ where $\mathscr{A}\geq\sqrt{2}$:

$$\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2\leq\frac{28\sigma_*^2}{T_b(\eta)}+\frac{2700\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2}{\eta^2\mu^2T_b(\eta)^2}\leq\frac{28\sigma_*^2}{T_b(\eta)}+\mathscr{A}^{-2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2.$$

Solving the recursion, we arrive at the bound:

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2 &\leq\frac{28\sigma_*^2}{T_b(\eta)}\sum_{n=1}^{\mathscr{E}}\mathscr{A}^{2-2n}+\mathscr{A}^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2\\ &\leq\frac{28\sigma_*^2}{(1-\mathscr{A}^{-2})\sqrt{2700}\mathscr{A}\mu^{-1}\eta^{-1}}+\mathscr{A}^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2.\end{aligned}$$

Similar to our approach as in the proof of Theorem 1.1.B in §1.3.2, we continue to choose $\mathscr{A}=e$ so the total complexity is minimized in an approximate sense. We also continue to choose the number of short epochs as

$$\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\,\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil.$$

This altogether indicates

$$\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2 \le e^{-1}\mu\eta\sigma_*^2 + e^{-2\mathscr{E}}\,\|\nabla F(\theta_0)\|_2^2 \le 2e^{-1}\mu\eta\sigma_*^2.$$

Substituting this initial condition into the bound (1.21) in Theorem 1.2.A, we obtain the final bound

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_T^{(\mathscr{E}+1)})\right\|_2^2 - \frac{\sigma_*^2}{T} \\
&\le \left(\frac{947}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} \\
&\quad + \frac{12\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2 + \frac{60}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2 \\
&\le \left(\frac{947}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} \\
&\quad + \frac{12\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot\sqrt{2e^{-1}\mu\eta\sigma_*^2} + \frac{60}{\eta\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot 2e^{-1}\mu\eta\sigma_*^2 \\
&= \left(\frac{947}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{(60)(2e^{-1})\mathscr{B}\sigma_*^2}{T^2} + \frac{(12)\mathscr{B}\sigma_*^2}{T^2}\cdot\sqrt{\frac{(2e^{-1})\ell_\Xi^2\eta}{\mu}} \\
&\le \left(\frac{947}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{48\mathscr{B}\sigma_*^2}{T^2}.
\end{aligned}$$

Noticing $\frac{48\mathscr{B}\sigma_*^2}{T^2} \le \frac{1192-947}{\sqrt{\eta\mu T}}\cdot\frac{\sigma_*^2}{T}$, we finally conclude

$$\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{T} = \mathbb{E}\left\|\nabla F(\theta_T^{(\mathscr{E}+1)})\right\|_2^2 - \frac{\sigma_*^2}{T} \le \left(\frac{1192}{\sqrt{\eta\mu T}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T},$$

which proves (1.22).

(ii) Choosing the step-size as $\eta = \frac{C\mu^{1/3}}{\ell_\Xi^{4/3}T^{1/3}} \wedge \frac{1}{4L}$ with $C = 0.49 \le \left(\frac{\sqrt{8}}{24}\right)^{1/3}$, we are able to both

(a) Verify in a straightforward fashion that our choice of η automatically satisfies $\eta \le \eta_{max} \le \frac{\mu}{8\ell_\Xi^2}$, and

(b) Balance the two terms of orders $\frac{1}{\sqrt{\eta\mu T}}$ and $\frac{\ell_\Xi^2\eta}{\mu}$ in (1.22).

Now to justify the convergence rate bound (1.22), we apply the elementary inequality that for $N \geq 2T_b(\eta)\mathscr{E}$, $\alpha \in (0,1]$,[12]

$$T^{-\alpha} = N^{-\alpha}\left(1 - \frac{T_b(\eta)\mathscr{E}}{N}\right)^{-\alpha} \leq N^{-\alpha}\left(1 + 2\alpha \cdot \frac{T_b(\eta)\mathscr{E}}{N}\right) \leq 2N^{-\alpha},$$

where from (1.21) we have $T = N - T_b(\eta)\mathscr{E}$. Therefore (1.22) gives

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 &\leq \left[1 + 2\left(\frac{1192}{\sqrt{\eta\mu N}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right]\frac{\sigma_*^2}{T} \\
&\leq \left[1 + 2\left(\frac{1192}{\sqrt{\eta\mu N}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right]\left(1 + \frac{2T_b(\eta)\mathscr{E}}{N}\right)\frac{\sigma_*^2}{N} \\
&\leq \left(1 + \frac{2T_b(\eta)\mathscr{E}}{N} + 4\left(\frac{1192}{\sqrt{\eta\mu N}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right)\frac{\sigma_*^2}{N} \\
&= \left(1 + \frac{(2)(7340)\mathscr{E}}{\eta\mu N} + 4\left(\frac{1192}{\sqrt{\eta\mu N}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right)\frac{\sigma_*^2}{N}.
\end{aligned}$$

By taking $\eta = \frac{C\mu^{1/3}}{\ell_\Xi^{4/3}N^{1/3}} \wedge \frac{1}{4L}$ with $C = 0.49$, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 &\leq \left(1 + \frac{(2)(7340)\mathscr{E}}{\eta\mu N} + 4\left(\frac{1192}{\sqrt{\eta\mu N}} + \frac{25\ell_\Xi^2\eta}{\mu} + \frac{630\log\left(\frac{N}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right)\frac{\sigma_*^2}{N} \\
&= \left(1 + \frac{(2)(7340)\mathscr{E}\ell_\Xi^{4/3}}{C\mu^{4/3}N^{2/3}} + \frac{(2)(7340)(4)L\mathscr{E}}{\mu N}\right. \\
&\quad \left. + 4\left(\frac{1192\ell_\Xi^{2/3}}{\sqrt{C\mu^{4/3}N^{2/3}}} + \frac{(2)(1192)\sqrt{L}}{\sqrt{\mu N}} + \frac{25C\ell_\Xi^{2/3}}{N^{1/3}\mu^{2/3}} + \frac{630\log\left(\frac{N}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\right)\frac{\sigma_*^2}{N} \\
&\leq \left(1 + (4)\left(\frac{1192}{\sqrt{C}} + 25C\right)\frac{\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \frac{(2)(7340)(4)L\mathscr{E}}{\mu N} + \frac{(2)(4)(1192)\sqrt{L}}{\sqrt{\mu N}}\right. \\
&\quad \left. + \frac{(2)(7340)\mathscr{E}\ell_\Xi^{4/3}}{C\mu^{4/3}N^{2/3}} + \frac{(630)(4)\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\frac{\sigma_*^2}{N} \\
&\leq \left(1 + \frac{6862\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \frac{29960\mathscr{E}\ell_\Xi^{4/3}}{\mu^{4/3}N^{2/3}} + \frac{2520\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N} + \frac{9536\sqrt{L}}{\sqrt{\mu N}} + \frac{58720L\mathscr{E}}{\mu N}\right)\frac{\sigma_*^2}{N}.
\end{aligned}$$

From the fact that $N \geq T_b\mathscr{E} = \frac{7340\mathscr{E}}{\eta\mu}$, we derive $\frac{1}{N} \leq \frac{\eta\mu}{7340\mathscr{E}}$. Further by recalling that $\eta = \frac{C\mu^{1/3}}{\ell_\Xi^{4/3}N^{1/3}} \wedge \frac{1}{4L}$, the above inequality leads to

[12] By Bernoulli's inequality, $(1+t)^\beta \leq 1 + \beta t$ holds for each $\beta \in (0,1]$ and $t \geq -1$, so for $|x| \leq \frac{1}{2}$ and such β, $(1-x)^{-\beta} \leq (1+2x)^\beta \leq 1 + 2\beta x \leq 2$.

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{N} &\le \left(\frac{6862\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \frac{29960\mathscr{E}\ell_\Xi^{4/3}}{\mu^{4/3}N^{2/3}} + \frac{2520\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N} + \frac{9536\sqrt{L}}{\sqrt{\mu N}} + \frac{58720L\mathscr{E}}{\mu N}\right)\frac{\sigma_*^2}{N} \\
&\le \left(\frac{6862\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \frac{29960\mathscr{E}\ell_\Xi^{4/3}}{\mu^{4/3}N^{1/3}}\cdot\frac{C^{1/2}\mu^{2/3}}{(7340\mathscr{E})^{1/2}\ell_\Xi^{2/3}} + \frac{2520\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N^{1/3}}\cdot\frac{C\mu^{4/3}}{7340\mathscr{E}\ell_\Xi^{4/3}}\right. \\
&\quad \left. + \frac{9536\sqrt{L}}{\sqrt{\mu N}} + \frac{58720L\mathscr{E}}{\mu N^{1/2}}\cdot\frac{\sqrt{\mu}}{\sqrt{(4)(7340)}\sqrt{\mathscr{E}L}}\right)\frac{\sigma_*^2}{N} \\
&\le \left(\left(6862 + \frac{29960\sqrt{\mathscr{E}C}}{\sqrt{7340}} + \frac{2520C\log\left(\frac{T}{\mathscr{B}}\right)}{7340\mathscr{E}}\right)\frac{\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}}\right. \\
&\quad \left. + \left(9536 + \frac{58720\sqrt{\mathscr{E}}}{\sqrt{(4)(7340)}}\right)\frac{\sqrt{L}}{\sqrt{\mu N}}\right)\frac{\sigma_*^2}{N} \\
&\le \left(\left(6862 + 500\sqrt{\mathscr{E}} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)}{5\mathscr{E}}\right)\frac{\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + \left(9536 + 343\sqrt{\mathscr{E}}\right)\frac{\sqrt{L}}{\sqrt{\mu N}}\right)\frac{\sigma_*^2}{N} \\
&\le \left(\left(7352\sqrt{\mathscr{E}} + \log\left(\frac{T}{\mathscr{B}}\right)\right)\frac{\ell_\Xi^{2/3}}{\mu^{2/3}N^{1/3}} + 9879\sqrt{\mathscr{E}}\frac{\sqrt{L}}{\sqrt{\mu N}}\right)\frac{\sigma_*^2}{N}.
\end{aligned}$$

This concludes (1.23) and hence the whole theorem.

1.3.5 Proof of Theorem 3.1

Here we provide a two-step proof of Theorem 3.1:

Proof (Proof of Theorem 3.1). We continue to adopt the v_t—z_t decomposition as earlier used, and we proceed with the proof in two steps:

(i) We first have the following single-epoch result, (1.41), that under the setting of Theorem 3.1 along with $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$, the single-epoch estimator produced by Algorithm 5 with burn-in time $\mathscr{B} = \left\lceil \frac{24}{\eta\mu} \right\rceil$, as $T \to \infty$, $\eta \to 0$ such that $\eta T \to \infty$ satisfies the following convergence in probability:

$$\sqrt{T} z_T - \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \varepsilon_s(\theta^*) \xrightarrow{p} 0. \tag{1.41}$$

Taking this as given, we now combine (1.41) with our multi-epoch design Algorithm 6 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 6 will be triggered with length $T = N - T_b\mathscr{E}$, and hence we apply (1.37) so for some $C \le 56$ we have the initial condition holds: $\mathbb{E}\|\nabla F(\theta_0^{(\eta)})\|_2^2 \le \frac{C\sigma_*^2}{T_b} = O(\eta\mu\sigma_*^2)$, achieved by the proof of Theorem 1.2.B in §1.3.4, so that as

$\eta T \to \infty$,

$$T\mathbb{E}\|v_T\|_2^2 \le O\left(\frac{\sigma_*^2}{\eta\mu T} + \frac{\eta\mu\sigma_*^2}{\eta^4\mu^4T^3}\right) \to 0.$$

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 $N - T_b\mathscr{E}$ we can apply Slutsky's rule of weak convergence and obtain the desired weak convergence: as $\eta \to 0, N \to \infty$ such that $\eta(N - T_b\mathscr{E}) \to \infty$

$$\begin{aligned}\sqrt{T}\nabla F(\theta_{T-1}^{(\eta)}) &= \sqrt{T}v_T - \sqrt{T}z_T \\ &= \sqrt{T}v_T - \left(\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^T \varepsilon_s(\theta^*)\right) - \frac{1}{\sqrt{T}}\sum_{s=1}^T \varepsilon_s(\theta^*) \xrightarrow{d} \mathcal{N}(0,\Sigma^*),\end{aligned}$$

concluding Theorem 3.1.

(ii) We proceed to prove (1.41) with the extra initialization condition $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$. By (1.37), (1.38), we have for $T \ge \mathscr{B}$ 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

$$\mathbb{E}\|z_T\|_2^2 \le \frac{2a_2}{T},$$

and consequently, we have from (1.37) that

$$\mathbb{E}\|v_T\|_2^2 \le \frac{752\sigma_*^2}{\eta\mu T^2} + \frac{69175}{\eta^4\mu^4T^4}\|\nabla F(\theta_0)\|_2^2 \le \frac{a_1}{T}\left(\frac{1}{\eta T} + \frac{\eta}{\eta^4T^3}\right) \le \frac{2a_1}{\eta T^2},$$

and hence

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \le 2\left(\mathbb{E}\|v_T\|_2^2 + \mathbb{E}\|z_T\|_2^2\right) \le \frac{4a_1}{\eta T^2} + \frac{4a_2}{T} \le \frac{4a_3}{T}.$$

Note from the definition in (1.27)

$$tz_t = \varepsilon_t(\theta_{t-1}) + (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2}))$$

we have by setting $A_t = (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)$, $Tz_T - \sum_{s=1}^T \varepsilon_s(\theta^*) = \sum_{s=1}^T A_s$ which is a martingale. We only need to show the following to conclude (1.41) as $T \to \infty$, $\eta \to 0$:

$$\mathbb{E}\left\|\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*)\right\|_2^2 = \frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\|A_s\|^2 \to 0. \tag{1.42}$$

Since we have

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{s=\mathscr{B}+1}^{T}(s-1)(\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2}))\right\|_2^2 = \sum_{s=\mathscr{B}+1}^{T}(s-1)^2\mathbb{E}\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2\\
&\le \ell_\Xi^2\sum_{s=\mathscr{B}+1}^{T}(s-1)^2\mathbb{E}\|\theta_{s-1}-\theta_{s-2}\|_2^2 = \eta^2\ell_\Xi^2\sum_{s=\mathscr{B}+1}^{T}(s-1)^2\mathbb{E}\|v_{s-1}\|_2^2\\
&\le \eta^2\ell_\Xi^2\sum_{s=\mathscr{B}+1}^{T}(s-1)^2\frac{2a_1}{\eta^4(s-1)^4} \le \frac{2a_1\ell_\Xi^2}{\eta^2\mathscr{B}}.
\end{aligned}$$

We note that

$$\begin{aligned}
\mathbb{E}\left\|\sum_{s=\mathscr{B}+1}^{T}(\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta^*))\right\|_2^2 &= \sum_{s=\mathscr{B}+1}^{T}\mathbb{E}\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta^*)\|_2^2\\
&\le \ell_\Xi^2\sum_{s=\mathscr{B}+1}^{T}\mathbb{E}\|\theta_{s-1}-\theta^*\|_2^2 \le \frac{\ell_\Xi^2}{\mu^2}\cdot 4a_3\log\left(\frac{T}{\mathscr{B}}\right).
\end{aligned}$$

Therefore, combining this with $\mathbb{E}\|A_t\|_2^2 = \mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\|_2^2 \le 2\ell_\Xi^2\eta^2(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$ we have as $T\to\infty$, $\eta\to 0$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|A_t\|_2^2 &\le \frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\|_2^2\\
&\le 2\ell_\Xi^2\eta^2\cdot\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\\
&\le 2\ell_\Xi^2\eta^2\cdot\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}(t-1)^2\frac{2a_1}{\eta(t-1)^2} + \frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T}\frac{4a_3}{t}\\
&= 4a_1\ell_\Xi^2\eta + \frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{4a_3\log\left(\frac{T}{\mathscr{B}}\right)}{T},
\end{aligned}$$

i.e. the limit (1.42) 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 (1.41).

	Algorithm	Complexity	Reference
ISC	SGD	$\max\left\{\frac{\ell_{max}}{\mu}\log\left(\frac{\ell_{max}}{\mu\varepsilon}\right), \frac{\ell_{max}^2\sigma_*^2}{\mu^2\varepsilon^2}\right\}$	(Nguyen et al., 2019)
	SCSG	$\frac{\ell_{max}\sigma_*^2}{\mu\varepsilon^2}\log\left(\frac{\ell_{max}}{\mu\varepsilon}\right)$	(Lei & Jordan, 2017)
	Inexact SARAH	$\max\left\{\frac{\ell_{max}}{\mu}, \frac{\sigma_*^2}{\varepsilon^2}\right\}\log\left(\frac{1}{\varepsilon}\right)$	(Nguyen et al., 2021)
	ROOT-SGD	$\max\left\{\frac{\ell_{max}}{\mu}\log\left(\frac{\ell_{max}}{\mu}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}$	(This work, Theorem 1.1.B)
LSN	SGD	$\max\left\{\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\log\left(\frac{L}{\mu\varepsilon}\right), \frac{L^2\sigma_*^2}{\mu^2\varepsilon^2}\right\}$	(Nguyen et al., 2019)
	SGD3	$\max\left\{\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\log\left(\frac{1}{\varepsilon}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}\times$ polylog	(Allen-Zhu, 2018)
	ROOT-SGD	$\max\left\{\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\log\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}$	(This work, Theorem 1.1.B)

Table 1.1 Comparison with existing results on the stochastic gradient complexity for our problem for finding $\|\nabla F(\theta)\| \le O(\varepsilon)$. For simplicity of comparison, we consider the representative regime $\varepsilon \ll \sigma_*/(\eta_{max}\mu)$ and $\|\nabla F(\theta_0)\|_2^2 \asymp \sigma_*^2 \asymp 1$. Here the polylog factor stands for $\log^\alpha\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)$ for some $\alpha \in [1,3]$, possibly different at each appearance when multiplying inside. Our (multi-epoch) ROOT-SGD complexity is detailed as in (1.20) of Theorem 1.1.B.

1.4 Comparisons with concurrent work

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. We summarize the comparable results in Table 1.1 where complexities are presented up to a constant factor.

(i) In order to compare ROOT-SGD with SGD (without averaging), we impose Assumption 14 that allows the noise variance to be unbounded. A recent analysis due to Nguyen et al. (2019), which builds upon earlier analysis surveyed by Bottou et al. (2018), makes a comparable noise assumption and applies to SGD in both **ISC** and **LSN** cases. In special, Nguyen et al. (2019) shows that for appropriate diminishing step sizes η_t we have $\mathbb{E}\|\theta_T^{\mathrm{SGD}} - \theta^*\|_2^2 \lesssim \frac{\sigma_*^2}{\mu^2 T}$. We observe that both for single-epoch design and both **ISC** and **LSN** cases, the convergence rate of SGD is in no regime better than that of ROOT-SGD presented in (1.17). Moreover, generalizing their analysis to appropriate multi-epoch design for SGD further allows the convergence to be valid for any T after the burn-in period, which presents the corresponding complexities in Table 1.1 after a straightforward metric conversion. Such a multi-epoch SGD is also in no regime better than the optimized multi-epoch ROOT-SGD complexity depicted in (1.20). Specially in the **ISC** case, the squared condition number factor $\frac{\ell_{max}^2}{\mu^2}$ of the optimal statistical risk term can be mitigated to $\frac{\ell_{max}}{\mu}$ by a variance-reduced design (Lei & Jordan, 2017).

(ii) For the **ISC** case, the Inexact SARAH algorithm developed by Nguyen et al. (2021) achieves an complexity upper bound of $O\left(\max\left\{\frac{\ell_{max}}{\mu}, \frac{\sigma_*^2}{\varepsilon^2}\right\}\log\left(\frac{\sigma_*^2}{\varepsilon}\right)\right)$ to achieve an estimator of gradient norm bounded by $O(\varepsilon)$. To the best of our knowledge, their result is the first that achieves a complexity of $O\left(\frac{\sigma_*^2}{\varepsilon}\right)$ to achieve an $O(\varepsilon)$-gradient in leading-order optimal risk term, up to a logarithmic factor. In comparison, our multi-epoch algorithm ROOT-SGD improves this complexity by removing the logarithmic factor in this term, as well as improving the logarithmic factor in the $\frac{\ell_{max}}{\mu}$ term. More importantly, our multi-epoch ROOT-SGD generalizes to the **LSN** case where it attain a near-optimal asymptotic guarantee.[13]

(iii) For the **LSN** case, Allen-Zhu (2018) developed a multi-epoch variant of SGD with averaging (called SGD3) via recursive regularization techniques that achieves a near-optimal rate for attaining an estimator of gradient norm of $O(\varepsilon)$. The convergence theory of ROOT-SGD requires the finiteness assumption of either ℓ_{max} or ℓ_Ξ, rendering worse dependency on these parameters in their set of (more relaxed) assumptions. Despite that, our ROOT-SGD is set under the noise assumption in statistical learning setting (Bottou et al., 2018; Nguyen et al., 2019) that imposes more stringent stochastic Lipschitzness condition while allowing the noise variance to break the boundedness.[14] Without sketching its proof, one can (optimistically at best) achieve via a fine-grained analysis and impose an *effective* variance bound $\sigma_*^2 + \frac{\varepsilon^2\ell_\Xi^2}{\mu^2}\log\left(\frac{1}{\varepsilon}\right)$, leading to a complexity upper bound as taking the maximal of the following two terms:

$$\frac{L}{\mu}\log\left(\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\cdot\frac{\|\nabla F(\theta_0)\|_2}{\sigma_*^2}\right)\log\left(\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}\right),$$

and

$$\max\left\{\frac{\ell_\Xi^2}{\mu^2}\log\left(\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}\right), \frac{\sigma_*^2}{\varepsilon^2}\right\}\log^3\left(\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\cdot\frac{\|\nabla F(\theta_0)\|_2}{\sigma_*^2}\right),$$

which *cannot* exactly match the complexity of our multi-epoch ROOT-SGD Algorithm 6 in (1.20) of Theorem 1.1.B, partly due to at least $\log^3(\cdot)$ extra poly-logarithmic factor in condition number in its leading-order optimal risk term, letting alone matching the near-unity prefactor in Theorem 1.2.B.[15]

---

[13] It is unclear at this point if these new guarantees by ROOT-SGD is achievable by some variant of Inexact SARAH, which is not reported by Nguyen et al. (2021) and is pending future studies. The algorithm of Nguyen et al. (2021) requires random output and burn-in batches that depend on ε, yielding an upper bound that is logarithmic prefactor multiples of the statistical risk lower bound.

[14] Note that Allen-Zhu (2018) still poses a ℓ_{max} without individually individually convex assumption, which easily generalizes to the **LSN** assumption as indicated here.

[15] Their setting and analysis can be translated to our set of assumptions in the **LSN** case, where the multi-epoch ROOT-SGD achieves a convergence rate upper bound that is no worse than their SGD3, since their variance bound scales locally as

1.5 Future directions

We have shown in this work that ROOT-SGD with appropriate specifications enjoys favorable asymptotic and nonasymptotic behavior for solving the stochastic optimization problem (3.1) in the smooth, strongly convex case. With this focus, several promising future directions are left to explore:

(i) It is natural to extend our results for ROOT-SGD to nonstrongly convex and nonconvex settings both in the nonasymptotic and asymptotic setting. It would also be highly 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 matching the minimax lower bound in all terms (Agarwal et al., 2012; Woodworth & Srebro, 2016).

(ii) From the statistical optimization viewpoint, while this work has been concentrating on the first-order smoothness setting, stacking up continuity conditions on Hessians might provide improved convergence rates. Specially, it is curious if the convergence behaviors of ROOT-SGD match the near-unity statistical rate with optimal leading lower-order efficiency under a slightly more stringent set of smoothness and noise, while mild for a wide range of statistical learning application, assumptions (for instance, the efficiency of the maximal likelihood estimator (van der Vaart, 2000; Frostig et al., 2015)).

(iii) For statistical inference using online samples, the near-unity nonasymptotic and asymptotic result presented in this work would potentially allow us to provde confidence intervals and other inferential assertions for the use of ROOT-SGD estimators. It is worth mentioning that several recent works have been focusing on the SGD with Ruppert-Polyak-Juditsky averaging (Chen et al., 2020; Su & Zhu, 2018), and it will be interesting to see its analogy for the ROOT-SGD algorithm, which shall be a promising direction building upon the asymptotic normality that our work has established.

In the appendix, we discuss more on related works [§2.4], as well as the proofs of auxiliary lemmas [§1.7 and §1.8].

1.6 More discussions on related works

We lay more discussions on related works in this section.

$$\sigma_*^2 + \ell_\Xi^2 \|\theta_0 - \theta^*\|_2^2 \leq \sigma_*^2 + \frac{\ell_\Xi^2}{\mu^2} \|\nabla F(\theta_0)\|_2^2.$$

However, this is optimistically the *best-case* translation, and a rigorous analysis is also missing and pending future research.

1.6.1 Ruppert-Polyak-Juditsky averaging

The step sizes that are possible with PRJ-averaged SGD are significantly smaller than the step sizes possible with ROOT-SGD. Indeed, for a generic smooth and strongly convex objective function, PRJ-averaged SGD need not converge if a constant η is used (Polyak & Juditsky, 1992; Ruppert, 1988). In general, the choice of step size for PRJ-averaged SGD reflects a tradeoff between terms involving the initialization and linearization error. As a result, a sequence of diminishing step sizes $\eta_t \to 0$ is generally unavoidable and hence the range of feasible step sizes is *limited* (Bach & Moulines, 2011; Dieuleveut et al., 2020). Comparatively, our ROOT-SGD implicitly adjusts and corrects the gradient estimation error, and hence the linearization error is completely avoided, allowing for the step size choices that forget the initial condition as quickly as possible. Indeed, the asymptotic result for constant step size ROOT-SGD for *general* stochastic optimization problems is exactly the same as that of the PRJ algorithm for linear problems, where the stochastic gradients are evaluated at θ^*. An important consequence of this large step size choice is that ROOT-SGD enjoys state-of-the-art nonasymptotic rates for gradient-norm minimization at the cost of a slightly stronger Lipschitz continuity condition on the stochastic gradients.

1.6.2 STORM and HSGD

It is important to distinguish our update rule in (3.5) from "momentum" terms that aim to achieve acceleration (Dieuleveut et al., 2017; Cutkosky & Orabona, 2019). First, viewed as a gradient estimator the momentum term is in general *biased*: it can be interpreted as a certain moving average of stochastic gradients at extrapolation points. Thinking of the case of all noise terms being zero, we rediscover the gradient descent algorithm which enjoys a sample complexity bounded by, up to a logarithmic factor, the condition number *in lieu to* the square-rooted condition number as Nesterov's accelerated gradient descent in the smooth and strongly convex setting.[16] Secondly, in contradistinction to momentum-based acceleration algorithms, ROOT-SGD is a two-timescale method that couples fast-slow iteration where the $\frac{1}{t}$ step size in the v_t update is asymptotically smaller than the η_t step size in the θ_t update as $t \to \infty$. Note in particular that Cutkosky & Orabona (2019) essentially put $\eta_t \asymp 1/\sqrt{t}$ in STORM updates. Existing theory on SGD with momentum acceleration suggests the choice of two sequences of step sizes be asymptotically of the same scale, modulo the prefactors (Nesterov, 2018).

Alternatively, we can also interpret v_t as an estimator of the gradient that is a hybrid of stochastic gradient (SG) and stochastic recursive gradient (SARAH) esti-

[16] In contrast to ROOT-SGD, a treatment of Ruppert-Polyak-Juditsky averaging is still required for stochastic gradient descent with Nesterov's momentum acceleration to achieve the Cramer-Rao lower bound.

mators (Tran-Dinh et al., 2021):

$$v_t = \frac{1}{t}\underbrace{\nabla f(\theta_{t-1};\xi_t)}_{\text{stochastic gradient}} + \frac{t-1}{t}\underbrace{(v_{t-1}+\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))}_{\text{stochastic recursive gradient}}. \tag{1.43}$$

In this expression, one observes that the ratio of the weights for the SG and SARAH estimators has the proportion $1:(t-1)$, yielding the dominance of the SARAH part as t becomes large. In this vein, Tran-Dinh et al. (2021) proposed (in the non-convex setting) a gradient estimator that linearly interpolates the two estimators. This requires two independent samples, however, doubling the sample complexity. In contrast, the update rule of ROOT-SGD use the same sample for both stochastic and stochastic recursive estimators and it enjoys both desirable asymptotic and nonasymptotic convergence behavior for the smooth and strongly convex setting.[17]

1.7 Proofs of auxiliary lemmas in §1.3.1

In this section, we provide proofs of the auxiliary lemmas for proving Theorem 1.1.A.

1.7.1 Proof of Lemma 1.4

The claim (1.29b) follows from the definition along with some basic probability. In order to prove the claim (1.29a), recall from the ROOT-SGD update rule for v_t in the first line of (3.5) that for $t \geq \mathscr{B}+1$ we have:

$$tv_t = (t-1)v_{t-1} + t\nabla f(\theta_{t-1};\xi_t) - (t-1)\nabla f(\theta_{t-2};\xi_t). \tag{1.44}$$

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

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

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

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

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

[17] Concurrent to this work, Liu et al. (2020) proposes the momentum SARAH method that resolves this issue and reduce two independent samples to one.

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

where in the last step follows from Young's inequality. Computing the constants out completes the proof of the claim (1.29a).

1.7.2 Proof of Lemma 1.5

Eq. (1.30a) follows in a straightforward manner by expanding the square and taking an expectation. As for the inequality (1.30b), from the update rule (1.9) for v_t, we have

$$tv_t - \nabla F(\theta_{t-1}) = t\nabla f(\theta_{t-1};\xi_t) + (t-1)\,[v_{t-1} - \nabla f(\theta_{t-2};\xi_t)] - \nabla F(\theta_{t-1})$$
$$= (t-1)v_{t-1} + (t-1)\,[\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)] + \varepsilon_t(\theta_{t-1}).$$

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

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

Further rearranging yields

$$\mathbb{E}\|tv_t - \nabla F(\theta_{t-1})\|_2^2 = (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2(t-1)^2\mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2$$
$$+ 2\mathbb{E}\,\|\varepsilon_t(\theta_{t-1})\|_2^2 + 2(t-1)^2\mathbb{E}\,\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle\,. \quad (1.45)$$

We split the remainder of our analysis into two cases, corresponding to the **LSN** Setting or the **ISC** Setting. 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 Setting:

From L-Lipschitz smoothness of F in Assumption 5, we have

$$\big\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\big\rangle = -\frac{1}{\eta}\big\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\big\rangle$$
$$\leq -\frac{1}{\eta L}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2. \quad (1.46)$$

Now consider the inner product term $\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle$ in Eq. (1.45). We split it into two terms, and upper bound them using equations (1.48) and (1.46)

respectively. Doing so yields:

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

From the condition (1.14b), 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 Setting:

We deal with the last summand in the last line of Eq. (1.45), where we use the iterated law of expectation to achieve

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

The update rule for v_t implies that $v_{t-1} = -\frac{\theta_{t-1}-\theta_{t-2}}{\eta}$ for all $t \ge \mathscr{B}+1$. The following analysis uses various standard inequalities (c.f. §2.1 in Nesterov (2018)) that hold for individually convex and ℓ_{max}-Lipschitz smooth functions. First, we have

$$
\begin{aligned}
\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle &= -\frac{1}{\eta}\,\langle \theta_{t-1} - \theta_{t-2}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle \\
&\le -\frac{1}{\eta\,\ell_{max}}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 ,
\end{aligned}
\tag{1.47}
$$

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

$$
\begin{aligned}
\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle &= -\frac{1}{\eta}\,\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle \\
&\le -\frac{\mu}{\eta}\|\theta_{t-1} - \theta_{t-2}\|_2^2 = -\eta\mu\|v_{t-1}\|_2^2 .
\end{aligned}
\tag{1.48}
$$

Plugging the bounds (1.47) and (1.48) into Eq. (1.45) yields

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

where in the last inequality relies on the fact that $\eta \in (0, \frac{1}{4\ell_{max}}]$ (see Eq. (1.14b)), leading to the bound (1.30b).

1.7.3 Proof of Lemma 1.6

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 7), or the functions are individually convex and smooth (cf. Assumption 4).

Analysis in the LSN Setting:

From the ℓ_Ξ-Lipschitz smoothness of the stochastic gradients (Assumption 7) and the μ-strong-convexity of F (Assumption 5), we have

$$\begin{aligned}
\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 &\leq 2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\
&\leq 2\ell_\Xi^2\mathbb{E}\|\theta_{t-1} - \theta^*\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\
&\leq \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2,
\end{aligned} \tag{1.49}$$

which establishes the claim.

Analysis in the ISC Setting:

Using Assumption 4 and standard inequalities for ℓ_{max}-smooth and convex functions yields

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

Taking expectations in this inequality and performing some algebra[18] yields

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

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

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

We also note that[19]

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

Finally, applying the argument of (1.49) yields the claim (1.31).

1.8 Proofs of auxiliary lemmas in §1.3.3

1.8.1 Proof of Lemma 1.7 (Sharp bound on v_t)

Proof (Proof of Lemma 1.7). Our main technical tools is the following Lemma 1.9, which bound the second moments of v_t based on other parameters.

First of all, combining (1.50) in Lemma 1.9 into Theorem 1.1.A, we obtain a bound for $\mathbb{E}\|v_t\|_2^2$:

Lemma 1.9 (v_t recursion). *Under the setting of Theorem 1.2.A, when $\eta \le \frac{1}{4L}\wedge\frac{\mu}{8\ell_\Xi^2}$, for $\mathscr{B}=24\mu^{-1}\eta^{-1}$ we have the following bound for $t\ge\mathscr{B}+1$*

$$t^2\mathbb{E}\|v_t\|_2^2 \le \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{4}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 5\sigma_*^2. \quad (1.50)$$

The detailed proof of Lemma 1.9 is relegated to §1.8.3.1.

[18] 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.

[19] This proof strategy is forklore and appears elsewhere in the variance-reduction literature; see, e.g., the proof of Theorem 1.1.A in Johnson & Zhang (2013), and also adopted by Nguyen et al. (2019, 2021)).

Now, to combine everything together, we conclude from (1.16) and (1.50) that

$$\begin{aligned} t^2\mathbb{E}\left\|v_t\right\|_2^2 &\leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{4}{\eta\mu}\left[\frac{2700\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2t^2}+\frac{28\,\sigma_*^2}{t}\right]+5\sigma_*^2 \\ &\leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3t^2}+117\sigma_*^2. \end{aligned} \tag{1.51}$$

We have from (1.51)

$$\begin{aligned} t^4\mathbb{E}\left\|v_t\right\|_2^2 &\leq \left(1-\frac{\eta\mu}{2}\right)t^2(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3}+117\sigma_*^2t^2 \\ &\leq \left(1-\frac{\eta\mu}{6}\right)(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3}+117\sigma_*^2t^2, \end{aligned}$$

since the following holds $\frac{t^2}{(t-1)^2}\leq\frac{1-\frac{\eta\mu}{6}}{(1-\frac{\eta\mu}{6})^3}\leq\frac{1-\frac{\eta\mu}{6}}{1-\frac{\eta\mu}{2}}$ for $t\geq\frac{6}{\eta\mu}$.[20] This gives, by solving the recursion,

$$\begin{aligned} &T^4\mathbb{E}\left\|v_T\right\|_2^2 \leq \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^2+\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\left(\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3}+117\sigma_*^2t^2\right) \\ &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^2+\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3}+\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}117\sigma_*^2t^2 \\ &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^2+\frac{6}{\eta\mu}\cdot\frac{(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3}+\frac{6}{\eta\mu}\cdot117\sigma_*^2T^2 \\ &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^2+\frac{(6)(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4}+\frac{(6)(117)\sigma_*^2}{\eta\mu}T^2, \end{aligned} \tag{1.52}$$

where the summand is increasing so

$$\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}t^2\leq\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}T^2\leq\frac{6}{\eta\mu}T^2.$$

All in all, this concludes

$$\mathbb{E}\left\|v_T\right\|_2^2\leq\left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\frac{\mathscr{B}^4}{T^4}\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^2+\frac{(6)(4)(2700)\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}+\frac{(6)(117)\sigma_*^2}{\eta\mu T^2}. \tag{1.53}$$

Lastly, for bound on the initialization scheme of ROOT-SGD in (3.5)

$$v_{\mathscr{B}}=\frac{1}{\mathscr{B}}\sum_{s=1}^{\mathscr{B}}\nabla f(\theta_0;\xi_s),$$

[20] Recall Bernstein's inequality gives $\left(1-\frac{\eta\mu}{6}\right)^3\geq1-\frac{\eta\mu}{2}$ since $\eta\mu\leq1$.

(1.29b) from Lemma 1.4 along with Lemma 1.6 immediately yields the following bounds

$$\mathscr{B}\mathbb{E}\,\|z_{\mathscr{B}}\|_2^2 \leq \left(\frac{\ell_\Xi}{\mu}\,\|\nabla F(\theta_0)\|_2 + \sigma_*\right)^2 \leq \frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2 + 2\sigma_*^2,$$

and hence

$$\mathscr{B}\mathbb{E}\,\|v_{\mathscr{B}}\|_2^2 \leq \mathscr{B}\,\|\nabla F(\theta_0)\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2 + 2\sigma_*^2.$$

Bringing the last display into Eq. (1.53) we have

$$\begin{aligned}
\mathbb{E}\,\|v_T\|_2^2 &\leq \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\frac{\mathscr{B}^4}{T^4}\mathbb{E}\,\|v_{\mathscr{B}}\|_2^2 + \frac{(6)(4)(2700)\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{(6)(117)\sigma_*^2}{\eta\mu T^2}\\
&\leq \frac{\mathscr{B}^3}{T^4}\left(\mathscr{B}\,\|\nabla F(\theta_0)\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2 + 2\sigma_*^2\right) + \frac{(6)(4)(2700)\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{(6)(117)\sigma_*^2}{\eta\mu T^2}\\
&\leq \frac{(25)(25)(7)\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{(6)(4)(2700)\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{752\sigma_*^2}{\eta\mu T^2}\\
&\leq \frac{69175\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{752\sigma_*^2}{\eta\mu T^2}.
\end{aligned}$$

1.8.2 Proof of Lemma 1.8 (Sharp bound on z_t)

In this section, we prove the refined bound on estimating $\mathbb{E}\,\|z_t\|_2^2$.

(i) 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

$$tz_t - (t-1)z_{t-1} = \varepsilon_t(\theta^*) + \underbrace{(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)}_{\equiv A_t}.$$

In words, the increment of tz_t splits into two parts: the additive part $\varepsilon_t(\theta^*)$ and the multiplicative part A_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

$$\begin{aligned}
t^2\mathbb{E}\|z_t\|_2^2 - (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 &= \mathbb{E}\,\|\varepsilon_t(\theta^*) + A_t\|_2^2\\
&= \mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 + \mathbb{E}\|A_t\|_2^2 + 2\mathbb{E}\,\langle\varepsilon_t(\theta^*), A_t\rangle\,.
\end{aligned}$$

Telescoping the above equality for $t = \mathscr{B}+1,\ldots,T$ gives

$$T^2\mathbb{E}\left\|z_T\right\|_2^2-\mathscr{B}^2\mathbb{E}\left\|z_{\mathscr{B}}\right\|_2^2=\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2+\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|A_t\|_2^2+2\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\left\langle\varepsilon_t(\theta^*),A_t\right\rangle. \tag{1.54}$$

(ii) We first find that $\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2=\sigma_*^2$ and standard Young's inequality gives[21]

$$\begin{aligned}\mathbb{E}\|A_t\|_2^2 &\le \left(\ell_\Xi(t-1)\sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta_{t-2}\right\|_2^2}+\ell_\Xi\sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta^*\right\|_2^2}\right)^2\\ &\le \left(\eta\ell_\Xi(t-1)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2}+\frac{\ell_\Xi}{\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}\right)^2\\ &\le 2\eta^2\ell_\Xi^2(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2.\end{aligned} \tag{1.55}$$

Now, we bound the partial sum of the expected inner product sequence $\mathbb{E}\left\langle\varepsilon_t(\theta^*),A_t\right\rangle$. Note bounding $\mathbb{E}\left\langle\varepsilon_t(\theta^*),A_t\right\rangle$ by the sum of $\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2$ and $\mathbb{E}\left\|A_t\right\|_2^2$ directly gets us a raw bound on z_t. The key technical contribution is $\ell_\Xi\eta(t-1)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2}$ is asymptotically larger than $\frac{\ell_\Xi}{\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}$.
We now estimate the cross term so its bound is significantly better than directly applying Cauchy-Schwarz. For any $T\ge\mathscr{B}+1$ in Algorithm ROOT-SGD, we have

$$\begin{aligned}&\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\left\langle\varepsilon_t(\theta^*),A_t\right\rangle\\ =&\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\left\langle\varepsilon_t(\theta^*),\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)+(t-1)\left(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right)\right\rangle\\ =&\sum_{t=\mathscr{B}+1}^{T}\left[t\mathbb{E}\left\langle\varepsilon_t(\theta^*),\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\right\rangle-(t-1)\mathbb{E}\left\langle\varepsilon_t(\theta^*),\varepsilon_t(\theta_{t-2})-\varepsilon_t(\theta^*)\right\rangle\right].\end{aligned}$$

Replacing $\varepsilon_t(\cdot)$ by $\varepsilon_T(\cdot)$ for its argument being $\mathscr{F}_{t-1}$-measurable, we have by Assumption 7

[21] This bound itself is useful for the proof of our asymptotic result, Theorem 3.1.

$$
\begin{aligned}
&\sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\,\langle \varepsilon_t(\theta^*), A_t\rangle \\
&= \sum_{t=\mathscr{B}+1}^{T} [t\mathbb{E}\,\langle \varepsilon_T(\theta^*), \varepsilon_T(\theta_{t-1}) - \varepsilon_T(\theta^*)\rangle - (t-1)\mathbb{E}\,\langle \varepsilon_T(\theta^*), \varepsilon_T(\theta_{t-2}) - \varepsilon_T(\theta^*)\rangle] \\
&= T\mathbb{E}\,\langle \varepsilon_T(\theta^*), \varepsilon_T(\theta_{T-1}) - \varepsilon_T(\theta^*)\rangle - \mathscr{B}\mathbb{E}\,\langle \varepsilon_T(\theta^*), \varepsilon_T(\theta_0) - \varepsilon_T(\theta^*)\rangle \\
&\le T\sqrt{\mathbb{E}\,\|\varepsilon_T(\theta^*)\|_2^2\,\mathbb{E}\,\|\varepsilon_T(\theta_{T-1}) - \varepsilon_T(\theta^*)\|_2^2} + \mathscr{B}\sqrt{\mathbb{E}\,\|\varepsilon_T(\theta^*)\|_2^2\,\mathbb{E}\,\|\varepsilon_T(\theta_0) - \varepsilon_T(\theta^*)\|_2^2} \\
&\le T\sqrt{\sigma_*^2\ell_\Xi^2\mathbb{E}\,\|\theta_{T-1} - \theta^*\|_2^2} + \mathscr{B}\sqrt{\sigma_*^2\ell_\Xi^2\,\|\theta_0 - \theta^*\|_2^2} \\
&\le \frac{\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\,\|\nabla F(\theta_0)\|_2\right),
\end{aligned} \tag{1.56}
$$

where we used $\theta_{\mathscr{B}-1} = \theta_0$ along with Assumptions 5 and 7. Combining (1.54), (1.55) and (1.56), we have for any iteration $T \ge \mathscr{B}+1$

$$
\begin{aligned}
&T^2\mathbb{E}\,\|z_T\|_2^2 - \mathscr{B}^2\mathbb{E}\,\|z_{\mathscr{B}}\|_2^2 \\
&\le (T-\mathscr{B})\sigma_*^2 + 2\eta^2\ell_\Xi^2\sum_{t=\mathscr{B}+1}^{T}(t-1)^2\mathbb{E}\,\|v_{t-1}\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 \\
&\quad + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\,\|\nabla F(\theta_0)\|_2\right).
\end{aligned} \tag{1.57}
$$

In order to further estimate (1.57), we proceed to bound each term in the following steps separately in a reversed order.

(iii) From (1.16), we have for $\sqrt{\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^2}$ using Pythagorean theorem $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for two positive reals:

$$
\begin{aligned}
T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} &\le T\left(\sqrt{\frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2}} + \sqrt{\frac{28\ \sigma_*^2}{T}}\right) \\
&\le \frac{52\,\|\nabla F(\theta_0)\|_2}{\eta\mu} + 6\sigma_*\sqrt{T}.
\end{aligned}
$$

Also, telescoping directly the bound (1.16), we have

$$
\begin{aligned}
\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 &\le \sum_{t=\mathscr{B}+1}^{T}\left(\frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2} + \frac{28\ \sigma_*^2}{t}\right) \\
&\le \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2\mathscr{B}} + 28\ \sigma_*^2\log\left(\frac{T}{\mathscr{B}}\right) \\
&= \frac{2700\mathscr{C}\ \|\nabla F(\theta_0)\|_2^2}{\eta\mu} + 28\ \sigma_*^2\log\left(\frac{T}{\mathscr{B}}\right),
\end{aligned}
$$

where we plug in the burn-in time $\mathscr{B} = \mathscr{C}^{-1}\mu^{-1}\eta^{-1}$ for some prescribed real positive $\mathscr{C} \leq \frac{1}{24}$ to be determined later.

(iv) Recalling the v_t bound in Lemma 1.9, we have by telescoping the following variant of (1.50)

$$t^2\mathbb{E}\|v_t\|_2^2 - (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 \leq -\frac{\eta\mu}{2}(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{4}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 5\sigma_*^2,$$

and arrive at

$$\frac{\eta\mu}{2}\sum_{t=\mathscr{B}+1}^{T}(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 \leq \mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 - T^2\mathbb{E}\|v_T\|_2^2 + \frac{4}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 5(T-\mathscr{B})\sigma_*^2. \tag{1.58}$$

Replacing them into the above (1.57) we have

$$\begin{aligned}
&T^2\mathbb{E}\|z_T\|_2^2 - \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 \\
&\leq (T-\mathscr{B})\sigma_*^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{\eta\mu}{2}\sum_{t=\mathscr{B}+1}^{T}(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 \\
&\quad + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\|\nabla F(\theta_0)\|_2\right) \\
&\leq (T-\mathscr{B})\sigma_*^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot\left[\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 - T^2\mathbb{E}\|v_T\|_2^2 + \frac{4}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 5(T-\mathscr{B})\sigma_*^2\right] \\
&\quad + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\|\nabla F(\theta_0)\|_2\right),
\end{aligned}$$

so by algebraic manipulations and rearranging

$$\begin{aligned}
&T^2\mathbb{E}\|z_T\|_2^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot T^2\mathbb{E}\|v_T\|_2^2 \\
&\leq \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 + (T-\mathscr{B})\sigma_*^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{4}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \\
&\quad + \frac{4\ell_\Xi^2\eta}{\mu}\cdot 5(T-\mathscr{B})\sigma_*^2 + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\|\nabla F(\theta_0)\|_2\right) \\
&\leq \left(1 + \frac{20\ell_\Xi^2\eta}{\mu}\right)(T-\mathscr{B})\sigma_*^2 + \mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2 + \frac{4\ell_\Xi^2\eta}{\mu}\cdot\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 \\
&\quad + \frac{18\ell_\Xi^2}{\mu^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + \mathscr{B}\|\nabla F(\theta_0)\|_2\right).
\end{aligned}$$

Hence by throwing away the second summand on LHS and dividing both sides by T^2

$$
\begin{aligned}
&\mathbb{E}\,\|z_T\|_2^2 \\
&\le \left(1+\frac{20\ell_\Xi^2\eta}{\mu}\right)\frac{T-\mathscr{B}}{T^2}\sigma_*^2+\frac{\mathscr{B}^2}{T^2}\mathbb{E}\,\|z_{\mathscr{B}}\|_2^2+\frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{\mathscr{B}^2}{T^2}\mathbb{E}\,\|v_{\mathscr{B}}\|_2^2 \\
&\quad+\frac{18\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T^2}\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2+\frac{2\ell_\Xi\sigma_*}{\mu}\cdot\frac{1}{T^2}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2}+\mathscr{B}\,\|\nabla F(\theta_0)\|_2\right).
\end{aligned}
\tag{1.59}
$$

(v) Now we finish the argument by combining the last three items, to obtain a z_T bound. For bound on the initial condition, (1.29b) from Lemma 1.4 along with Lemma 1.6 yields

$$
\mathscr{B}\mathbb{E}\,\|z_{\mathscr{B}}\|_2^2\le\left(\frac{\ell_\Xi}{\mu}\,\|\nabla F(\theta_0)\|_2+\sigma_*\right)^2\le\frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+2\sigma_*^2,
$$

and

$$
\mathscr{B}\mathbb{E}\,\|v_{\mathscr{B}}\|_2^2\le\mathscr{B}\,\|\nabla F(\theta_0)\|_2^2+\frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+2\sigma_*^2.
$$

Plugging the above bounds back in (1.59) we obtain via standard estimations

$$
\begin{aligned}
&\mathbb{E}\,\|z_T\|_2^2 \\
&\le\left(1+\frac{20\ell_\Xi^2\eta}{\mu}\right)\cdot\frac{T-\mathscr{B}}{T^2}\cdot\sigma_*^2+\frac{\mathscr{B}}{T^2}\cdot\underbrace{\mathscr{B}\mathbb{E}\,\|z_{\mathscr{B}}\|_2^2}+\frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot\underbrace{\mathscr{B}\mathbb{E}\,\|v_{\mathscr{B}}\|_2^2} \\
&\quad+\frac{18\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T^2}\underbrace{\sum_{t=\mathscr{B}+1}^{T}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2}+\frac{2\ell_\Xi\sigma_*}{\mu}\cdot\frac{1}{T^2}\left(\underbrace{T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2}}+\mathscr{B}\,\|\nabla F(\theta_0)\|_2\right) \\
&\le\left(1+\frac{20\ell_\Xi^2\eta}{\mu}\right)\cdot\frac{T-\mathscr{B}}{T^2}\cdot\sigma_*^2+\frac{\mathscr{B}}{T^2}\underbrace{\left(\frac{\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+\frac{2\ell_\Xi\sigma_*}{\mu}\,\|\nabla F(\theta_0)\|_2+\sigma_*^2\right)} \\
&\quad+\frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\underbrace{\left(\mathscr{B}\,\|\nabla F(\theta_0)\|_2^2+\frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+2\sigma_*^2\right)} \\
&\quad+\frac{18\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T^2}\underbrace{\left(\frac{2700\mathscr{C}\,\|\nabla F(\theta_0)\|_2^2}{\eta\mu}+28\,\sigma_*^2\log\left(\frac{T}{\mathscr{B}}\right)\right)} \\
&\quad+\frac{2\ell_\Xi\sigma_*}{\mu}\cdot\frac{1}{T^2}\left(\underbrace{\frac{52\,\|\nabla F(\theta_0)\|_2}{\eta\mu}+6\sigma_*\sqrt{T}}+\mathscr{B}\,\|\nabla F(\theta_0)\|_2\right),
\end{aligned}
$$

which further gives

$$
\begin{aligned}
\mathbb{E}\,\|z_T\|_2^2\le{}&\frac{T-\mathscr{B}}{T^2}\cdot\sigma_*^2+\frac{20\ell_\Xi^2\eta}{\mu}\cdot\frac{T-\mathscr{B}}{T^2}\cdot\sigma_*^2+\frac{\mathscr{B}}{T^2}\cdot\frac{\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+\frac{\mathscr{B}}{T^2}\cdot\frac{2\ell_\Xi\sigma_*}{\mu}\,\|\nabla F(\theta_0)\|_2+\frac{\mathscr{B}}{T^2}\cdot\sigma_*^2 \\
&+\frac{4\ell_\Xi^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left(\mathscr{B}\,\|\nabla F(\theta_0)\|_2^2+\frac{2\ell_\Xi^2}{\mu^2}\,\|\nabla F(\theta_0)\|_2^2+2\sigma_*^2\right)
\end{aligned}
$$

$$+\frac{18\ell_{\Xi}^2}{\mu^2}\cdot\frac{1}{T^2}\left(\frac{2700\mathscr{C}\ \|\nabla F(\theta_0)\|_2^2}{\eta\mu}+28\ \sigma_*^2\log\left(\frac{T}{\mathscr{B}}\right)\right)$$
$$+\frac{2\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{1}{T^2}\left((52\mathscr{C}+1)\mathscr{B}\|\nabla F(\theta_0)\|_2+6\sigma_*\sqrt{T}\right).$$

Collecting terms and using $\mathscr{B}=\mathscr{C}^{-1}\mu^{-1}\eta^{-1}$, we arrive at

$$\begin{aligned}
\mathbb{E}\|z_T\|_2^2 &\leq \frac{\sigma_*^2}{T}+\frac{20\ell_{\Xi}^2\eta}{\mu}\cdot\frac{T-\mathscr{B}}{T^2}\cdot\sigma_*^2+\frac{\ell_{\Xi}^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2^2+\frac{2\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2\\
&\quad+\frac{4\ell_{\Xi}^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot\mathscr{B}\|\nabla F(\theta_0)\|_2^2+\frac{4\ell_{\Xi}^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot\frac{2\ell_{\Xi}^2}{\mu^2}\|\nabla F(\theta_0)\|_2^2+\frac{4\ell_{\Xi}^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T^2}\cdot 2\sigma_*^2\\
&\quad+\frac{18\ell_{\Xi}^2}{\mu^2}\cdot\frac{1}{T^2}\cdot\frac{2700\mathscr{C}\ \|\nabla F(\theta_0)\|_2^2}{\eta\mu}+\frac{18\ell_{\Xi}^2}{\mu^2}\cdot\frac{1}{T^2}\cdot 28\ \sigma_*^2\log\left(\frac{T}{\mathscr{B}}\right)\\
&\quad+\frac{2\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{1}{T^2}\cdot(52\mathscr{C}+1)\mathscr{B}\|\nabla F(\theta_0)\|_2+\frac{2\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{1}{T^2}\cdot 6\sigma_*\sqrt{T}\\
&\leq\left(1+\frac{20\ell_{\Xi}^2\eta}{\mu}\cdot\frac{T-\mathscr{B}}{T}+\frac{8\ell_{\Xi}^2\eta}{\mu}\cdot\frac{\mathscr{B}}{T}+\frac{12\ell_{\Xi}}{\mu\sqrt{T}}+\frac{(18)(28)\ell_{\Xi}^2}{\mu^2}\cdot\frac{\log\left(\frac{T}{\mathscr{B}}\right)}{T}\right)\frac{\sigma_*^2}{T}\\
&\quad+\frac{\left(1+\frac{4}{\mathscr{C}}+(18)(2700)\mathscr{C}^2+\frac{8\ell_{\Xi}^2\eta}{\mu}\right)\ell_{\Xi}^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2^2\\
&\quad+\frac{2(52\mathscr{C}+2)\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2\\
&\leq\left(1+\frac{20\ell_{\Xi}^2\eta}{\mu}+\frac{12\ell_{\Xi}}{\mu\sqrt{T}}+\frac{(18)(28)\log\left(\frac{T}{\mathscr{B}}\right)\ell_{\Xi}^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T}\\
&\quad+\frac{2(\frac{52}{24}+2)\ell_{\Xi}\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2+\frac{\left(1+(4)(24)+\frac{(18)(2700)}{24^2}+1\right)\ell_{\Xi}^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2^2,
\end{aligned}$$

where in the final step we plugged in $\mathscr{C}=\frac{1}{24}$ and bound $\frac{8\ell_{\Xi}^2\eta}{\mu}\leq 1$ due to $\eta\in(0,\eta_{max}]$. This completes the proof of (1.38) and hence Lemma 1.8.

1.8.3 Proofs of secondary lemmas

1.8.3.1 Proof of Lemma 1.9

Proof (Proof of Lemma 1.9). From the update dynamics of ROOT-SGD, we obtain the iterative update rule of tv_t By definition, we note that

$$tv_t=(t-1)v_{t-1}+t\nabla f(\theta_{t-1};\xi_t)-(t-1)\nabla f(\theta_{t-2};\xi_t). \tag{1.60}$$

Subtracting off a $\nabla F(\theta_{t-1})$ term from both sides of Eq. (1.60), we have

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

Taking the expected squared norms on both sides, we have

$$\begin{aligned}
&\mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^2 \\
&= \mathbb{E}\,\|(t-1)v_{t-1} + (t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^2 \\
&= (t-1)^2\mathbb{E}\,\|v_{t-1}\|_2^2 + \underbrace{\mathbb{E}\,\|(t-1)\,(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^2}_{\mathrm{I}} \\
&\quad + \underbrace{2(t-1)^2\mathbb{E}\,\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle + 2(t-1)\mathbb{E}\,\langle v_{t-1}, \varepsilon_t(\theta_{t-1})\rangle}_{\mathrm{II}}
\end{aligned} \tag{1.61}$$

(i) For Term I, we have from Young's inequality

$$\begin{aligned}
\mathrm{I} &= \mathbb{E}\,\|(t-1)\,(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\|_2^2 \\
&\le 2\mathbb{E}\,\|(t-1)\,(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t))\|_2^2 + 2\mathbb{E}\,\|\varepsilon_t(\theta_{t-1})\|_2^2 \\
&\le 2(t-1)^2\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + 2\eta^2\ell_\Xi^2(t-1)^2\mathbb{E}\,\|v_{t-1}\|_2^2 + \frac{4\ell_\Xi^2}{\mu^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 4\sigma_*^2
\end{aligned} \tag{1.62}$$

where we use the property of martingale and the fact that

$$\begin{aligned}
&\mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 = \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + \mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2 \\
&\le \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + \ell_\Xi^2\mathbb{E}\,\|\theta_{t-1} - \theta_{t-2}\|_2^2 \\
&= \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + \eta^2\ell_\Xi^2\mathbb{E}\,\|v_{t-1}\|_2^2
\end{aligned}$$

along with Lemma 1.6 that

$$\mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\,\|\varepsilon_t(\theta_{t-1})\|_2^2 \le \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2$$

(ii) For Term II, its second summand $2(t-1)\mathbb{E}\,\langle v_{t-1}, \varepsilon_t(\theta_{t-1})\rangle$ is zero due to iterated law of expectation; similar argument leads us to $\mathbb{E}\,\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle = \mathbb{E}\,\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle$. Therefore we have

$$\begin{aligned}
\mathrm{II} &= 2(t-1)^2\mathbb{E}\left\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\rangle + 2(t-1)\mathbb{E}\left\langle v_{t-1}, \varepsilon_t(\theta_{t-1})\right\rangle \\
&= 2(t-1)^2\mathbb{E}\left\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle \\
&= -\frac{2}{\eta}(t-1)^2\mathbb{E}\left\langle \theta_{t-1}-\theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle \\
&\le -\frac{1}{2\eta L}(t-1)^2\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 - \frac{3\mu}{2\eta}(t-1)^2\mathbb{E}\left\|\theta_{t-1}-\theta_{t-2}\right\|_2^2 \\
&= -\frac{1}{2\eta L}(t-1)^2\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 - \frac{3\eta\mu}{2}(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2
\end{aligned} \tag{1.63}$$

(iii) Combining Eq. (1.62) and (1.63) into (1.61), we have

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

For $\eta \le \eta_{max}$ satisfying (1.14a), we have both $1 - \frac{3\eta\mu}{2} + 2\eta^2\ell_\Xi^2 \le 1 - \eta\mu$ and $2 - \frac{1}{2\eta L} \le 0$ and hence

$$\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^2 \le (1-\eta\mu)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{4\ell_\Xi^2}{\mu^2}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 4\sigma_*^2 \tag{1.64}$$

Further noting that Young's inequality gives (a different coefficient from the Cauchy+Young analysis in pp. 18 in Theorem 1.1.A is adopted)

$$\begin{aligned}
\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^2 &= t^2\mathbb{E}\left\|v_t\right\|_2^2 + \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 - 2t\mathbb{E}\left\langle v_t, \nabla F(\theta_{t-1})\right\rangle \\
&\ge \left(1 - \frac{\eta\mu}{2}\right)t^2\mathbb{E}\left\|v_t\right\|_2^2 - \frac{2}{\eta\mu}\left(1 - \frac{\eta\mu}{2}\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2
\end{aligned}$$

Combining this with (1.64) we have from $1 - \eta\mu \le \left(1 - \frac{\eta\mu}{2}\right)^2$

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

Now we multiply both sides by $\left(1-\frac{\eta\mu}{2}\right)^{-1}$ which lies in $[1,\frac{8}{7}]$, rearranging, and have

$$
\begin{aligned}
&t^2\mathbb{E}\|v_t\|_2^2 \\
&\leq \frac{2}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+\left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+\frac{\left(1-\frac{\eta\mu}{2}\right)^{-1}(4)\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+\left(1-\frac{\eta\mu}{2}\right)^{-1}(4)\sigma_*^2 \\
&\leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+\left(\frac{2}{\eta\mu}+\frac{\left(\frac{8}{7}\right)(4)\ell_\Xi^2}{\mu^2}\right)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+\left(\frac{8}{7}\right)(4)\sigma_*^2 \\
&\leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2+\frac{4}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+5\sigma_*^2,
\end{aligned}
$$

where $\eta\leq\eta_{max}$ with (1.14a) allowing $\frac{\left(\frac{8}{7}\right)(4)\ell_\Xi^2}{\mu^2}\leq\frac{2}{\eta\mu}$. This finishes the proof of (1.37), and hence the whole Lemma 1.7.

Chapter 2
Stacking up Lipschitzness on Hessians

We revisit the classical problem solving for strongly-convex and smooth M-estimators using stochastic optimization algorithms, and establish sharp non-asymptotic results that matches the exact behavior of the optimal statistical risk. In particular, we show that the ROOT-SGD algorithm introduced by Li et al. (2020) with constant step-size converges non-asymptotically in gradient norm to the asymptotically optimal risk with near-unity pre-factor and exponentially decaying additional terms. When a one-point Hessian-Lipschitz condition is imposed in addition, we establish upper bounds, whose leading term exactly matches the asymptotically optimal one with unity pre-factor. Moreover, through a refined analysis, we show that the ROOT-SGD algorithm with constant step-size converges non-asymptotically in gradient norm to the asymptotically optimal risk with near-unity pre-factor and exponentially decaying additional terms. The additional term in the bound scales as $O(N^{-3/2})$ with a sharp dependency on problem parameters, which shares the same efficiency as the maximum likelihood estimation.
Key words: Stochastic convex optimization, asymptotic normality, Cramér-Rao lower bound, variance reduction, averaging SGD.

2.1 Introduction

Stochastic first-order methods for optimization are widely used in large-scale machine learning. For parametric estimation tasks such as linear regression, logistic regression and deep neural networks, instead of directly minimizing the empirical risk using batch methods, a stochastic optimization algorithm processes the data with a single pass in a streaming manner, leading to substantial savings in storage and computation. In the mean time, it is desirable that the estimator produced by a stochastic optimization algorithm share the same *optimal statistical properties* typically possessed by the empirical risk minimizer. The notion of statistical efficiency, in both asymptotic and non-asymptotic forms, allows for assessment of optimality. Moreover, local asymptotic minimax theorems further show that the optimal asymptotic

distribution, under any bowl-shaped loss function, takes a Gaussian form (van der Vaart, 2000; Duchi & Ruan, 2021). The asymptotic covariance provides a form of local complexity, and it is desirable to achieve this optimal bound with a *unity* prefactor. Under relatively mild conditions, the empirical risk minimizer itself does so.

In contrast, our understanding of which first-order stochastic algorithms are optimal (or non-optimal) in this fine-grained way remains far from complete. Most existing performance guarantees are too coarse for this purpose, as the convergence rates are measured with worst-case problem parameters, and bounds are given up to universal constants instead of unity in the asymptotic limit.

In particular, 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

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{for a function of the form } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big]. \tag{2.1}$$

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:

$$\widehat{\theta}_N^{\mathrm{ERM}} := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{N}\sum_{i=1}^N f(\theta;\xi_i) \text{ satisfies } \sqrt{N}\left(\widehat{\theta}_N^{\mathrm{ERM}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right). \tag{2.2}$$

Furthermore, the asymptotic distribution (3.2) is known to be locally optimal (see van der Vaart (2000) and Duchi & Ruan (2021) 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 Eq.* (3.2)*?*

An affirmative answer to this question at least qualitatively, is provided by the seminal work by Polyak & Juditsky (1992); Polyak (1990); Ruppert (1988). 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 (3.2), as the number of samples grows to infinity. This algorithm lays the foundations of online statistical inference (Chen et al., 2020) and fine-grained error guarantees for stochastic optimization algorithms (Bach & Moulines, 2011; Dieuleveut et al., 2020). 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:

(i) 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 (c.f. Polyak & Juditsky (1992); Duchi & Ruan (2021)). However, the

asymptotic guarantee (3.2) 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 (c.f. van der Vaart (2000)). 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 (Le Cam, 1970; van der Vaart, 2000). *Is there a stochastic optimization algorithm that achieves the asymptotic guarantee* (3.2) *under the mildest smoothness conditions including that the Hessian is continuous but not Hölder continuous at its global optimum?*

(ii) 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 (3.2) with additional terms that decays faster as $N \to \infty$. For example, under the one-point Hessian Lipschitz condition, Bach & Moulines (2011); Gadat & Panloup (2017) established bounds in the form of

$$\mathbb{E}\left\|\widehat{\theta}_N^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \leq \frac{1}{N}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + \text{higher-order terms}, \tag{2.3}$$

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 2.3 for details). *The question then arises of whether the $O(N^{-3/2})$ higher-order term can be achieved by stochastic approximation algorithms with a sharp dependency on problem-dependent constants.*

In this chapter, we answer both questions affirmatively using a variance-reduced stochastic optimization algorithm called ROOT-SGD (Li et al., 2020).

2.1.1 Contributions

We summarize the contributions of this chapter:

(i) On the asymptotic side, under a set of slightly stronger distributional assumption we prove that ROOT-SGD is asymptotically efficient; that is, it outputs an estimator that converges asymptotically to a normal distribution whose covariance achieves the Cramér-Rao lower bound plus a correction term that vanishes when the step size tends to zero (Theorem 2.2). Notably, ROOT-SGD provides an algorithm whose asymptotic distribution matches that of stochastic gradient descent with Polyak-Ruppert-Juditsky averaging Polyak & Juditsky (1992); Ruppert (1988), while allowing a significantly larger range of choice of step size. This provides the first result establishing asymptotic optimality among variance-reduced gradient methods.

(ii) Nonasymptotically, the one-point Hessian Lipschitz at the global optimum θ^* and certain fourth-moment conditions are assumed, in Theorem 2.3, we show an upper bound on the mean-squared error (MSE) in the form of (3.3). 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-dependent prefactor, which tends to zero as $\ell_\Xi \to 0$.

2.1.2 More related works

In the field of smooth and convex stochastic optimization, variance-reduced gradient methods represented by, but not limited to, SAG (Le Roux et al., 2012), SDCA (Shalev-Shwartz & Zhang, 2013), SVRG (Johnson & Zhang, 2013), SCSG (Lei & Jordan, 2017), SAGA (Defazio et al., 2014), SARAH (Nguyen et al., 2017) and Inexact SARAH (Nguyen et al., 2021) have been proposed to improve the theoretical convergence rate of (stochastic) gradient descent. Under self-concordance conditions, Frostig et al. (2015) provide function-value bounds for a variant of the SVRG algorithm that matches the asymptotic behavior of the empirical risk minimizer, while the corresponding nonasymptotic rates can have worse dependency on the condition number compared to SGD. More recent accelerated variants of SGD provide further improvements in convergence rate (Lin et al., 2015; Shalev-Shwartz, 2016; Allen-Zhu, 2017; Lan & Zhou, 2018; Kulunchakov & Mairal, 2020; Lan et al., 2019). Additionally, a variety of recursive variance-reduced stochastic approximation methods (Fang et al., 2018; Zhou et al., 2020; Wang et al., 2019; Nguyen et al., 2021; Pham et al., 2020) have been studied in the nonconvex stochastic optimization literature. These algorithms, as well as their hybrid siblings (Cutkosky & Orabona, 2019; Tran-Dinh et al., 2021), achieve state-of-the-art convergence rates and in particular are faster than SGD under mild additional smoothness assumption on the stochastic gradients.

On the asymptotic efficiency side, SGD with Ruppert-Polyak-Juditsky averaging achieves a variety of asymptotic and nonasymptotic results have been obtained in the recent literature (Bach & Moulines, 2013; Duchi & Ruan, 2021; Dieuleveut et al., 2017, 2020; Jain et al., 2017, 2018; Asi & Duchi, 2019; Mou et al., 2020). The general idea of iteration averaging is based on the analysis of two-time-scale iterations techniques and it achieves asymptotic normality with an optimal covariance (Ruppert, 1988; Polyak & Juditsky, 1992). Turning to the case of constant-stepsize linear stochastic approximation, Mou et al. (2020) characterizes the asymptotic normal distribution for constant-step-size linear stochastic approximation, delineating the asymptotic covariance that adds onto the Cramér-Rao covariance and which vanishes as $\eta \to 0$ (Dieuleveut et al., 2020). The asymptotic efficiency of variance-reduced stochastic approximation methods, however, has been less studied, and we establish for the first time the asymptotic behavior in our ROOT-SGD analysis for strongly convex objectives. Instead of averaging the iteration, our ROOT-SGD al-

gorithm averages the past stochastic gradients with proper de-bias corrections and achieves asymptotic efficiency.[1]

On the non-asymptotic bounds side, when the objective has additional smoothness, the current bounds for either Ruppert-Polyak-Juditsky iteration averaging (Ruppert, 1988; Polyak & Juditsky, 1992) or variance-reduced stochastic approximation (Li et al., 2020). Bach & Moulines (2011) have presented a nonasymptotic analysis of SGD with iteration averaging showing that, after processing T samples, the algorithm achieves a nonasymptotic rate that matches the Cramér-Rao lower bound with a prefactor equal to one. Later Gadat & Panloup (2017) improves the additional term in the Cramér-Rao bound of Bach & Moulines (2011) under a different set of assumptions.**Junchi — DC** For online variance-reduced gradient algorithms, Frostig et al. (2015) proposes an online variant of the SVRG algorithm (Johnson & Zhang, 2013) and establishes a nonasymptotic upper bound on the excess risk, whose leading term matches the optimal asymptotics, under certain "self-concordant condition" posed on the objective function. Arnold et al. (2019) proposes a method of *Implicit Gradient Transportation* to reduce the variance of the algorithm. We also note that iteration averaging provides robustness and adaptivity (Lei & Jordan, 2020).

Organization. This chapter is organized as follows. §2.2 presents the asymptotic normality result and its nonasymptotic upper-bound correspondence of ROOT-SGD under the Hessian continuity assumption at the minimizer. We present the proofs of main results in §2.3. Additional discussion on comparison of our results with concurrent ones is provided in §2.4. Delayed proofs are provided in the appendix.

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

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

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

For the proof of asymptotic results, we adopt some tensor notations. For two matrices A, B, we use $A \otimes B$ to denote their Kronecker product. When it is clear from the context, we slightly overload the notation to let $A \otimes B$ denote the 4-th-order tensor produced by taking the tensor product of A and B. Note that Kronecker product is just a flattened version of the tensor. For a k-th order tensor T, matrix

[1] A related but fundamentally different idea was proposed in Nesterov (2009); Xiao (2010); Lee et al. (2012) in terms of dual averaging for the regularized/proximal case. See also Duchi & Ruan (2021); Tripuraneni et al. (2018) for first-order optimization methods over the Riemannian manifold.

Algorithm 3 ROOT-SGD

Input: initialization θ_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: θ_T

Algorithm 4 ROOT-SGD, multi-epoch version

Input: initialization θ_0; step-size sequence $(\eta_t)_{t\geq 1}$; burn-in time $\mathscr{B}$; short epochs length $T_b \geq \mathscr{B}$; short epochs number $\mathscr{E}$
Set initialization for first epoch $\theta_0^{(1)} = \theta_0$
for $b = 1, 2, \cdots, \mathscr{E}$ **do**
  Run ROOT-SGD (Algorithm 5) with burn-in time $\mathscr{B}$ (meaning $\eta_t = 0$ for $t \leq \mathscr{B} - 1$) for T_b iterates
  Set the initialization $\theta_0^{(b+1)} := \theta_{T_b}^{(b)}$ for the next epoch
end for
Run ROOT-SGD (Algorithm 5) for $T := N - T_b\mathscr{E}$ iterates with burn-in time $\mathscr{B}$
Output: The final iterate estimator $\theta_N^{\text{final}} := \theta_T^{(\mathscr{E}+1)}$

M and vector v, we use $T[M]$ to denote the $(k-2)$-th order tensor obtained by applying T to matrix M, and similarly, we use $T[v]$ to denote the $(k-1)$-th order tensor obtained by applying T to vector v.

2.2 Main results

In this section, we describe the algorithm and explain the connection and differences between our results and Li et al. (2020).

2.2.1 The ROOT-SGD algorithm revisited

For the stochastic optimization problem in the strongly-convex and smooth setup, Li et al. (2020) recently proposed a variance-reduced 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

$$v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right) \tag{2.4a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t, \tag{2.4b}$$

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

For the purposes of stabilizing the iterates, Algorithm (3.5) is initialized with a *burn-in* phase of length $\mathscr{B} > 1$, 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 $t = 1, \ldots, \mathscr{B}$, and compute

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

After T steps the last iterate θ_T is used as the output of the algorithm.

Li et al. (2020) analyzed this algorithm when it is run with a constant step-size, 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 step-size choice. In this chapter, we provide a sharper analysis that yields non-asymptotic bounds matching the asymptotic behavior in its leading-order term, with lower-order additional terms being sharp and state-of-the-art. Our work is also motivated by the practical question of step-size schedule in ROOT-SGD. The asymptotic and non-asymptotic guarantees are established for a spectrum of rate of decaying step-sizes. The optimal trade-off between fast convergence and well-behaved limiting variance is also addressed, leading to the optimal choice of step-size sequences under different regimes.

Building upon the proof techniques in the non-asymptotic bounds of Li et al. (2020), our work provide fine-grained guarantees for ROOT-SGD, addressing both aforementioned questions immediately before introducing ROOT-SGD with affirmative answers. 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.

2.2.2 Asymptotic results under Lipschitz continuity of the stochastic Hessians

In this section, we focus on the **LSN** setting, and present asymptotic guarantees for the single-loop version of Algorithm 3. We first recall the assumptions listed in Li et al. (2020):

Assumption 5 (Strong convexity and smoothness) *The population objective objective function F is twice continuously differentiable, μ-strongly-convex and L-smooth for some $0 < \mu \leq L < \infty$:*

$$\left\|\nabla F(\theta) - \nabla F(\theta')\right\|_2 \leq L \left\|\theta - \theta'\right\|_2,$$

$$\left\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \right\rangle \geq \mu \left\|\theta - \theta'\right\|_2^2,$$

Algorithm 5 ROOT-SGD

1: **Input:** initialization θ_0; step-size sequence $(\eta_t)_{t\geq 1}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right)$
4: $\quad \theta_t = \theta_{t-1} - \eta_t v_t$
5: **end for**
6: **Output:** θ_T

Algorithm 6 ROOT-SGD, multi-epoch version

1: **Input:** initialization θ_0; step-size sequence $(\eta_t)_{t\geq 1}$; burn-in time $\mathscr{B}$; short epochs length $T_b \geq \mathscr{B}$; short epochs number $\mathscr{E}$
2: Set initialization for first epoch $\theta_0^{(1)} = \theta_0$
3: **for** $b = 1, 2, \cdots, \mathscr{E}$ **do**
4: $\quad$ Run ROOT-SGD (Algorithm 5) with burn-in time $\mathscr{B}$ (meaning $\eta_t = 0$ for $t \leq \mathscr{B} - 1$) for T_b iterates
5: $\quad$ Set the initialization $\theta_0^{(b+1)} := \theta_{T_b}^{(b)}$ for the next epoch
6: **end for**
7: Run ROOT-SGD (Algorithm 5) for $T := N - T_b\mathscr{E}$ iterates with burn-in time $\mathscr{B}$
8: **Output:** The final iterate estimator $\theta_N^{\text{final}} := \theta_T^{(\mathscr{E}+1)}$

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

Assumption 6 (Finite variance at optimality) *At the minimizer* θ^**, the stochastic gradient* $\nabla f(\theta^*; \xi)$ *has a positive definite covariance matrix* $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*; \xi)(\nabla f(\theta^*; \xi))^\top\right]$*, and its trace* $\sigma_*^2 := \mathbb{E}\|\nabla f(\theta^*; \xi)\|_2^2$ *is finite.*

Assumption 7 (Lipschitz stochastic noise) *The noise function* $\theta \mapsto \varepsilon(\theta; \xi)$ *in the associated stochastic gradients satisfies the bound*

$$\mathbb{E}\left\|\varepsilon(\theta; \xi) - \varepsilon(\theta'; \xi)\right\|_2^2 \leq \ell_\Xi^2 \left\|\theta - \theta'\right\|_2^2 \qquad \text{for all pairs } \theta, \theta' \in \mathbb{R}^d. \tag{2.5}$$

We first introduce our one-point Hessian continuity assumption as follows:

Assumption 8 (One-point Hessian continuity) *The Hessian mapping* $\nabla^2 F(\theta)$ *is continuous at the minimizer* θ^**, i.e.,*

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

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

We conclude in the following Theorem 2.1 that ROOT-SGD with constant step-size that properly scales down with N converges asymptotically to the optimal Gaussian limit as $N \to \infty$.:

Theorem 2.1 (Asymptotic normality, Li et al. (2020)). *Under Assumptions 5, 14, 7 and 15, for any* $\alpha \in (0, 1)$*, the multi-epoch estimator produced by Algorithm 6 with burn-in time* $\mathscr{B} = \frac{24}{\eta\mu}$*, short-epoch length* $T_b = \frac{7340}{\eta\mu}$ *and number of short*

epoches $\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$. *Then as* $N - T_b\mathscr{E} \to \infty$, $\eta \to 0$ *such that* $\eta(N - T_b\mathscr{E}) \to \infty$ *the following weak convergence holds:*

$$\sqrt{T}\left(\theta_N^{final,(\eta)} - \theta^*\right) \xrightarrow{d} \mathscr{N}\left(0, [\nabla^2 F(\theta^*)]^{-1}\Sigma^*[\nabla^2 F(\theta^*)]^{-1}\right), \tag{2.6}$$

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.*

Theorem 2.1 have essentially been proved in Li et al. (2020), and we forgo its proof. 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 matrices at θ^*, another standard condition for asymptotic normality.

Now, in order obtain asymptotic results for the constant step size algorithms, we impose the following slightly stronger assumptions on the smoothness of stochastic gradients and moments:

(CLT.A) For any $\theta \in \mathbb{R}^d$ we have

$$\sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left\|(\nabla^2 f(\theta;\xi) - \nabla^2 f(\theta^*;\xi))v\right\|_2^2 \le \beta^2\|\theta - \theta^*\|_2^2. \tag{2.7a}$$

(CLT.B) The fourth moments of the stochastic gradient vectors at θ^* exist, and in particular we have

$$\mathbb{E}\|\nabla f(\theta^*;\xi)\|_2^4 < \infty, \qquad \text{and} \qquad \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left\|\nabla^2 f(\theta^*;\xi)v\right\|_2^4 \le \ell_\Xi'^4 < \infty. \tag{2.7b}$$

Note that both conditions are imposed in a one-point fashion with respect to the optimal point θ^*, instead of 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:

$$\Lambda H^* + H^*\Lambda - \eta\mathbb{E}\left[\Xi(\theta^*)\Lambda\Xi(\theta^*)\right] - \eta H^*\Lambda H^* = \eta\Sigma^*. \tag{2.8}$$

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.

Theorem 2.2. *Suppose that Assumptions 5, 14, and 7 are satisfied, as are (CLT.A) and (CLT.B). Then there exists constants* c_1, c_2, *given the step size* $\eta \in \left(0, c_1(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L} \wedge \frac{\mu^{1/3}}{\ell_\Xi'^{4/3}})\right)$, *and burn-in time* $\mathscr{B} = \frac{c_2}{\mu\eta}$, *we have:*

$$\sqrt{T}(\theta_T - \theta^*) \xrightarrow{d} \mathscr{N}\left(0, (H^*)^{-1}\left(\Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right)(H^*)^{-1}\right).$$

See §2.3.1 for the proof of this theorem.

A few remarks are in order. First, we note that the asymptotic covariance is the sum of the optimal covariance $(H^*)^{-1}\Sigma^*(H^*)^{-1}$ and an additional correction term defined in Eq. (2.8). The correction term is exactly the same as that of the constant step size version of the Polyak-Juditsky-Ruppert algorithm derived in Mou et al. (2020), while our theorem applies to more general nonlinear stochastic approximation problems. As the step size decreases, the correction terms tends to zero, which follows from the following trace bound (see Mou et al. (2020)):

$$\mathrm{Tr}\left((H^*)^{-1}\mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right](H^*)^{-1}\right) \le \eta\frac{\ell_\Xi^2\sigma_*^2}{\mu^3}.$$

Indeed, using a slowly decreasing step size sequence, one can prove an asymptotic result that matches the Cramér-Rao lower bound. Second, we note that Theorem 2.2 has an additional requirement on the step size, needing it to be upper bounded by $\frac{\mu^{1/3}}{\ell_\Xi'^{4/3}}$. We note that 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.

2.2.3 Non-asymptotic upper bounds under Hölder continuity of the Hessians

We need the following *one-point Hölder continuity condition* for the Hessian, as a quantitative counterpart of the continuity assumption:

Assumption 9 (Hölder continuous Hessians with exponent γ) *There exists an exponent $\gamma \in (0,1]$ and a constant $L_\gamma > 0$ such that*

$$\left\|\nabla^2 F(\theta) - \nabla^2 F(\theta^*)\right\|_{op} \le L_\gamma \|\theta-\theta^*\|_2^\gamma. \tag{2.9}$$

We also need the following stronger fourth moment conditions for technical reasons. Note that these conditions are also exploited in prior works Bach & Moulines (2011); Gadat & Panloup (2017).

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

$$\sqrt{\mathbb{E}\|\varepsilon(\theta_1;\xi)-\varepsilon(\theta_2;\xi)\|_2^4} \le \tilde{\ell}_\Xi^2\|\theta_1-\theta_2\|_2^2, \quad \textit{for all pairs } \theta_1,\theta_2 \in \mathbb{R}^d. \tag{2.10}$$

Assumption 11 *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has a positive definite covariance matrix, and $\tilde{\sigma}_*^2 := \sqrt{\mathbb{E}\|\nabla f(\theta^*;\xi)\|_2^4}$ is finite.*

These assumptions are the higher-order correspondence to Assumptions 14 and 7, separately. We will observe later that a Hölder continuity condition characterizes to what degree our algorithm resembles linear stochastic approximation. We conclude the following main result:

Theorem 2.3.A (Improved nonasymptotic result, single-epoch ROOT-SGD, Hölder continuous Hessians). *For F satisfying Assumptions 5, 14 and 7 and 9, there exists a global constant C such that the estimator produced by Algorithm 5 satisfies the following convergence rate upper bound: for $T \geq \mathscr{B}+1$*

$$\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\eta\mu T} + \frac{\log\left(\frac{eT}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}} + \frac{\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2T^2} + \frac{CL_\gamma\left\|\nabla F(\theta_0)\right\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(7+4\gamma)/2}T^{(5+2\gamma)/2}}. \tag{2.11}$$

Junchi — Add comments Junchi — Here, we absorb the cross term in its left and right neighborhood terms so that $\frac{2\ell_\Xi\sigma_*}{\mu T}\cdot\frac{\mathscr{B}}{T}\left\|\nabla F(\theta_0)\right\|_2 \leq \frac{6\ell_\Xi^2\sigma_*^2}{\mu^2T^2} + \frac{\mathscr{B}^2}{6T^2}\left\|\nabla F(\theta_0)\right\|_2^2$ Junchi — OLD

$$\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left(\frac{\eta\ell_\Xi^2}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\eta\mu T} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + C\frac{L_\nu\tilde{\sigma}_*^{2+\nu}}{\mu^{3/2+\nu}\eta^{1/2}T^{(3+\nu)/2}} + C\left(\frac{1}{\eta^2\mu^2T^2}\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{\ell_\Xi\sigma_*}{\eta\mu^2T^2}\left\|\nabla F(\theta_0)\right\|_2\right) + C\frac{L_\nu\left\|\nabla F(\theta_0)\right\|_2^{2+\nu}}{\eta^{5/2+\nu}\mu^{7/2+2\nu}T^{5/2+\nu}} \tag{2.11'}$$

Theorem 2.3.B (Improved upper bound on gradient norm, multi-epoch ROOT-SGD, Hölder continuous Hessians). *For F satisfying Assumptions 5, 14 and 7, and let the number of short epochs $\mathscr{E} = \left\lceil \frac{1}{2}\log\left(\frac{e\left\|\nabla F(\theta_0)\right\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$, the burn-in time $\mathscr{B} = \frac{24}{\eta\mu}$, and the small epoch length $T_b(\eta) = \frac{7340}{\eta\mu}$. Then the multi-epoch estimator produced by Algorithm 6 satisfies the following bound: for $N \geq T_b(\eta)\mathscr{E}+1$*

$$\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left(\frac{1}{\eta\mu T} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\ell_\Xi^2\eta}{\mu} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}}, \tag{2.12}$$

where $T := N - T_b(\eta)\mathscr{E}$, and $C>0$ is a global constant. By further taking the step-size $\eta = \frac{c}{\mu\left(\ell_\Xi^2/\mu^2\right)^{1-\alpha}N^\alpha} \wedge \frac{1}{4L}$ where $c = 9.49$, $\alpha := \frac{1+2\gamma}{3} \wedge \frac{1}{2}$, we have

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_N^{final})\right\|_2^2 - \frac{\sigma_*^2}{N} &\leq C\left(\left(\frac{\ell_\Xi}{\mu\sqrt{N}}\right)^{2\alpha} + \frac{\log\left(\frac{eN}{\mathscr{B}}\right)\ell_\Xi}{\mu\sqrt{N}} + \frac{L}{\mu N}\right)\frac{\sigma_*^2}{N} \\ &\quad + \frac{C(\mu^{(2\alpha-1)/2}\ell_\Xi^{1-\alpha}N^{\alpha/2} + L^{1/2})L_\gamma\tilde{\sigma}_*^{2+\gamma}}{\mu^{(3+2\gamma)/2}N^{(3+\gamma)/2}} \\ &\leq C\left(\log\left(\frac{eN}{\mathscr{B}}\right)\left(\frac{\ell_\Xi}{\mu\sqrt{N}}\right)^{2\alpha} + \frac{L}{\mu N}\right)\frac{\sigma_*^2}{N} \\ &\quad + \frac{C(\mu^{(2\alpha-1)/2}\ell_\Xi^{1-\alpha}N^{\alpha/2} + L^{1/2})L_\gamma\tilde{\sigma}_*^{2+\gamma}}{\mu^{(3+2\gamma)/2}N^{(3+\gamma)/2}}.\end{aligned} \tag{2.13}$$

See §2.3.2 and §2.3.3 for the proof of the above two theorems, separately. We remark that in our proof of Theorem 2.3.B, the leading-order higher-order term spells $\tilde{O}\left(\left(\frac{\ell_\Xi}{\mu\sqrt{N}}\right)^{2\alpha}\frac{1}{N}\right)$ as in gradient norm squared in (2.13). When $\gamma = 1$, we obtain the optimal upper bound in squared gradient norm with a higher-order $O(N^{-3/2})$ additional term on top of the optimal statistical risk, which matches the MLE efficiency (van der Vaart, 2000).

2.3 Convergence rate analysis (Proof of main results)

We provide the convergence rate analysis and the proofs of our theorems in this section. We inherit from the analysis in Li et al. (2020) and utilize the central object in our analysis is the *tracking error process*, defined as

$$z_t := v_t - \nabla F(\theta_{t-1}), \qquad \text{for } t \geq \mathscr{B}. \tag{2.14}$$

In our analysis we heavily use the fact that the process $(tz_t)_{t\geq\mathscr{B}}$ is a martingale adapted to the natural filtration.

2.3.1 Proof of Theorem 2.2

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:

$$t(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) = t\int_0^1 \Xi_t\left(\gamma\theta_{t-2} + (1-\gamma)\theta_{t-1}\right)(\theta_{t-1} - \theta_{t-2})d\gamma.$$

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 $\{tv_t\}_{t \geq \mathscr{B}}$ to a stationary distribution, then the asymptotic result follows from the Birkhoff ergodic theorem and a martingale CLT. While the process $\{tv_t\}_{t \geq \mathscr{B}}$ 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 $\{y_t\}_{t \geq \mathscr{B}}$, initialized as $y_{\mathscr{B}} = \mathscr{B} v_{\mathscr{B}}$ and updated as

$$y_t = y_{t-1} - \eta H_t(\theta^*) y_{t-1} + \varepsilon_t(\theta^*), \quad \text{for all } t \geq \mathscr{B}+1. \tag{2.15}$$

Note that $\{y_t\}_{t \geq \mathscr{B}}$ is a time-homogeneous Markov process that is coupled to $\{(\theta_t, v_t, z_t)\}_{t \geq \mathscr{B}}$. We have the following coupling estimate:

Lemma 2.4. *Supposing that Assumptions 5, 14 and 7, as well as Conditions (CLT.A) and (CLT.B) hold, then for any iteration $t \geq \mathscr{B}$ and any step size $\eta \in (0, \frac{1}{2L} \wedge \frac{\mu}{2\ell_\Xi^2})$, 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, ℓ'_Ξ, μ, β and the step size η, but independent of t.

See §2.5.1 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 2.5. *Suppose that Assumptions 5, 14 and 7, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq \mathscr{B}$, any step size $\eta \in (0, \frac{1}{2L} \wedge \frac{\mu}{\ell_\Xi^2})$***Junchi — This η condition is weird** *and for any random vector $\tilde{\theta}_{t-1} \in \mathscr{F}_{t-1}$, we have*

$$\mathbb{E} \left\| \left[H_t(\tilde{\theta}_{t-1}) - H_t(\theta^*) \right] y_{t-1} \right\|_2^2 \leq c_1 \sqrt{\mathbb{E} \left\| \tilde{\theta}_{t-1} - \theta^* \right\|_2^2},$$

where c_1 is a constant independent of t and the choice of $\tilde{\theta}_{t-1}$.

See §2.5.2 for the proof of this lemma.

Finally, the following lemma characterizes the behavior of the process $\{y_t\}_{t \geq \mathscr{B}}$ defined in Eq. (2.15):

Lemma 2.6. *Suppose that Assumptions 5, 14 and 7, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq \mathscr{B}$ and any step size $\eta \in (0, \frac{1}{2L} \wedge \frac{\mu}{16\ell_\Xi^2} \wedge \frac{\mu^{1/3}}{6(\ell'_\Xi)^{4/3}})$, we have*

$$\mathbb{E}(y_t) = 0 \quad \textit{for all } t \geq \mathscr{B} \qquad \textit{and} \qquad \sup_{t \geq \mathscr{B}} \mathbb{E} \|y_t\|_2^4 < a',$$

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

$$H^* Q_\eta + Q_\eta H^* - \eta(H^* Q_\eta H^* + \mathbb{E}(\Xi(\theta^*) Q_\eta \Xi(\theta^*))) = \tfrac{1}{\eta}\Sigma^*.$$

See §2.5.3 for the proof of this lemma.

Taking these three lemmas as given, we now proceed with the proof of Theorem 2.2. We first define two auxiliary processes:

$$N_T := \sum_{t=\mathscr{B}+1}^{T} \varepsilon_t(\theta^*), \qquad \Upsilon_T := \eta \sum_{t=\mathscr{B}+1}^{T} \Xi_t(\theta^*) y_{t-1}.$$

Observe that both N_T and Υ_T are martingales adapted to $(\mathscr{F}_t)_{t\geq\mathscr{B}}$. 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 1.2 in Li et al. (2020), define $a_0 := \frac{28\sigma_*^2}{\mu^2} + \frac{2700}{\eta^2\mu^4\mathscr{B}}\|\nabla F(\theta_0)\|_2^2$, we have:

$$\mathbb{E}\,\|\theta_t - \theta^*\|_2^2 \leq \frac{1}{\mu^2}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \frac{2700\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^4(t+1)^2} + \frac{28\,\sigma_*^2}{\mu^2(t+1)} \leq \frac{a_0}{t+1}, \quad \text{for all } t \geq \mathscr{B}. \tag{2.16}$$

Applying the bound (2.16) with Assumption 7, we have

$$\mathbb{E}\,\|M_T - N_T\|_2^2 = \sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 \leq \ell_\Xi^2 \sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\,\|\theta_{t-1} - \theta^*\|_2^2 \leq a_0 \ell_\Xi^2 \log T. \tag{2.17}$$

For the process Υ_T, by Cauchy-Schwartz inequality, we have:

$$\begin{aligned}
\mathbb{E}\,\|\Psi_T - \Upsilon_T\|_2^2 &= \sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\,\|\eta\,\Xi_t(\theta^*) y_{t-1} - (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2}))\|_2^2 \\
&\leq \eta^2 \sum_{t=\mathscr{B}+1}^{T} \mathbb{E} \int_0^1 \|\Xi_t(\theta^*) y_{t-1} - \Xi_t(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2}; \xi_t)(t-1) v_{t-1}\|_2^2 d\gamma \\
&\leq I_I + I_2,
\end{aligned}$$

where we define

$$\begin{aligned}
I_1 &:= 2\eta^2 \sum_{t=\mathscr{B}+1}^{T} \mathbb{E} \int_0^1 \|(\Xi_t(\theta^*) - \Xi_t(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2}))\, y_{t-1}\|_2^2 d\gamma, \quad \text{and} \\
I_2 &:= 2\eta^2 \sum_{t=\mathscr{B}+1}^{T} \mathbb{E} \int_0^1 \|\Xi_t(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2})(y_{t-1} - (t-1)v_{t-1})\|_2^2 d\gamma.
\end{aligned}$$

We bound each of these two terms in succession.

Bound on I_1:

In order to bound the term I_1, we apply Lemma 2.5 with the choice

$$\tilde{\theta}_{t-1} = \gamma\theta_{t-1} + (1-\gamma)\theta_{t-2} \in \mathscr{F}_{t-1},$$

so as to obtain

$$\mathbb{E}\left\|\left(H_t(\tilde{\theta}_{t-1}) - H_t(\theta^*)\right) y_{t-1}\right\|_2^2 \leq c_1\sqrt{\mathbb{E}\left\|\tilde{\theta}_t - \theta^*\right\|_2^2}.$$

Applying the Cauchy-Schwartz inequality yields

$$\mathbb{E}\left\|\left(\nabla^2 F(\tilde{\theta}_{t-1}) - \nabla^2 F(\theta^*)\right) y_{t-1}\right\|_2^2 \leq \mathbb{E}\left\|\left(H_t(\tilde{\theta}_{t-1}) - H_t(\theta^*)\right) y_{t-1}\right\|_2^2 \leq c_1\sqrt{\mathbb{E}\left\|\tilde{\theta}_t - \theta^*\right\|_2^2}.$$

Putting the two bounds together, we obtain:

$$\begin{aligned}\mathbb{E}\left\|\left(\Xi_t(\tilde{\theta}_{t-1}) - \Xi_t(\theta^*)\right) y_{t-1}\right\|_2^2 &\leq 2\mathbb{E}\left\|\left(H_t(\tilde{\theta}_{t-1}) - H_t(\theta^*)\right) y_{t-1}\right\|_2^2 + 2\mathbb{E}\left\|\left(\nabla^2 F(\tilde{\theta}_{t-1}) - \nabla^2 F(\theta^*)\right) y_{t-1}\right\|_2^2 \\ &\leq 4c_1\sqrt{\mathbb{E}\left\|\tilde{\theta}_t - \theta^*\right\|_2^2}.\end{aligned}$$

Thus, we find that

$$\begin{aligned}\mathbb{E}\left\|(\Xi_t(\theta^*) - \Xi_t(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2}))\, y_{t-1}\right\|_2^2 &\leq 4c_1\sqrt{\mathbb{E}\left\|\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2} - \theta^*\right\|_2^2} \\ &\leq 4c_1\left(\sqrt{\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^2} + \sqrt{\mathbb{E}\left\|\theta_{t-2} - \theta^*\right\|_2^2}\right) \\ &\leq 4c_1\sqrt{a_0}\left(\frac{1}{\sqrt{t-1}} + \frac{1}{\sqrt{t-2}}\right) \\ &\leq \frac{16c_1\sqrt{a_0}}{\sqrt{t}},\end{aligned}$$

where in the last step we used the inequality (2.16). Summing over $t \in [\mathscr{B}, T]$ yields the bound

$$I_1 \leq 2\eta^2 \sum_{t=\mathscr{B}+1}^{T} \frac{16c_1\sqrt{a_0}}{\sqrt{t}} \leq 64\eta^2 c_1\sqrt{a_0 T}.$$

Bound on I_2:

Turning to the term I_2, by Assumption 7 and Lemma 2.4, we note that:

$$I_2 \le 2\eta^2 \sum_{t=\mathscr{B}+1}^{T} \ell_\Xi^2 \mathbb{E}\|y_{t-1} - (t-1)v_{t-1}\|_2^2 \le 2\eta^2 \ell_\Xi^2 \sum_{t=\mathscr{B}+1}^{T} \frac{c_0}{\sqrt{t}} \le 4\eta^2 \ell_\Xi^2 c_0 \sqrt{T}.$$

Putting these inequalities together, we conclude that:

$$\mathbb{E}\|\Psi_T - \Upsilon_T\|_2^2 \le (64\eta^2 c_1 \sqrt{a_0} + 4\eta^2 \ell_\Xi^2 c_0)\sqrt{T}. \tag{2.18}$$

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 $v_t := \varepsilon_t(\theta^*) + \eta \Xi_t(\theta^*) y_{t-1}$. By definition, $N_T + \Upsilon_T = \sum_{t=\mathscr{B}}^{T} v_t$, and we have:

$$\begin{aligned}\mathbb{E}(v_t v_t^\top) = \mathbb{E}(\varepsilon_t(\theta^*)\varepsilon_t(\theta^*)^\top) + \mathbb{E}\left(\Xi_t(\theta^*) y_{t-1} y_{t-1}^\top \Xi_t(\theta^*)^\top\right) \\ + \mathbb{E}\left(\varepsilon_t(\theta^*) y_{t-1}^\top \Xi_t(\theta^*)^\top\right) + \mathbb{E}\left((\Xi_t(\theta^*) y_{t-1} \varepsilon_t(\theta^*)^\top\right).\end{aligned}$$

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 2.6, we note that the time-homogeneous Markov process $\{y_t\}_{t \ge \mathscr{B}}$ converges asymptotically to a stationary distribution with covariance Q_η. Invoking the Birkhoff ergodic theorem, we have:

$$\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\left(\Xi_t(\theta^*) y_{t-1} y_{t-1}^\top \Xi_t(\theta^*)^\top \mid \mathscr{F}_{t-1}\right) = \mathbb{E}(\Xi(\theta^*) \otimes \Xi(\theta^*)) \left[\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} y_{t-1} y_{t-1}^\top\right] \xrightarrow{p} \mathbb{E}\left(\Xi(\theta^*) Q_\eta \Xi(\theta^*)^\top\right).$$

For the cross terms, we note that:

$$\mathbb{E}\left(\varepsilon_t(\theta^*) y_{t-1}^\top \Xi_t(\theta^*)^\top | \mathscr{F}_{t-1}\right) = \mathbb{E}(\varepsilon(\theta^*) \otimes \Xi(\theta^*))[y_{t-1}].$$

Note that by Lemma 2.6, we have $\mathbb{E}(y_t) = 0$ for any $t \ge \mathscr{B}$. By the weak law of large numbers, we have $\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} y_t \xrightarrow{p} 0$. Putting together these inequalities, we find that

$$\begin{aligned}\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\left(v_t v_t^\top \mid \mathscr{F}_{t-1}\right) = \frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} \Big(\Sigma^* + \eta^2 \mathbb{E}(\Xi(\theta^*) \otimes \Xi(\theta^*))[y_t y_t^\top] \\ + \eta \mathbb{E}(\varepsilon(\theta^*) \otimes \Xi(\theta^*))[y_{t-1}] + \eta \mathbb{E}(\Xi(\theta^*) \otimes \varepsilon(\theta^*))[y_{t-1}]\Big),\end{aligned}$$

and hence the random matrix $\frac{1}{T}\sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\left(v_t v_t^\top \mid \mathscr{F}_{t-1}\right)$ 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 Hall & Heyde (1980)). It remains to verify the conditional Lindeberg condition. Indeed, for any $\varepsilon > 0$, a straightforward calculation yields:

$$R_T(\varepsilon) := \sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\left(\left\| \frac{v_t}{\sqrt{T}} \right\|_2^2 \mathbf{1}_{\left\| \frac{v_t}{\sqrt{T}} \right\|_2 > \varepsilon} | \mathscr{F}_{t-1} \right)$$
$$\overset{(i)}{\leq} \frac{1}{T} \sum_{t=\mathscr{B}+1}^{T} \sqrt{\mathbb{E}\left(\|v_t\|_2^4 | \mathscr{F}_{t-1} \right)} \cdot \sqrt{\mathbb{P}\left(\|v_t\|_2 > \varepsilon\sqrt{T} | \mathscr{F}_{t-1} \right)} \overset{(ii)}{\leq} \frac{1}{T} \sum_{t=\mathscr{B}+1}^{T} \frac{1}{(\varepsilon\sqrt{T})^2} \mathbb{E}\left(\|v_t\|_2^4 | \mathscr{F}_{t-1} \right).$$

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:

$$\mathbb{E}\left(\|v_t\|_2^4 | \mathscr{F}_{t-1} \right) \leq 8\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + 8(\ell'_\Xi)^4 \|y_{t-1}\|_2^4.$$

Plugging back to the upper bound for $R_T(\varepsilon)$, and applying Lemma 2.6, as $T \to \infty$, we have:

$$\mathbb{E}[R_T(\varepsilon)] \leq \frac{8}{T\varepsilon^2}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + \frac{8(\ell'_\Xi)^4}{T^2\varepsilon^2} \sum_{t=\mathscr{B}+1}^{T} \mathbb{E}\|y_{t-1}\|_2^4 \leq \frac{8}{T\varepsilon^2}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + \frac{8(\ell'_\Xi)^4}{T\varepsilon^2} a' \to 0.$$

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:

$$\frac{N_T + \Upsilon_T}{\sqrt{T}} \xrightarrow{d} \mathscr{N}\left(0, \Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)\right]\right).$$

By the second-moment estimates (2.17) and (2.18), we have:

$$\frac{\|\Upsilon_T - \Psi_T\|_2}{\sqrt{T}} \xrightarrow{p} 0, \quad \frac{\|M_T - N_T\|_2}{\sqrt{T}} \xrightarrow{p} 0.$$

With the burn-in time $\mathscr{B}$ fixed, we also have $\frac{\mathscr{B}}{T} z_{\mathscr{B}} \xrightarrow{p} 0$. By Slutsky's theorem, we have:

$$\sqrt{T} z_T \xrightarrow{d} \mathscr{N}\left(0, \Sigma^* + \mathbb{E}\left(\Xi(\theta^*)\Lambda_\eta \Xi(\theta^*)^\top\right)\right).$$

Note that $\nabla F(\theta_{t-1}) = v_t - z_t$. By Lemma 2.4 and Lemma 2.6, 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

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

By Assumption (CLT.A), we have:

$$\forall v\in\mathbb{S}^{d-1},\ \theta\in\mathbb{R}^d\quad\left\|(\nabla^2F(\theta)-\nabla^2F(\theta^*))v\right\|_2^2\le\mathbb{E}\left\|(\nabla^2f(\theta;\xi)-\nabla^2f(\theta^*;\xi))v\right\|_2^2\le\beta^2\|\theta-\theta^*\|_2^2.$$

Consequently, we have the bound:

$$\|\nabla F(\theta)-H^*(\theta-\theta^*)\|_2\le\|\theta-\theta^*\|_2\cdot\sup_{\|\theta'-\theta^*\|_2\le\|\theta-\theta^*\|_2}\ \sup_{v\in\mathbb{S}^{d-1}}\left\|(\nabla^2F(\theta')-H^*)v\right\|_2\le\beta\|\theta-\theta^*\|_2^2.$$

By Eq (2.16), we have $\sqrt{T}\|\nabla F(\theta_T)-H^*(\theta_T-\theta^*)\|_2\xrightarrow{p}0$. Invoking Slutsky's theorem, this leads to $\sqrt{T}H^*(\theta_T-\theta^*)\xrightarrow{d}\mathcal{N}\left(0,\Sigma^*+\mathbb{E}\left(\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)^\top\right)\right)$, and consequently,

$$\sqrt{T}(\theta_T-\theta^*)\xrightarrow{d}\mathcal{N}\left(0,(H^*)^{-1}\left(\Sigma^*+\mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)^\top\right]\right)(H^*)^{-1}\right),$$

which finishes the proof.

2.3.2 Proof of Theorem 2.3.A

We prepare to prove Theorem 2.3.A. The following result is a higher-order-moment analogy to the one in Theorem 1.2 in Li et al. (2020):

Lemma 2.7 (Bound on $\nabla F(\theta_{t-1})$). *Under Assumptions 5, 10 and 11. We use burn-in time $\mathscr{B}=\frac{1}{\eta\mu}$. Then for any $T\ge\mathscr{B}$, the iterates θ_T from the ROOT-SGD algorithm satisfies the bound*

$$\left(\mathbb{E}\|\nabla F(\theta_T)\|_2^4\right)^{1/2}\le\frac{140\tilde{\sigma}_*^2}{T+1}+\frac{6250\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2(T+1)^2}.\tag{2.19}$$

Proof can be found in §2.6.1.

Remark 2.1. On a very intuitive and high-level viewpoint, when comparing with the Polyak-Ruppert-Juditsky analysis, we can think of the $(\eta t v_t : t\ge 0)$ 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. Although driving by the same stochasticity, they are "asymptotically independent" 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. Also, we can inject randomness on θ_0 as long as it is $\mathscr{F}_0$-measurable.

Lemma 2.8 (Sharp bound on v_t, Hölder continuous Hessians). *Under the setting of Theorem 2.3 we have the following bound for $T \geq \mathscr{B}+1$*

$$\sqrt{\mathbb{E}\|v_T\|_2^4} \leq \frac{4484\tilde{\sigma}_*^2}{\eta\mu T^2} + \frac{1359375}{\eta^4\mu^4T^4}\|\nabla F(\theta_0)\|_2^2. \tag{2.20}$$

Proof can be found in §2.6.2.

By definition of z_t, we obtain the equation that decomposes $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$ into three terms:

$$\mathbb{E}\|\nabla F(\theta_{t-1})\|^2 = \mathbb{E}\|v_t - z_t\|^2 = \mathbb{E}\|v_t\|^2 + \mathbb{E}\|z_t\|^2 - 2\mathbb{E}\langle v_t, z_t\rangle \tag{2.21}$$

In the next steps, 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. The main focus will be on bounding the cross term:

(i) For general non-quadratic case, we have

$$\begin{aligned} tv_t &= (t-1)v_{t-1} + (t-1)\left(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right) + \nabla f(\theta_{t-1};\xi_t) \\ tz_t &= (t-1)z_{t-1} + (t-1)\left(\delta_t(\theta_{t-1}) - \delta_t(\theta_{t-2})\right) + \delta_t(\theta_{t-1}) - \delta_t(\theta^*) + \delta_t(\theta^*). \end{aligned}$$

We subtract off a $(t-\tilde{T}^*)z_{t-\tilde{T}^*}$ term the tz_t expression above, and decompose the absolute value of the cross term $|\mathbb{E}\langle v_t, tz_t\rangle|$ as:

$$|\mathbb{E}\langle tz_t, v_t\rangle| \leq (t-\tilde{T}^*)\underbrace{\left|\mathbb{E}\left\langle z_{t-\tilde{T}^*}, v_t\right\rangle\right|}_{\mathrm{I}_1} + \underbrace{\left|\mathbb{E}\left\langle tz_t - (t-\tilde{T}^*)z_{t-\tilde{T}^*}, v_t\right\rangle\right|}_{\mathrm{I}_2} \tag{2.22}$$

For bounding I_2 we make use of the recursive rule of tz_t and conclude that

$$\begin{aligned} &\left|\mathbb{E}\left\langle tz_t - (t-\tilde{T}^*)z_{t-\tilde{T}^*}, v_t\right\rangle\right| \leq \sqrt{\mathbb{E}\left\|tz_t - (t-\tilde{T}^*)z_{t-\tilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\|v_t\|_2^2} \\ &\leq \sqrt{\mathbb{E}\|v_t\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\sum_{s=t-\tilde{T}^*+1}^{t}\left[(s-1)(\delta_s(\theta_{s-1}) - \delta_s(\theta_{s-2})) + \delta_s(\theta_{s-1}) - \delta_s(\theta^*) + \delta_s(\theta^*)\right]\right\|_2^2} \\ &\leq c\sqrt{\mathbb{E}\|v_t\|_2^2}\cdot\sqrt{\sum_{s=t-\tilde{T}^*+1}^{t}(s-1)^2\eta^2\ell_\Xi^2\mathbb{E}\|v_{s-1}\|_2^2 + \frac{\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{s-1})\|_2^2 + \sigma_*^2} \\ &\leq c\sqrt{\mathbb{E}\|v_t\|_2^2}\cdot\sqrt{\tilde{T}^*\cdot\left(\sigma_*^2 + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3t^2}\right)} \end{aligned}$$

$$\leq c\sqrt{\tilde{T}^*}\left(\frac{\sigma_*}{\sqrt{\eta\mu t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2 t^2}\right)\left(\sigma_*+\frac{\|\nabla F(\theta_0)\|_2}{\eta^{1.5}\mu^{1.5}t}\right)\leq c\left(\frac{\sigma_*^2}{\eta\mu t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 t^3}\right)\sqrt{\log t}$$

The bound for the term I_1 in the decomposition 2.22 is given by the following analysis:

$$\begin{aligned}\left|\mathbb{E}\left\langle z_{t-\tilde{T}^*}, v_t\right\rangle\right| &\leq \left|\mathbb{E}\left\langle z_{t-\tilde{T}^*}, \mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\rangle\right| \\ &\leq \sqrt{\mathbb{E}\left\|z_{t-\tilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\|_2^2},\end{aligned} \tag{2.23}$$

where the last inequality comes from applying the Cauchy-Schwarz Inequality. For estimating $\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\|_2^2}$, we note that

$$\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})=\int_0^1 \nabla^2 F\left(\lambda\theta_{t-2}+(1-\lambda)\theta_{t-1}\right)(\theta_{t-1}-\theta_{t-2})d\lambda,$$

which leads to the following bound under the Hölder's continuity condition for the Hessians (Assumption 9):

$$\begin{aligned}&\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})-\nabla^2 F(\theta^*)(\theta_{t-1}-\theta_{t-2})\right\|_2 \\ &=\int_0^1\left\|\left(\nabla^2 F\left(\lambda\theta_{t-2}+(1-\lambda)\theta_{t-1}\right)-\nabla^2 F(\theta^*)\right)(\theta_{t-1}-\theta_{t-2})\right\|_2 d\lambda \\ &\leq \eta L_\gamma\|v_{t-1}\|_2\int_0^1\|\lambda(\theta_{t-2}-\theta^*)+(1-\lambda)(\theta_{t-1}-\theta^*)\|_2^\gamma d\lambda \\ &\leq \eta L_\gamma\|v_{t-1}\|_2\cdot\max\left(\|\theta_{t-1}-\theta^*\|_2^\gamma,\|\theta_{t-2}-\theta^*\|_2^\gamma\right),\end{aligned} \tag{2.24}$$

where we applied $(a+b)^\gamma\leq a^\gamma+b^\gamma$ with a,b being two arbitrary positives so $\|v+w\|^\gamma\leq(\|v\|+\|w\|)^\gamma\leq\|v\|^\gamma+\|w\|^\gamma$ with v,w being two arbitrary vectors, due to the mononoticity and convexity of function $-x^\gamma$ for $\gamma\in(0,1]$. Since $H^*=\nabla^2F(\theta^*)$ we have

$$\begin{aligned}&t\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\|_2 \\ &=\left\|\mathbb{E}\left((t-1)\left(v_{t-1}+\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right)+\nabla F(\theta_{t-1}) \mid \mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2 \\ &=\|\mathbb{E}\left((t-1)\left(v_{t-1}+H^*(\theta_{t-1}-\theta_{t-2})\right.\right. \\ &\quad\left.\left.+\ \nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})-H^*(\theta_{t-1}-\theta_{t-2})\right)+\nabla F(\theta_{t-1}) \mid \mathscr{F}_{t-\tilde{T}^*}\right)\|_2 \\ &\leq\left\|\mathbb{E}\left((t-1)\left(v_{t-1}+H^*(\theta_{t-1}-\theta_{t-2})\right) \mid \mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2+\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\|_2 \\ &\quad+\left\|\mathbb{E}\left((t-1)\left(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})-H^*(\theta_{t-1}-\theta_{t-2})\right) \mid \mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2.\end{aligned} \tag{??}$$

Junchi — Fix equation number Further by rearranging the terms, and dividing both sides by $(t-1)$, we obtain

$$\begin{aligned}\left\|\mathbb{E}(v_t \mid \mathscr{F}_{t-\tilde{T}^*})\right\|_2 &\leq\left\|\mathbb{E}\left(v_{t-1}+H^*(\theta_{t-1}-\theta_{t-2}) \mid \mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2 \\ &\quad+\left\|\mathbb{E}\left(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})-H^*(\theta_{t-1}-\theta_{t-2})\right) \mid \mathscr{F}_{t-\tilde{T}^*}\right\|_2\end{aligned}$$

$$\leq (1-\eta\mu)\left\|\mathbb{E}\left(v_{t-1}\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2 + \eta L_\gamma \mathbb{E}\left(\|v_{t-1}\|_2\cdot\max\left(\|\theta_{t-1}-\theta^*\|^\gamma,\|\theta_{t-2}-\theta^*\|^\gamma\right)\mid\mathscr{F}_{t-\tilde{T}^*}\right),$$

where in the last inequality we apply the result in Eq. (2.24). Next by calculating the second moment of both the RHS and the LHS of the above quantity and the Hölder's inequality, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2} \leq (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_{t-1}\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\\
&+\eta L_\gamma\sqrt{\mathbb{E}\left\|\mathbb{E}\left(\|v_{t-1}\|_2^2\cdot\max\left(\|\theta_{t-1}-\theta^*\|_2^{2\gamma},\|\theta_{t-2}-\theta^*\|_2^{2\gamma}\right)\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\\
&\leq (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_{t-1}\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\\
&+\eta L_\gamma\sqrt{\left(\mathbb{E}\|v_{t-1}\|_2^{2+2\gamma}\right)^{\frac{1}{1+\gamma}}\max\left(\mathbb{E}\|\theta_{t-1}-\theta^*\|_2^{2+2\gamma},\mathbb{E}\|\theta_{t-2}-\theta^*\|_2^{2+2\gamma}\right)^{\frac{\gamma}{1+\gamma}}}\\
&\leq (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_{t-1}\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}+\eta L_\gamma\sqrt{\left(\mathbb{E}\left\|v_{t-\tilde{T}^*}\right\|_2^{2+2\gamma}\right)^{\frac{1}{1+\gamma}}}\cdot\sqrt{\left(\mathbb{E}\left\|\theta_{t-\tilde{T}^*}-\theta^*\right\|_2^{2+2\gamma}\right)^{\frac{\gamma}{1+\gamma}}}.
\end{aligned}$$

Recursively applying the above inequality from $t-\tilde{T}^*$ to t and we have that

$$\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\leq(1-\eta\mu)^{\tilde{T}^*}\mathbb{E}\left\|v_{t-\tilde{T}^*}\right\|_2^2+\frac{L_\gamma}{\mu}\sqrt{\left(\mathbb{E}\left\|v_{t-\tilde{T}^*}\right\|_2^{2+2\gamma}\right)^{\frac{1}{1+\gamma}}}\cdot\sqrt{\left(\mathbb{E}\left\|\theta_{t-\tilde{T}^*}-\theta^*\right\|_2^{2+2\gamma}\right)^{\frac{\gamma}{1+\gamma}}}.$$

(eq:condition_vt_bound)

We recall from Lemmas 2.7 and 2.8 the following

$$\left(\mathbb{E}\|v_T\|_2^{2+2\gamma}\right)^{\frac{1}{1+\gamma}}\leq C\left(\frac{\tilde{\sigma}_*{}^2}{\eta\mu T^2}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right),\qquad\left(\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^{2+2\gamma}\right)^{\frac{1}{1+\gamma}}\leq C\left(\frac{\tilde{\sigma}_*{}^2}{T}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2}\right).$$

Bringing this into Eq. (eq:condition_vt_bound) and we have that

$$\begin{aligned}
&\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\leq(1-\eta\mu)^{\tilde{T}^*}\mathbb{E}\left\|v_{t-\tilde{T}^*}\right\|_2^2\\
&+\frac{cL_\gamma}{\mu}\left(\frac{\tilde{\sigma}_*}{\sqrt{\eta\mu}(t-\tilde{T}^*)}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2(t-\tilde{T}^*)^2}\right)\left(\frac{\tilde{\sigma}_*}{\mu\sqrt{t-\tilde{T}^*}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2(t-\tilde{T}^*)}\right)^\gamma
\end{aligned}$$

By taking $\tilde{T}^*$ in $[c\eta^{-1}\mu^{-1}\log T,\frac{T}{2}]$ satisfying $\tilde{T}^*\leq\frac{T}{2}$ and $(1-\eta\mu)^{\tilde{T}^*}\leq e^{-\eta\mu\tilde{T}^*}\leq\frac{1}{t^c}$, the above inequality reduces as follows:

$$\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}\leq\frac{cL_\gamma}{\mu}\left(\frac{\tilde{\sigma}_*}{\sqrt{\eta\mu}t}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2t^2}\right)\left(\frac{\tilde{\sigma}_*}{\mu\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2t}\right)^\gamma$$

Bringing this back to the inequality (2.23) and by utilizing the z_t bound by Lemma 1.8, we have

$$\mathbb{E}\|z_{t-1}\|_2^2 \le \left(1+\frac{20\ell_\Xi^2\eta}{\mu}+\frac{12\ell_\Xi}{\mu t^{1/2}}+\frac{504\log\left(\frac{t}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 t}\right)\frac{\sigma_*^2}{t}+\frac{9\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{t^2}\|\nabla F(\theta_0)\|_2+\frac{183\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{t^2}\|\nabla F(\theta_0)\|_2^2$$
$$\le C\left(\frac{\sigma_*^2}{t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2}\right),$$

and thus

$$\left|\mathbb{E}\left\langle z_{t-\tilde{T}^*},v_t\right\rangle\right| \le \sqrt{\mathbb{E}\left\|z_{t-\tilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\tilde{T}^*}\right)\right\|_2^2}$$
$$\le \frac{cL_\gamma}{\mu}\left(\frac{\sigma_*}{\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu t}\right)\left(\frac{\tilde{\sigma}_*}{\sqrt{\eta\mu}t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2}\right)\left(\frac{\tilde{\sigma}_*}{\mu\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta\mu^2t}\right)^\gamma$$
$$\le \frac{cL_\gamma}{\mu}\left(\frac{\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(1+2\gamma)/2}t^{(3+\gamma)/2}}+\frac{\|\nabla F(\theta_0)\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(5+4\gamma)/2}t^{(5+2\gamma)/2}}\right).$$

Combining the bound for I_1 and I_2 together, we estimate the cross term as:

$$|\mathbb{E}\langle tz_t,v_t\rangle| \le c(t-\tilde{T}^*)\frac{L_\gamma}{\mu}\left(\frac{\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(1+2\gamma)/2}t^{(3+\gamma)/2}}+\frac{\|\nabla F(\theta_0)\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(5+4\gamma)/2}t^{(5+2\gamma)/2}}\right)$$
$$+c\left(\frac{\sigma_*^2}{\eta\mu t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4t^3}\right)\sqrt{\log t}. \tag{eq:last_two}$$

We conclude by dividing both sides of Eq. (eq:last_two) by T and arrive at the following bound:

$$|\mathbb{E}\langle v_T,z_T\rangle| \le c\left(\frac{\sigma_*^2}{\eta\mu T^2}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right)\sqrt{\log t}+cL_\gamma\left(\frac{\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}}+\frac{\|\nabla F(\theta_0)\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(7+4\gamma)/2}T^{(5+2\gamma)/2}}\right).$$

This finishes our bound on the cross term.

(ii) Now we come to obtain the final rate of convergence. Recall that we have sharp bound for z_t in (1.38) of Lemma 1.8 as follows

$$\mathbb{E}\|z_T\|_2^2-\frac{\sigma_*^2}{T}$$
$$\le\left(\frac{20\ell_\Xi^2\eta}{\mu}+\frac{12\ell_\Xi}{\mu\sqrt{T}}+\frac{504\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2T}\right)\frac{\sigma_*^2}{T}+\frac{9\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2+\frac{183\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2^2$$
$$\le C\left(\frac{\ell_\Xi^2\eta}{\mu}+\frac{\ell_\Xi}{\mu\sqrt{T}}+\frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2T}\right)\frac{\sigma_*^2}{T}+C\left(\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2+\frac{\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\|\nabla F(\theta_0)\|_2^2\right).$$

Adding up the three terms, we have

$$\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^2-\frac{\sigma_*^2}{T}=\mathbb{E}\left\|v_T-z_T\right\|_2^2-\frac{\sigma_*^2}{T}=\left(\mathbb{E}\left\|z_T\right\|_2^2-\frac{\sigma_*^2}{T}\right)+\mathbb{E}\left\|v_T\right\|_2^2-2\mathbb{E}\left\langle v_T,z_T\right\rangle$$
$$\leq C\left(\frac{\ell_\Xi^2\eta}{\mu}+\frac{\ell_\Xi}{\mu\sqrt{T}}+\frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2T}\right)\frac{\sigma_*^2}{T}+C\left(\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2+\frac{\ell_\Xi^2}{\mu^2}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$
$$+C\left(\frac{\sigma_*^2}{\eta\mu T^2}+\frac{\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}\right)+6C_0L_\gamma\left(\frac{\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}}+\frac{\left\|\nabla F(\theta_0)\right\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(7+4\gamma)/2}T^{(5+2\gamma)/2}}\right)$$
$$\leq C\left(\frac{\ell_\Xi^2\eta}{\mu}+\frac{\ell_\Xi}{\mu\sqrt{T}}+\frac{1}{\eta\mu T}+\frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2T}\right)\frac{\sigma_*^2}{T}+\frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}}$$
$$+C\left(\frac{\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2T^2}+\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{\mathscr{B}}{T^2}\left\|\nabla F(\theta_0)\right\|_2\right)+C\frac{L_\gamma\left\|\nabla F(\theta_0)\right\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(7+4\gamma)/2}T^{(5+2\gamma)/2}}.$$

This gives Eq. (2.11) and concludes Theorem 2.3.A.

2.3.3 Proof of Theorem 2.3.B

Now we turn to the proof of multi-loop results.

(i) Invoking Eq. (2.19) in Lemma 2.7, we obtain for $b=1,2,\cdots,\mathscr{E}$ the bound for $T_b\geq c\mathscr{B}$:

$$\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2\leq\frac{1}{e^2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2+\frac{c\sigma_*^2}{T_b},\qquad\text{and}\quad\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^4}\leq\frac{1}{e^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^4}+\frac{c\tilde{\sigma}_*^2}{T_b},$$

where our setting of T_b gives a discount factor of $1/e^2$. Solving the recursion, we arrive at the bound:

$$\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2\leq\frac{c\sigma_*^2}{T_b}+e^{-2\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2,\qquad\text{and}\quad\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^4}\leq\frac{c\tilde{\sigma}_*^2}{T_b}+e^{-2\mathscr{E}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}.$$

Our take is $\mathscr{E}\geq\log\frac{T_b\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}}{c\sigma_*^2}$ such that $e^{-2\mathscr{E}}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2\leq\frac{\sigma_*^2}{T_b}$ and $e^{-2\mathscr{E}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}\leq\frac{\tilde{\sigma}_*^2}{T_b}$ both hold. Finally, we have

$$\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^2\leq e^{-2\mathscr{E}}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2+\frac{c\sigma_*^2}{T_b}\leq\frac{c'\sigma_*^2}{T_b},$$
$$\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^4}\leq e^{-2\mathscr{E}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}+\frac{c\tilde{\sigma}_*^2}{T_b}\leq\frac{c'\tilde{\sigma}_*^2}{T_b},$$
$$\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^{2+\gamma}\leq\left(\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2^4\right)^{\frac{2+\gamma}{4}}\leq\frac{c'\tilde{\sigma}_*^{2+\gamma}}{T_b^{(2+\gamma)/2}},$$

$$\mathbb{E}\left\|\nabla F(\theta_0^{(\mathscr{E}+1)})\right\|_2 \leq \frac{c\sigma_*}{T_b^{1/2}},$$

where constants c, c' change from line to line. Substituting this initial condition into the bound (2.11), we obtain the final bound:

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_T^{(\mathscr{E}+1)})\right\|_2^2 - \frac{\sigma_*^2}{T} \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{1}{\eta\mu T} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}} \\
&\quad + C\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{\ell_\Xi\sigma_*\|\nabla F(\theta_0)\|_2}{\eta\mu^2T^2}\right) + \frac{CL_\gamma\|\nabla F(\theta_0)\|_2^{2+\gamma}}{\eta^{(5+2\gamma)/2}\mu^{(7+4\gamma)/2}T^{(5+2\gamma)/2}} \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{1}{\eta\mu T} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}} \\
&\quad + C\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{L_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{(3+\gamma)/2}\mu^{(5+3\gamma)/2}T^{(5+2\gamma)/2}}\right) \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{1}{\eta\mu T} + \frac{\log\left(\frac{T}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}T^{(3+\gamma)/2}},
\end{aligned}$$

which proves the bound (2.12).

(ii) Finally, substituting T by the final epoch length $N - \mathscr{E}T_b$ and adopt similar reasoning as the previous one, we arrive at the conclusion:

$$\mathbb{E}\left\|\nabla F(\theta_N^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{N} \leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu N^{1/2}} + \frac{1}{\eta\mu N} + \frac{\log\left(\frac{N}{\mathscr{B}}\right)\ell_\Xi^2}{\mu^2 N}\right)\frac{\sigma_*^2}{N} + \frac{CL_\gamma\tilde{\sigma}_*^{2+\gamma}}{\eta^{1/2}\mu^{(3+2\gamma)/2}N^{(3+\gamma)/2}},$$

which proves the bound (2.12). Finally, plugging η as given we prove (2.13) and hence Theorem 2.3.B. **Junchi — Fill in as ch1.tex**

2.4 Comparison with related works and lower bound on leading second-order term

In the Lipschitz continuous Hessian case, we can achieve asymptotic unity. We ccompare with Bach & Moulines (2011) and Gadat & Panloup (2017).

(i) For SGD with PRJ averaging, Bach & Moulines (2011) 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 $\eta_t = Ct^{-\alpha}$ for $\alpha \in (1/2, 1)$, Bach & Moulines (2011) show that the following bound holds true for the averaged iterates $\bar{\theta}_T$ for the PRJ:

Algorithm	Assumption	Additional Term	Reference
PRJ	Hessian Lipschitz	$O\left(\frac{1}{T^{7/6}}\right)$	(Bach & Moulines, 2011)
PRJ	Hessian Lipschitz	$O\left(\frac{1}{T^{5/4}}\right)$	(Gadat & Panloup, 2017)
Streaming SVRG	Self-concordant	$O\left(\frac{1}{T^{3/2}}\right)$	(Frostig et al., 2015)
ROOT-SGD	Hessian Lipschitz	$O\left(\frac{1}{T^{3/2}}\right)$	(This work)
ROOT-SGD	Hessian ν-Hölder	$O\left(\frac{1}{T^{(4+2\nu)/3\wedge 3/2}}\right)$	(This work)

Table 2.1 For the unity result, we only describe the additional term on top of the optimal risk. **Junchi — Maybe add one column on convergence criteria/metric; ours can work for any metric Junchi — DC this part, Gadat. Put non-unity result into the tables Junchi — When self-concordant condition is not satisfied for streaming SVRG, the whole convergence rate will inflate by a factor of ℓ_{max}/μ (not only the additional term)**

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

Junchi — DC THIS RATE?! which corresponds to an $O(N^{-7/6})$ additional term in our metrics. Here, the constant c_0 depends on smoothness and strong convexity parameters of second and third order derivatives, as well as higher-order moments of the noise. More recently, Gadat & Panloup (2017) further improves the higher-order term from $O(N^{-7/6})$ to $O(N^{-5/4})$. **Junchi — Discuss more** The convergence rate of (single-loop) ROOT-SGD is similar to SGD with PRJ averaging in the nature of the leading term and the high-order terms, but the rate of ROOT-SGD is much cleaner and easier to interpret.

(ii) In Frostig et al. (2015), a variant of SVRG is proposed that provides nonasymptotic guarantees in terms of the objective gap. In the **ISC** case when the number of samples grows, under an additional self-concordance condition their objective gap bound asymptotically matches the Cramér-Rao lower bound achieved by the empirical risk minimizer.**Junchi — This manuscript does not present ISC case in any corner** However, to get the corresponding nonasymptotic guarantees under such a setting, their bound require a scaling condition $T \gtrsim \frac{\ell_{max}^2}{\mu^2}$, which is larger than our burn-in sample size. Without the self-concordance condition, Frostig et al. (2015) also achieves a bound whose leading term is off by a prefactor $\alpha \in \left[1, \frac{\ell_{max}}{\mu}\right]$. In comparison, our algorithm is based on recursive variance reduction, and our theory does not require the self-concordance condition. Convergence measured by gradient norm squared, our nonasymptotic theoretical result achieves the optimal dependency on $\frac{\ell_{max}}{\mu}$.

Junchi — More remarks, future works, etc.

In this appendix, we do XXXX **Junchi — Indexing the appendix sections**

2.5 Proofs of auxiliary lemmas in §2.3.1

In this section, we prove the three auxiliary lemmas used in the proof of Theorem 2.2. Note that the proofs of the lemmas have inter-dependencies. In the following, we first prove Lemma 2.4 assuming Lemma 2.5, and then prove Lemma 2.5 assuming Lemma 2.6. Finally, we give a self-contained proof for Lemma 2.6.

2.5.1 Proof of Lemma 2.4

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

$$Q_1(t) := e_{t-1} - \eta\int_0^1 \nabla^2 f(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2};\xi_t)e_{t-1}d\gamma, \qquad Q_2(t) := (\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)),$$

$$Q_3(t) := \eta\int_0^1 \big(\nabla^2 f(\gamma\theta_{t-1} + (1-\gamma)\theta_{t-2};\xi_t) - \nabla^2 f(\theta^*;\xi_t)\big)\, y_{t-1}d\gamma.$$

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.

Junchi — Paragraphs to itemize?!

Upper bound on $\mathbb{E}\,\|Q_1(t)\|_2^2$:

Assumption 5 and Assumption 7 together imply that

$$
\begin{aligned}
&\mathbb{E}\,\|Q_1(t)\|_2^2 \\
&= \mathbb{E}\,\|e_{t-1}\|_2^2 - 2\eta\mathbb{E}\int_0^1 e_{t-1}^\top \nabla^2 F(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2})e_{t-1}d\gamma \\
&\quad + \eta^2\int_0^1 \mathbb{E}\left\|\nabla^2 f(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2};\xi_t)e_{t-1}\right\|_2^2 d\gamma \\
&= \mathbb{E}\,\|e_{t-1}\|_2^2 - \mathbb{E}\int_0^1 e_{t-1}^\top \left(2\eta\nabla^2 F(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2}) - \eta^2(\nabla^2 F(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2}))^2\right) e_{t-1}d\gamma \\
&\quad + \eta^2\int_0^1 \mathbb{E}\,\|\Xi_t(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2})e_{t-1}\|_2^2 d\gamma \\
&\overset{(i)}{\le} \mathbb{E}\,\|e_{t-1}\|_2^2 - (2\eta-\eta^2 L)\int_0^1 e_{t-1}^\top \nabla^2 F(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2})e_{t-1}d\gamma + \eta^2\ell_\Xi^2\int_0^1 \|e_{t-1}\|_2^2 d\gamma \\
&\overset{(ii)}{\le} \mathbb{E}\,\|e_{t-1}\|_2^2 - \mu\left(2\eta-\eta^2 L\right)\mathbb{E}\,\|e_{t-1}\|_2^2 + \ell_\Xi^2\eta^2\mathbb{E}\,\|e_{t-1}\|_2^2.
\end{aligned}
$$

In step (i), we are using the fact that $0 \preceq \nabla^2 F(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2}) \preceq LI_d$, 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 7 and Eq (2.16), 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 1.2.

Upper bound on $\mathbb{E}\,\|Q_3(t)\|_2^2$:

Applying Lemma 2.5 with $\tilde{\theta}_{t-1} := \gamma\theta_{t-1}+(1-\gamma)\theta_{t-2} \in \mathscr{F}_{t-1}$, we have:

$$
\begin{aligned}
&\mathbb{E}\,\|(H_t(\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2}) - H_t(\theta^*))\,y_{t-1}\|_2^2 \le c_1\sqrt{\mathbb{E}\,\|\gamma\theta_{t-1}+(1-\gamma)\theta_{t-2}-\theta^*\|_2^2} \\
&\le c_1\left(\sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta^*\|_2^2} + \sqrt{\mathbb{E}\,\|\theta_{t-2}-\theta^*\|_2^2}\right) \le c_1\sqrt{a_0}\left(\frac{1}{\sqrt{t-1}}+\frac{1}{\sqrt{t-2}}\right) \le \frac{16c_1\sqrt{a_0}}{\sqrt{t}}.
\end{aligned}
$$

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:

$$\sqrt{\mathbb{E}\left\|e_T\right\|_2^2} \leq (4c_1^{1/2}a_0^{1/4} + \ell_\Xi\sqrt{a_0}) \sum_{s=\mathscr{B}+1}^{t} s^{-\frac{1}{4}} \exp\left(-\frac{\mu\eta}{2}(T-s)\right) + e^{-\frac{\mu\eta(T-\mathscr{B})}{2}} \sqrt{\mathbb{E}\left\|e_{\mathscr{B}}\right\|_2^2}.$$

For the first term, we note that:

$$\begin{aligned}\sum_{s=\mathscr{B}+1}^{T} s^{-\frac{1}{4}} \exp\left(-\frac{\mu\eta}{2}(T-s)\right) &\leq \sum_{s=1}^{T/2} \exp\left(-\frac{\mu\eta}{2}T\right) + \frac{1}{(T/2)^{1/4}} \sum_{s=T/2}^{T} e^{-\frac{\mu\eta(T-s)}{2}} \\ &\leq \frac{T}{2}e^{-\frac{\mu\eta T}{2}} + \frac{4}{\mu\eta T^{1/4}}.\end{aligned}$$

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 $(a_0, c_1, a', \eta, \mu, \mathscr{B})$ but independent of t, such that

$$\mathbb{E}\left\|tv_t - y_t\right\|_2^2 \leq \frac{c_0}{\sqrt{t}},$$

which finishes the proof.

2.5.2 *Proof of Lemma 2.5*

Observe that Assumption (CLT.A) guarantees that

$$\mathbb{E}\left(\left\|(H_t(\theta^*) - H_t(\tilde{\theta}_{t-1}))y_{t-1}\right\|_2^2 | \mathscr{F}_{t-1}\right) \leq \beta^2 \left\|\tilde{\theta}_{t-1} - \theta^*\right\|_2^2 \cdot \left\|y_{t-1}\right\|_2^2$$

On the other hand, by Assumption 7, we have:

$$\mathbb{E}\left(\left\|(H_t(\theta^*) - H_t(\tilde{\theta}_{t-1}))y_{t-1}\right\|_2^2 | \mathscr{F}_{t-1}\right) \leq 4\ell_\Xi^2 \left\|y_{t-1}\right\|_2^2.$$

Taking a geometric average and applying the tower law yields the bound

$$\begin{aligned}\mathbb{E}\left\|\left(H_t(\tilde{\theta}_{t-1}) - H_t(\theta^*)\right)y_{t-1}\right\|_2^2 &\leq 2\ell_\Xi\beta\mathbb{E}\left(\left\|\tilde{\theta}_{t-1} - \theta^*\right\|_2 \cdot \left\|y_{t-1}\right\|_2^2\right) \\ &\overset{(i)}{\leq} 2\ell_\Xi\beta\sqrt{\mathbb{E}\left\|\tilde{\theta}_{t-1} - \theta^*\right\|_2^2} \cdot \sqrt{\mathbb{E}\left\|y_{t-1}\right\|_2^4},\end{aligned}$$

where step (i) follows from the Cauchy-Schwarz inequality. Applying Lemma 2.6, we are guaranteed the existence of a constant $a' > 0$ such that

$$\sup_{t \geq \mathscr{B}} \mathbb{E}\left\|y_t\right\|_2^4 \leq a' < \infty.$$

Setting $c_1 = 2\ell_\Xi\beta\sqrt{a'}$ completes the proof of the claim.

2.5.3 Proof of Lemma 2.6

Throughout this section, we adopt the shorthand notation $H_t := H_t(\theta^*)$ and $\Xi_t := \Xi_t(\theta^*)$. Beginning with the proof of the first claim, we take expectations on both sides of Eq. (2.15), thereby finding that

$$\mathbb{E}(y_t) = \mathbb{E}\left(y_{t-1} - \eta H_t(\theta^*) y_{t-1} + \varepsilon_t(\theta^*)\right) = (I - \eta H^*)\mathbb{E}(y_{t-1}) = (I - \eta H^*)^{t-\mathscr{B}}\mathbb{E}(y_{\mathscr{B}}) = 0.$$

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:

$$\begin{aligned}
&\mathbb{E}\left(\|(I - \eta H_t) y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) \\
&\leq \|(I - \eta H^*) y_{t-1}\|_2^4 + 4\mathbb{E}\left(\langle \eta \Xi_t y_{t-1}, (I - \eta H^*) y_{t-1}\rangle \|(I - \eta H^*) y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\|\eta \Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) \\
&\qquad + 6\mathbb{E}\left(\|(I - \eta H^*) y_{t-1}\|_2^2 \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) + 4\mathbb{E}\left(\langle \eta \Xi_t y_{t-1}, (I - \eta H^*) y_{t-1}\rangle \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) \\
&\leq \|(I - \eta H^*) y_{t-1}\|_2^4 + \eta^4 \mathbb{E}\left(\|\Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) + 6\eta^2 \ell_\Xi^2 \|y_{t-1}\|_2^4 \\
&\qquad + 2\mathbb{E}\left(\|\eta \Xi_t y_{t-1}\|_2^4 \mid \mathscr{F}_{t-1}\right) + 2\mathbb{E}\left(\|(I - \eta H^*) y_{t-1}\|_2^2 \cdot \|\eta \Xi_t y_{t-1}\|_2^2 \mid \mathscr{F}_{t-1}\right) \\
&\leq (1 - 3\eta\mu)\|y_{t-1}\|_2^4 + 8\eta^2 \ell_\Xi^2 \|y_{t-1}\|_2^4 + 3\eta^4 \ell_\Xi'^4 \|y_{t-1}\|_2^4.
\end{aligned}$$

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 $\mathbb{E}\|y_{\mathscr{B}}\|_2^4 = 0$. Solving this recursion leads to the bound

$$\sup_{t \geq \mathscr{B}} \mathbb{E}\|y_t\|_2^4 \leq \frac{157}{(\mu\eta)^4}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4.$$

Let $a' = \frac{157}{(\eta\mu)^4}$, we prove the second claim.

Finally we study the stationary covariance of the process $\{y_t\}_{t\geq\mathscr{B}}$. The existence and uniqueness of the stationary distribution was established in Mou et al. (2020). Let π_η denote the stationary distribution of $(y_t)_{t\geq\mathscr{B}}$, 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.

2.6 Proofs of auxiliary lemmas in §2.3.2 and §2.3.3

2.6.1 Proof of Lemma 2.7

Proof (Proof of Lemma 2.7). Recalling 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

$$\begin{aligned}
\mathbb{E}\|tz_t\|_2^4 &= \mathbb{E}\|(t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&= \mathbb{E}\|(t-1)z_{t-1}\|_2^4 + \mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&\quad + 4\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^3\|(t-1)z_{t-1}\|_2 \\
&\quad + 6\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^2\|(t-1)z_{t-1}\|_2^2
\end{aligned} \tag{2.25}$$

By Hölder's inequality and Young's inequality, we bound the third term and the fourth term of the RHS as

$$
\begin{aligned}
&\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^3\left\|(t-1)z_{t-1}\right\|_2\\
&\le\left(\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{3/4}\left(\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4\right)^{1/4}\\
&\le\frac{1}{2}\left(\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4\right)^{1/2}\\
&\quad+\frac{1}{2}\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^2\left\|(t-1)z_{t-1}\right\|_2^2\\
&\le\left(\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4\right)
\end{aligned}
$$

Thus Eq. (2.25) can be rewritten as

$$
\begin{aligned}
\mathbb{E}\left\|tz_t\right\|_2^4&\le\mathbb{E}\left\|(t-1)z_{t-1}+(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\
&=\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4+3\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\
&\quad+8\left(\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4\right)\\
&\le\left(\sqrt{\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4}+4\sqrt{\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4}\right)^2
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\
&\le 27(t-1)^4\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4+27\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\right\|_2^4+27\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^4\\
&\le 27\tilde{\ell}_\Xi^4\eta^4(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4+\frac{27\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+27\tilde{\sigma}_*^4
\end{aligned}
$$

Then

$$
t^2\sqrt{\mathbb{E}\left\|z_t\right\|_2^4}\le\sqrt{\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4}+12\sqrt{3}\tilde{\ell}_\Xi^2\eta^2\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+\frac{12\sqrt{3}\tilde{\ell}_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+12\sqrt{3}\tilde{\sigma}_*^2
$$

Combining this with Eq. (2.32) in Lemma 2.9 that

$$
\sqrt{\mathbb{E}\left\|tv_t\right\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+\frac{5}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+14\tilde{\sigma}_*^2
$$

By the choice of η satisfying $\eta\le\frac{1}{56L}\wedge\frac{\mu}{64\tilde{\ell}_\Xi^2}$, we have $\frac{\tilde{\ell}_\Xi^2}{\mu^2}\le\frac{1}{64\eta\mu}$ and

$$
t^2\sqrt{\mathbb{E}\left\|z_t\right\|_2^4}+t^2\sqrt{\mathbb{E}\left\|v_t\right\|_2^4}\le\sqrt{\mathbb{E}\left\|(t-1)z_{t-1}\right\|_2^4}+\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+\frac{6}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+35\tilde{\sigma}_*^2
$$

Recursively applying the above inequality and by observing that $\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4} \leq 2\sqrt{\mathbb{E}\,\|z_t\|_2^4} + 2\sqrt{\mathbb{E}\,\|v_t\|_2^4}$, we have

$$\begin{aligned} T^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^4} &\leq 2T^2\sqrt{\mathbb{E}\,\|z_T\|_2^4} + 2T^2\sqrt{\mathbb{E}\,\|v_T\|_2^4} \\ &\leq 2\sqrt{\mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^4} + 2\sqrt{\mathbb{E}\,\|\mathscr{B}v_{\mathscr{B}}\|_2^4} + \frac{12}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4} + 70(T-\mathscr{B})\tilde{\sigma}_*^2 \end{aligned} \tag{2.26}$$

Further for $\mathscr{B}z_{\mathscr{B}}$ and $\mathscr{B}v_{\mathscr{B}}$ we note that by applying Khintchine's inequality as well as Young's inequality we have:

$$\mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^4 = \mathbb{E}\left\|\sum_{t=1}^{\mathscr{B}}\varepsilon_t(\theta_0)\right\|_2^4 = \mathbb{E}\left(\sum_{t=1}^{\mathscr{B}}\|\varepsilon_t(\theta_0)\|_2^2\right)^2 \leq \mathscr{B}\mathbb{E}\sum_{t=1}^{\mathscr{B}}\|\varepsilon_t(\theta_0)\|_2^4 \leq 8\mathscr{B}^2\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4 + \tilde{\sigma}_*^4\right) \tag{2.27}$$

and

$$\begin{aligned} \mathbb{E}\,\|\mathscr{B}v_{\mathscr{B}}\|_2^4 &= \mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^4 + \mathbb{E}\,\|\mathscr{B}\nabla F(\theta_0)\|_2^4 + 4\mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^3\|\mathscr{B}\nabla F(\theta_0)\|_2 + 6\mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^2\|\mathscr{B}\nabla F(\theta_0)\|_2^2 \\ &\leq 7\mathbb{E}\,\|\mathscr{B}v_{\mathscr{B}}\|_2^4 + 5\mathbb{E}\,\|\mathscr{B}\nabla F(\theta_0)\|_2^4 \leq 56\mathscr{B}^2\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4 + \tilde{\sigma}_*^4\right) + 5\mathscr{B}^4\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4 \end{aligned} \tag{2.28}$$

Taking squared root on Eq. (2.27) and (2.28) and recalling that $\eta \leq \frac{\mu}{64\tilde{\ell}_\Xi^2}$, we have

$$\sqrt{\mathbb{E}\,\|\mathscr{B}z_{\mathscr{B}}\|_2^4} \leq 2\sqrt{2}\mathscr{B}\left(\frac{\tilde{\ell}_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4} + \tilde{\sigma}_*^2\right), \tag{2.29}$$

$$\sqrt{\mathbb{E}\,\|\mathscr{B}v_{\mathscr{B}}\|_2^4} \leq (\sqrt{5}+1/8)\mathscr{B}^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4} + 8\mathscr{B}\tilde{\sigma}_*^2. \tag{2.30}$$

Bringing Eq. (2.29) and (2.30) into Eq. (2.26), we arrive at the following:

$$\begin{aligned} T^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_{T-1})\|_2^4} &\leq 4\sqrt{2}\mathscr{B}\left(\frac{\tilde{\ell}_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4} + \tilde{\sigma}_*^2\right) + (2\sqrt{5}+1/4)\mathscr{B}^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4} \\ &\quad + 16\mathscr{B}\tilde{\sigma}_*^2 + \frac{12}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4} + 70(T-\mathscr{B})\tilde{\sigma}_*^2 \\ &\leq 5\mathscr{B}^2\sqrt{\mathbb{E}\,\|\nabla F(\theta_0)\|_2^4} + \frac{12}{\eta\mu}\sum_{t=\mathscr{B}+1}^{T}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4} + 70T\tilde{\sigma}_*^2 \end{aligned} \tag{2.31}$$

Dividing both sides by T^2 and summing up Eq. (2.31) from $T = \mathscr{B}+1$ to $T^* \geq \mathscr{B}+1$ we have $\eta \leq \frac{\mu}{64\tilde{\ell}_\Xi^2}, \mathscr{B} \geq 2$

$$\sum_{T=\mathscr{B}+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4}\le 5\mathscr{B}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+\frac{12}{\eta\mu\mathscr{B}}\sum_{t=\mathscr{B}+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+70\tilde{\sigma}_*^2\log\left(\frac{T^*}{\mathscr{B}}\right)$$

Taking $\mathscr{B}=\left\lceil\frac{24}{\eta\mu}\right\rceil$, we have

$$\sum_{T=\mathscr{B}+1}^{T^*}\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4}\le 10\mathscr{B}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+140\tilde{\sigma}_*^2\log\left(\frac{T^*}{\mathscr{B}}\right)$$

Again by Eq. (2.31), we have

$$\begin{aligned}
&T^2\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4}\\
&\le 5\mathscr{B}^2\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+\frac{12}{\eta\mu}\left(10\mathscr{B}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+140\tilde{\sigma}_*^2\log\left(\frac{T}{\mathscr{B}}\right)\right)+70T\tilde{\sigma}_*^2\\
&\le 10\mathscr{B}^2\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+70\mathscr{B}\tilde{\sigma}_*^2\log\left(\frac{T}{\mathscr{B}}\right)+70T\tilde{\sigma}_*^2
\end{aligned}$$

Dividing both sides by T^2 we conclude that

$$\begin{aligned}
\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^4}&\le\frac{10\mathscr{B}^2}{T^2}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+70\left(1+\frac{\mathscr{B}}{T}\log\left(\frac{T}{\mathscr{B}}\right)\right)\frac{\tilde{\sigma}_*^2}{T}\\
&\le\frac{10\mathscr{B}^2}{T^2}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+\frac{140\tilde{\sigma}_*^2}{T}
\end{aligned}$$

which finishies our proof of Lemma 2.7.

2.6.2 *Proof of Lemma 2.8*

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

Lemma 2.9 (v_t recursion). *Under the setting of Theorem 2.3, when $\eta\le\frac{1}{56L}\wedge\frac{\mu}{64\ell_\Xi^2}$, we have the following bound for $t\ge\mathscr{B}+1$*

$$\sqrt{\mathbb{E}\|tv_t\|_2^4}\le\left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4}+\frac{5}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}+14\tilde{\sigma}_*^2 \tag{2.32}$$

The detailed proof is relegated to §2.6.3.1. We are ready for the proof.

Following the exact same argument as in the proof of Lemma **??Junchi — Fix this ref**, we derive the following lemma

Lemma 2.10.

$$\sqrt{\mathbb{E}\|v_T\|_2^4} \le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}} \frac{\mathscr{B}^4}{T^4}\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \frac{187500\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{2604\tilde{\sigma}_*^2}{\eta\mu T^2}$$

Indeed, from (2.19) and (2.32)

$$\begin{aligned} t^2\sqrt{\mathbb{E}\|v_t\|_2^4} &\le \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\|v_{t-1}\|_2^4} + \frac{5}{\eta\mu}\left[\frac{6250\,\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2} + \frac{140\,\tilde{\sigma}_*}{t}\right] + 14\tilde{\sigma}_*^2 \\ &\le \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\|v_{t-1}\|_2^4} + \frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3t^2} + 714\tilde{\sigma}_*^2 \end{aligned} \tag{2.33}$$

We have from (2.33)

$$\begin{aligned} t^4\sqrt{\mathbb{E}\|v_t\|_2^4} &\le \left(1-\frac{\eta\mu}{2}\right)t^2(t-1)^2\sqrt{\mathbb{E}\|v_{t-1}\|_2^4} + \frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + 714\tilde{\sigma}_*^2t^2 \\ &\le \left(1-\frac{\eta\mu}{6}\right)(t-1)^4\sqrt{\mathbb{E}\|v_{t-1}\|_2^4} + \frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + 714\tilde{\sigma}_*^2t^2 \end{aligned}$$

since the following holds $\frac{t^2}{(t-1)^2} \le \frac{1-\frac{\eta\mu}{6}}{(1-\frac{\eta\mu}{6})^3} \le \frac{1-\frac{\eta\mu}{6}}{1-\frac{\eta\mu}{2}}$ This gives, by solving the recursion,

$$\begin{aligned} T^4\sqrt{\mathbb{E}\|v_T\|_2^4} &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\left[\frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + 714\tilde{\sigma}_*^2t^2\right] \\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + \sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}714\tilde{\sigma}_*^2t^2 \\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \frac{6}{\eta\mu}\cdot\frac{31250\,\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + \frac{6}{\eta\mu}T^2\cdot 714\tilde{\sigma}_*^2 \\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\mathscr{B}^4\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \frac{187500\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4} + \frac{(714)(6)\tilde{\sigma}_*^2}{\eta\mu}T^2 \end{aligned} \tag{2.34}$$

where the summand is increasing so

$$\sum_{t=\mathscr{B}+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}t^2 \le \frac{6}{\eta\mu}T^2$$

All in all, this concludes

$$\begin{aligned} \sqrt{\mathbb{E}\|v_T\|_2^4} &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\frac{\mathscr{B}^4}{T^4}\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \frac{187500\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{(714)(6)\tilde{\sigma}_*^2}{\eta\mu T^2} \\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-\mathscr{B}}\frac{\mathscr{B}^4}{T^4}\sqrt{\mathbb{E}\|v_{\mathscr{B}}\|_2^4} + \frac{187500\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4} + \frac{(714)(6)\tilde{\sigma}_*^2}{\eta\mu T^2} \end{aligned}$$

and hence Lemma 2.10.

Bringing the burn-in upper bounds 2.30 and we arrive at our final result for bounding $\sqrt{\mathbb{E}\|v_T\|_2^4}$:

$$\begin{aligned}\sqrt{\mathbb{E}\|v_T\|_2^4} &\leq \left(\frac{3\mathscr{B}^4}{T^4}\sqrt{\mathbb{E}\|\nabla F(\theta_0)\|_2^4}+\frac{8\mathscr{B}^3}{T^4}\tilde{\sigma}_*^2\right)+\frac{187500\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\tilde{\sigma}_*^2}{\eta\mu T^2}\\ &\leq \frac{1359375\,\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}+\frac{4484\tilde{\sigma}_*^2}{\eta\mu T^2}\end{aligned}$$

2.6.3 Proofs of secondary lemmas

2.6.3.1 Proof of Lemma 2.9

Proof (Proof of Lemma 2.9). 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 forth moments on both sides, we have:

$$\begin{aligned}&\mathbb{E}\|tv_t-\nabla F(\theta_{t-1})\|_2^4\\ &=\mathbb{E}\|(t-1)v_{t-1}+(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\|_2^4\\ &=(t-1)^4\mathbb{E}\|v_{t-1}\|_2^4+\underbrace{4\mathbb{E}\left[\|(t-1)v_{t-1}\|_2^2\langle(t-1)v_{t-1},(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\delta_t(\theta_{t-1})\rangle\right]}_{\text{II}}\\ &\quad+\underbrace{6\mathbb{E}\left[\|(t-1)v_{t-1}\|_2^2\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\delta_t(\theta_{t-1})\|_2^2\right]}_{\text{I}}\\ &\quad+\underbrace{4\mathbb{E}\left[\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\delta_t(\theta_{t-1})\|_2^3\|(t-1)v_{t-1}\|_2\right]}_{\text{III}}\\ &\quad+\mathbb{E}\left[\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\delta_t(\theta_{t-1})\|_2^4\right]\end{aligned}\tag{2.35}$$

For bounding I, we apply the Hölder's inequality and have

$$\text{I}\leq 6\left(\mathbb{E}\|(t-1)v_{t-1}\|_2^4\right)^{1/2}\left(\mathbb{E}\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\delta_t(\theta_{t-1})\|_2^4\right)^{1/2}\tag{2.36}$$

For Bounding III, we again apply the Hölder's inequality:

$$\begin{aligned}\mathrm{III} &\le 4\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{3/4}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/4}\\ &\le 2\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2}\\ &\quad+2\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\end{aligned} \tag{2.37}$$

For bounding II, we first take expectation with respect to ξ_t and have

$$\mathrm{II}=4\mathbb{E}\left[\left\|(t-1)v_{t-1}\right\|_2^2\left\langle(t-1)v_{t-1},(t-1)\left(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right)\right\rangle\right]$$

where

$$\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\rangle\le-\frac{1}{\eta L}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2$$

and

$$\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\rangle\le-\frac{\mu}{\eta}\left\|\theta_{t-1}-\theta_{t-2}\right\|_2^2$$

holds true for any μ-strongly convex and L-smooth F. Then we have

$$\begin{aligned}\mathrm{II}&\le-(t-1)^4\mathbb{E}\left[\left\|v_{t-1}\right\|_2^2\left(\frac{1}{\eta L}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2+3\eta\mu\left\|v_{t-1}\right\|_2^2\right)\right]\\ &=-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\mathbb{E}\left\|v_{t-1}\right\|_2^2\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^2\\ &\le-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}\end{aligned} \tag{2.38}$$

Combining Eq. (2.36), (2.38) and (2.37) into Eq. (2.35) we have

$$\begin{aligned}&\mathbb{E}\left\|tv_t-\nabla F(\theta_{t-1})\right\|_2^4\\ &\le(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4+3\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\ &\quad+8\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2}\\ &\quad-3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4-\frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}\end{aligned} \tag{2.39}$$

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:

$$\begin{aligned}&\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t)-\nabla f(\theta_{t-2};\xi_t))+\varepsilon_t(\theta_{t-1})\right\|_2^4\\ &\le\mathbb{E}\left\|(t-1)(\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}))+(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)+\varepsilon_t(\theta^*)\right\|_2^4\\ &\le8(t-1)^4\underbrace{\mathbb{E}\left\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2})\right\|_2^4}_{I_1}+8\underbrace{\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)+\varepsilon_t(\theta^*)\right\|_2^4}_{I_2}\end{aligned} \tag{2.40}$$

We claim that

$$I_1 \leq 5\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 7\tilde{\ell}_\Xi^4 \eta^4 \mathbb{E}\left\|v_{t-1}\right\|_2^4 \tag{2.41}$$

and

$$I_2 \leq 8\mathbb{E}\left\|\delta_t(\theta_{t-1}) - \delta_t(\theta^*)\right\|_2^4 + 8\mathbb{E}\left\|\delta_t(\theta^*)\right\|_2^4 \leq \frac{8\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 8\tilde{\sigma}_*^4 \tag{2.42}$$

Combining Eq. (2.40), (2.41) and (2.42) we have the bound

$$\begin{aligned}
&\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4 \\
\leq\ & 40(t-1)^4\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 56\tilde{\ell}_\Xi^4\eta^4(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + \frac{64\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 64\tilde{\sigma}_*^4
\end{aligned} \tag{2.43}$$

Then, we bring Eq. (2.43) into Eq. (2.39) and have

$$\begin{aligned}
&\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4 \\
\leq\ & (t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + 120(t-1)^4\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 168\tilde{\ell}_\Xi^4\eta^4(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 \\
&+ \frac{192\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\tilde{\sigma}_*^4 + 8\sqrt{40}(t-1)^4\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} \\
&+ 64\tilde{\ell}_\Xi^2\eta^2(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + 64\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2} \\
&- 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} \\
\leq\ & \left(1 - 3\eta\mu + 64\tilde{\ell}_\Xi^2\eta^2 + 168\tilde{\ell}_\Xi^4\eta^4\right)\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4 \\
&+ \left(8\sqrt{40} - \frac{1}{\eta L} + 120L^2\eta^2\right)(t-1)^4\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} \\
&+ 64\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2} + \frac{192\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\tilde{\sigma}_*^4 \\
\leq\ & (1-\eta\mu)^2\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4 + 64\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2} \\
&+ \frac{192\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\tilde{\sigma}_*^4
\end{aligned}$$

where the last inequality is due to the choice of $\eta \leq \frac{\mu}{64\tilde{\ell}_\Xi^2}$ and $\eta \leq \frac{1}{56L}$ such that

$$168\tilde{\ell}_\Xi^4\eta^4 \leq \eta^2\mu^2, \tilde{\ell}_\Xi^2\eta^2 \leq \eta\mu \qquad \text{and} \quad 8\sqrt{40} - \frac{1}{\eta L} + 120L^2\eta^2 \leq 0$$

Taking squared root on both sides, we have

$$\sqrt{\mathbb{E}\left\|tv_t-\nabla F(\theta_{t-1})\right\|_2^4}\leq(1-\eta\mu)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4}+32\left(\frac{\tilde{\ell}_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+\tilde{\sigma}_*^2\right) \tag{2.44}$$

Furthermore, Young's inequality gives (a different coefficient from the Cauchy+Young analysis in §18 in Li et al. (2020) is adopted)

$$\begin{aligned}
&\mathbb{E}\left\|tv_t-\nabla F(\theta_{t-1})\right\|_2^4\\
&=t^4\mathbb{E}\left\|v_t\right\|_2^4+\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+6\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2-4\mathbb{E}\left\|tv_t\right\|_2^3\left\|\nabla F(\theta_{t-1})\right\|_2-4\mathbb{E}\left\|tv_t\right\|_2\left\|\nabla F(\theta_{t-1})\right\|_2^3\\
&\geq t^4\mathbb{E}\left\|v_t\right\|_2^4+\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+6\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2-2\mathbb{E}\left[\frac{2}{\eta\mu}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2+\frac{\eta\mu}{2}\left\|tv_t\right\|_2^4\right]\\
&\quad-2\mathbb{E}\left[\frac{\eta\mu}{2}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2+\frac{2}{\eta\mu}\left\|\nabla F(\theta_{t-1})\right\|_2^4\right]\\
&\geq(1-\eta\mu)\,\mathbb{E}\left\|tv_t\right\|_2^4+\left(1-\frac{4}{\eta\mu}\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+\left(6-\frac{4}{\eta\mu}-\eta\mu\right)\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2\\
&\geq(1-\eta\mu)\,\mathbb{E}\left\|tv_t\right\|_2^4-\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4-(1-\eta\mu)\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2
\end{aligned}$$

Combining this we have

$$\begin{aligned}
&(1-\eta\mu)\,\mathbb{E}\left\|tv_t\right\|_2^4-\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4-(1-\eta\mu)\left(\frac{4}{\eta\mu}-1\right)\mathbb{E}\left\|tv_t\right\|_2^2\left\|\nabla F(\theta_{t-1})\right\|_2^2\\
&\leq(1-\eta\mu)^2\,\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4+64\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+\tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\\
&\quad+\frac{192\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4+192\tilde{\sigma}_*^4
\end{aligned}$$

Now we multiply both sides by $(1-\eta\mu)^{-1}$, noting that $(1-\eta\mu)^{-1}\leq(1-\eta L)^{-1}\leq\frac{56}{55}$, rearranging, and have

$$\begin{aligned}
\mathbb{E}\|tv_t\|_2^4 &\leq (1-\eta\mu)\,\mathbb{E}\|(t-1)v_{t-1}\|_2^4 + \frac{(56)(64)}{(55)}\left(\frac{\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2} \\
&\quad + \frac{\frac{(56)(192)}{(55)}\tilde{\ell}_\Xi^4}{\mu^4}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \frac{(56)(192)}{(55)}\tilde{\sigma}_*^4 \\
&\quad + \frac{\frac{(4)(56)}{(55)}}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + \frac{4}{\eta\mu}\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2 \\
&\leq \left(1-\frac{\eta\mu}{2}\right)^2\mathbb{E}\|(t-1)v_{t-1}\|_2^4 + \frac{7}{55\eta^2\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + (56)(192)/(55)\tilde{\sigma}_*^4 \\
&\quad + \frac{56}{55}\left(\frac{1}{64\eta^2\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4 + 64\tilde{\sigma}_*^4\right)^{1/2}\left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{1/2} + \frac{4}{\eta\mu}\mathbb{E}\|tv_t\|_2^2\|\nabla F(\theta_{t-1})\|_2^2.
\end{aligned}$$

Rearranging and taking squared root on both sides we conclude that

$$\sqrt{\mathbb{E}\|tv_t\|_2^4} - \frac{2}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} \leq \left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4} + \frac{3}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 14\tilde{\sigma}_*^2$$

Further rearranging, we have

$$\sqrt{\mathbb{E}\|tv_t\|_2^4} \leq \left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\|(t-1)v_{t-1}\|_2^4} + \frac{5}{\eta\mu}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4} + 14\tilde{\sigma}_*^2$$

which concludes our proof.

Proof of Eq. (2.41): We use similar decomposition as in the decomposition in Eq. (2.35) and have

$$\begin{aligned}
I_1 &= \mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + \mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^4 \\
&\quad + 4\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^3\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2 \\
&\quad + 6\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2
\end{aligned}$$

Where we note that we used the fact that one of the cross terms in the fourth moment decomposition $\mathbb{E}\left[\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2\langle\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}), \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\rangle\right] = 0$.

Further utilizing the Hölder's inequality, we have

$$
\begin{aligned}
I_1 &\leq \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + \mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^4 \\
&\quad + 4\left(\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^4\right)^{3/4} \left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2\right)^{1/4} \\
&\quad + 6\left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2} \left(\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2\right)^{1/2} \\
&\leq \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 3\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^4 \\
&\quad + 8\left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2} \left(\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^4\right)^{1/2} \\
&\leq \mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 3\tilde{\ell}_\Xi^4 \eta^4 \mathbb{E}\,\|v_{t-1}\|_2^4 \\
&\quad + 8\tilde{\ell}_\Xi^2 \eta^2 \left(\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4\right)^{1/2} \left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \\
&\leq 5\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 7\tilde{\ell}_\Xi^4 \eta^4 \mathbb{E}\,\|v_{t-1}\|_2^4
\end{aligned}
$$

Thus proving Eq. (2.41).

Chapter 3
ROOT-SGD with Diminishing Stepsize

In this chapter, we revisit ROOT-SGD, an innovative method for stochastic optimization to bridge the gap between stochastic optimization and statistical efficiency. The proposed method enhances the performance and reliability of ROOT-SGD by integrating a carefully designed *diminishing stepsize strategy*. This approach addresses key challenges in optimization, providing robust theoretical guarantees and practical benefits. Our analysis demonstrates that ROOT-SGD with diminishing stepsize achieves optimal convergence rates while maintaining computational efficiency. By dynamically adjusting the learning rate, ROOT-SGD ensures improved stability and precision throughout the optimization process. The findings of this study offer valuable insights for developing advanced optimization algorithms that are both efficient and statistically robust.
Key words: Stochastic Optimization, ROOT-SGD Algorithm, Statistical Efficiency, Diminishing Stepsize, Non-Asymptotic Bounds

3.1 Introduction

Stochastic optimization has become a cornerstone in machine learning and statistical learning, particularly for large-scale and high-dimensional data. Among the various stochastic optimization techniques, *stochastic gradient descent* (SGD) stands out due to its simplicity and effectiveness Robbins et al. (1951). However, the performance of SGD can be significantly influenced by the stepsize schedule, which determines the balance between convergence speed and stability. In special, the *diminishing stepsize* strategy has been proposed to address the limitations of fixed stepsize schemes, offering a way to enhance the efficiency and robustness of SGD. This strategy allows for adaptive learning rates that decrease over time, facilitating 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 **?**, a novel optimization framework that improves both the convergence and stability of stochastic gradient methods. ROOT-SGD is designed to be theoretically optimal and practically effective, providing a comprehensive solution to the inherent trade-offs in stochastic optimization. In the mean time, the estimator produced by the ROOT-SGD algorithm share the same *optimal statistical properties* typically possessed by the empirical risk minimizer. The notion of statistical efficiency, in both asymptotic and non-asymptotic forms, allows for assessment of optimality. The (Bayesian) Cramér-Rao lower bounds relate the fundamental limit of the *mean-squared error* (MSE) of an estimator to the Fisher information;[1] moreover, local asymptotic minimax theorems further show that the optimal asymptotic distribution, under any bowl-shaped loss function, takes a Gaussian form **?**Duchi & Ruan (2021). The asymptotic covariance provides a form of local complexity, and it is desirable to achieve this optimal bound with a *unity* pre-factor. Under relatively mild conditions, the empirical risk minimizer itself does so.

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.[2] 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.

In particular, 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

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{for a function of the form } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big] \tag{3.1}$$

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:

$$\widehat{\theta}_N^{\text{ERM}} := \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^N f(\theta, \xi_i) \quad \text{satisfies} \quad \sqrt{N}\left(\widehat{\theta}_N^{\text{ERM}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right) \tag{3.2}$$

[1] The vanilla Cramér-Rao lower bounds are valid only for unbiased estimator; the Bayesian Cramér-Rao lower bound, on the other hand, gives lower bound on the Bayes risk for *any* estimator **?**

[2] To motivate the readers on the importance of unity pre-factor, consider the following *thought experiment*: an algorithm that randomly discards half of the training data is undesirable in practice, but this cannot be captured by any performance metric that ignores constant multiplicative factors.

Furthermore, the asymptotic distribution (3.2) is known to be *locally optimal*—see **?** and Duchi & Ruan (2021) 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* (3.2)*?*

An affirmative answer to this question at least qualitatively, is provided by the seminal work by Polyak & Juditsky (1992); Polyak (1990); Ruppert (1988). 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 (3.2), as the number of samples grows to infinity. This algorithm lays the foundations of online statistical inference Chen et al. (2020); **?** and fine-grained error guarantees for stochastic optimization algorithms **?**Dieuleveut et al. (2020). 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., Polyak & Juditsky (1992); Duchi & Ruan (2021)). However, the asymptotic guarantee (3.2) 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., **?**). 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 Le Cam (1970); **?**. *Is there a (single-loop) stochastic optimization algorithm that achieves the asymptotic guarantee* (3.2) *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 (3.2) with additional terms that decays faster as $N \to +\infty$. For example, under the one-point Hessian Lipschitz condition, **??**Gadat & Panloup (2017) established bounds in the form of

$$\mathbb{E}\left\|\widehat{\theta}_N^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \leq \frac{1}{N}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + \text{high order terms} \tag{3.3}$$

  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 3.4 for details). For quadratic objectives, the ar-

gument of **?** allows one to achieve an $O(N^{-3/2})$ higher-order term

$$\mathbb{E}\left\|\widehat{\theta}_N^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \leq \frac{1}{N}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + O\left(\frac{1}{N^{3/2}}\right) \tag{3.4}$$

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. *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 **?**.

The ROOT-SGD algorithm with varying stepsizes

For the stochastic optimization problem in the strongly-convex and mean-squared smooth setup, **?** 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

$$v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right) \tag{3.5a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t \tag{3.5b}$$

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

For the purposes of stabilizing the iterates, Algorithm (3.5) is initialized with a *burn-in* phase of length $\mathscr{B} > 1$, 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 $t = 1, \ldots, \mathscr{B}$, and compute

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

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

? 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 lower-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 **?**, 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 (3.5) 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.

3.1.1 Contribution and organization

Let us summarize the contributions of this chapter:

- On the asymptotic side, we show in Theorem 3.1 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 3.2 serving as complementary to the asymptotic Theorem 3.1.

- Under the same set of assumptions, in Theorem 3.3, 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 (arbi-

trarily 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 3.4, we show an upper bound on the mean-squared error (MSE) in the form of (3.3). 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 3.5.

This chapter is organized as follows. §3.2 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. §3.3 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 §3.4. In §3.5, we prove the asymptotic results, establishing optimality of ROOT-SGD and sub-optimality of Polyak-Ruppert averaging without high-order smoothness conditions. Additional related works are discussed in §3.6. We finalize the chapter with some discussions in §3.7.

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

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

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

3.2 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

$$\varepsilon(\theta;\xi) = \nabla_\theta f(\theta;\xi) - \nabla F(\theta) \tag{3.6}$$

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

Assumption 13 *The noise function $\theta \mapsto \nabla_\theta f(\theta,\xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\,\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\|_2^2 \le \ell_\Xi^2\,\|\theta_1 - \theta_2\|_2^2 \qquad \textit{for all pairs } \theta_1,\theta_2 \in \mathbb{R}^d \tag{3.7}$$

Assumption 14 *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 15 *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 13 (sometimes referred to as *mean-squared-smoothness*) as well as Assumptions 14 and 15 are standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see, e.g., **?**, 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 12 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 13, 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 **?**, which requires weaker conditions than the *individual smooth and convex* (ISC) setting in their chapter.

3.2.1 Asymptotic normality

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

Theorem 3.1. *Under Assumptions 12, 13 and 14, there exists universal constants $c, c_1 > 0$, such that for any $\alpha \in (0,1)$, ROOT-SGD with burn-in time $\mathscr{B} = c(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2})$ and stepsize sequence $\eta_t = \frac{1}{\mu \mathscr{B}^{1-\alpha} t^\alpha}$ for $t \geq \mathscr{B}$ satisfies the asymptotic limit:*

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

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

See §3.5.1 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 $(\theta_t, v_t, z_t)_{t \geq \mathscr{B}}$, where a central object in our analysis is the *tracking error process*:

$$z_t := v_t - \nabla F(\theta_{t-1}) \qquad \text{for } t \geq \mathscr{B} \tag{3.8}$$

See Proposition 3.1 for details.

A few remarks are in order. First, we note that this limiting distribution is locally asymptotically optimal (see, e.g., Duchi & Ruan (2021)). This result for diminishing stepsize sequence is complementary to the constant-stepsize result in the chapter **?**, 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 3.4 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 Polyak & Juditsky (1992); Ruppert (1988), Theorem 3.1 requires no quantitative Lipschitz or Hölder assumptions on the Hessian matrix ∇F, while requiring a stochastic continuity condition (Assumption 13) 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.

[3] In the meantime, the asymptotic normality result for multi-loop ROOT-SGD in **?** 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 prior knowledge of N.

3.2.2 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 3.1, this exhibits an asymptotic separation between Polyak-Ruppert averaging and ROOT-SGD.

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

$$\theta_t = \theta_{t-1} - \eta_t \nabla f(\theta, \xi) \qquad \text{for } t = 1, 2, \ldots \tag{3.9a}$$

$$\bar{\theta}_T = \frac{1}{T - \mathscr{B}} \sum_{t=\mathscr{B}}^{T-1} \theta_t \tag{3.9b}$$

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 (3.9a) a time-homogeneous Markov process, which is amendable to our analysis. Such a simplification has been employed in existing literature Bach (2014); Dieuleveut et al. (2020), 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 (3.9), even if started from the optimum, for any choice of burn-in period and step size.

Theorem 3.2. *There exists a function $F : \mathbb{R} \to \mathbb{R}$ that satisfies Assumptions 12 and 15 with constants $(\mu = 1, L = 2)$ and noise model $f(\cdot, \xi)$ satisfying Assumptions 13 and 14 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 (3.9) starting from $\theta_0 = \theta^*$, with step size $\eta = \eta_0 T^{-\alpha}$ and burn-in time $\mathscr{B} = S_0 T^\beta$ leads to the following limit:*

$$\lim_{T \to +\infty} T \cdot \mathbb{E} \left\| \bar{\theta}_T - \theta^* \right\|_2^2 = +\infty \tag{3.10}$$

See §3.5.2 for the proof of this theorem.

Note that Theorem 3.2 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 **?**, 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 3.1 but also the non-asymptotic bounds on the gradient norm in Theorem 3.3 works. Moreover, note that Polyak & Juditsky (1992) 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 3.2 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.

3.3 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 §3.2, while the latter requires additional conditions, as with existing literature Gadat & Panloup (2017); **?**.

3.3.1 Upper bounds on the gradient norm

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 3.1. *Under Assumptions 12, 13, and 14, there exist universal constants $c_1, c_2, C > 0$, using burn-in time $\mathscr{B} \geq C(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu})$, 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 $t > \mathscr{B}$. We have the following bounds for any $T \geq 2\mathscr{B}\log\mathscr{B}$:*

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

See §3.4.1 for the proof of this claim.

By the decomposition $\nabla F(\theta_t) = v_{t+1} - z_{t+1}$, it is easy to see that Proposition 3.1 implies the following bound on the gradient norm of the last iterate:

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

When taking largest possible stepsize $\eta = c(\frac{1}{L} \wedge \frac{\mu}{\ell_\Xi^2})$, this bound matches the gradient norm bound in the original ROOT-SGD chapter **?**, 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 **?**. 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 $(z_t)_{t\geq\mathscr{B}}$,[4] one could extract the main part of the variance and bound the additional parts using Proposition 3.1. The multiplicative constant in such bounds will only contribute to the high-order terms in the final conclusion. See Theorem 3.3 and its proofs for details.

Note that the bounds in Proposition 3.1 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 **??**, 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 3.3. *Under above set-up, given $\alpha\in(0,1)$, there exists constants $c_1>0$ depending only on α, such that the iterates* (3.5) *with any burn-in time $\mathscr{B}\geq c\big(\frac{\ell_\xi^2}{\mu^2}+\frac{L}{\mu}\big)$ and stepsize sequence $\eta_t=\frac{1}{c\mu\mathscr{B}^{1-\alpha}t^\alpha}$ satisfies the bound:*

$$\mathbb{E}\,\|\nabla F(\theta_T)\|_2^2\leq\left(1+c\left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\right)\cdot\frac{\sigma_*^2}{T}+c\log T\cdot\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\|\nabla F(\theta_0)\|_2^2 \tag{3.11a}$$

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

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_N)\right\|_2^2\leq\left(1+c\left(\frac{\mathscr{B}}{N}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\log^2 N\right)\frac{\sigma_*^2}{N} \tag{3.11b}$$

See §3.4.2 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.

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_N)\right\|_2^2\leq(1+\omega)\frac{\mathrm{Tr}(\Sigma^*)}{N}\qquad\text{for } N\geq\frac{c}{\omega^3}\left(\frac{L}{\mu}+\frac{\ell_\xi^2}{\mu^2}\right)\log^3\frac{\mathscr{B}}{\omega} \tag{3.12}$$

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 thresh-

[4] It can be shown that the process $(tz_t)_{t\geq\mathscr{B}}$ is a martingale adapted to the natural filtration (see §3.8 for details).

old 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 13. This establish a gradient-norm result complementary to the function value bound in Frostig et al. (2015). The gradient norm bound does not require the self-concordant condition needed in Frostig et al. (2015), 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 (3.12), 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 (3.12) lead to an $\widetilde{O}(N^{-4/3})$ bound on the additional term, achieved by the stepsize choice $\eta_t = \frac{1}{c\mu\mathscr{B}^{2/3}t^{1/3}}$. 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 Hessian Lipschitz condition on the objective function F, we can obtain an improved $\widetilde{O}(N^{-3/2})$ bound on the additional term.

3.3.2 Upper bounds on the estimation error

To obtain a precise upper bound for the estimation error $\mathbb{E}\|\theta_T - \theta^*\|_2^2$ that matches the asymptotic limit, we need the following one-point Hessian Lipschitz condition, as a quantitative counterpart of the continuity Assumption 15:

Assumption 15′ *There exists $L_2 > 0$, such that for any $\theta \in \mathbb{R}^d$, we have:*

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

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 $\frac{1}{N}\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)$. 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 **?**Gadat & Panloup (2017).

Assumption 13′ *The noise function $\theta \mapsto \nabla_\theta f(\theta,\xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\|_2^4 \le \widetilde{\ell_\Xi}^4 \|\theta_1 - \theta_2\|_2^4 \qquad \textit{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d \quad (3.13)$$

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

Assumption 14′ *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has bounded fourth moment: $\widetilde{\sigma_*}^4 := \mathbb{E}\|\nabla f(\theta^*;\xi)\|_2^4$ is finite.*

By Hölder's inequality, it is clear that the constants in Assumptions 13′ and 14′ are larger than their second-moment counterparts, i.e., $\ell_\Xi \le \widetilde{\ell_\Xi}$ and $\sigma_* \le \widetilde{\sigma_*}$.

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 3.1.

Proposition 3.2. *Under Assumptions 12, 13′, and 14′, there exist universal constants $c_1, c_2, C > 0$, using burn-in time $\mathscr{B} \ge C(\frac{\widetilde{\ell_\Xi}^2}{\mu^2} + \frac{L}{\mu})$, 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 $t > \mathscr{B}$. We have the following bounds for any $T \ge 2\mathscr{B}\log\mathscr{B}$:*

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

See §3.4.3 for the proof of this claim.

Compared to Proposition 3.1, the variance parameters (σ_*, ℓ_Ξ) are replaced with their fourth-moment counterparts $(\widetilde{\sigma_*}, \widetilde{\ell_\Xi})$. 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 3.1, these terms appear only in the high-order terms of Theorem 3.4.

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:

$$\mathscr{H}_T^{(\nabla)} := \log T \cdot \frac{\sigma_*^2}{T}\left(\frac{\mathscr{B}}{T}\right)^{\alpha\wedge 1-\alpha} + \log T \cdot \mathbb{E}\|\nabla F(\theta_0)\|_2^2\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \tag{3.14a}$$

$$\widetilde{r}_T := \frac{\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{\mathscr{B}}{T}} \cdot \left(\mathbb{E}\|\nabla F(\theta_0)\|_2^4\right)^{1/4} \quad \text{and} \tag{3.14b}$$

$$\mathscr{H}_N^{(\sigma)} := \frac{\sigma_*^2\log^2 N}{\lambda_{\min}(H^*)^2 N}\left(\frac{\mathscr{B}}{N}\right)^{\alpha\wedge 1-\alpha} + \frac{L_2\widetilde{\sigma_*}^3\log^2 N}{\lambda_{\min}(H^*)\mu^3 N^{3/2}} + \frac{L_2^2\widetilde{\sigma_*}^4\log^2 N}{\lambda_{\min}(H^*)^2\mu^4 N^2} \tag{3.14c}$$

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 $(z_t)_{t\ge\mathscr{B}}$ and $(v_t)_{t\ge\mathscr{B}}$, 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 15′, 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 **??**.

Theorem 3.4. *Under Assumptions 12, 13′, 14′ and 15′, there exists universal constant $c, c_1 > 0$, for burn-in-time $\mathscr{B} = c\left(\frac{\widetilde{\ell_\Xi}^2}{\mu^2} + \frac{L}{\mu}\right)$ and stepsize $\eta_t = \frac{c_1}{\mu \mathscr{B}^{1-\alpha_t \alpha}}$ for $t \geq \mathscr{B}$, we have the following bounds holding true for $t \geq 2\mathscr{B}\log\mathscr{B}$:*

$$\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2 \leq \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{T} + \frac{c\mathscr{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)^2} + \frac{cL_2\widetilde{r}_T^3}{\lambda_{\min}(H^*)} + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)^2} \tag{3.15a}$$

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

$$\mathbb{E}\left\|\widehat{\theta}_N - \theta^*\right\|_2^2 \leq \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{N} + c\mathscr{H}_N^{(\sigma)} \tag{3.15b}$$

See §3.4.4 for the proof of this theorem.

A few remarks are in order. First, we note that the asymptotically optimal $\frac{1}{N}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)$ 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 (3.15a) and (3.15b). 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 **?** and the $O(N^{-5/4})$ bound in the chapter **?**Gadat & Panloup (2017). It is easy to see that the bound (3.15b) 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 **??**. 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 $\widetilde{O}\left(\frac{\sigma_*^2\sqrt{\mathscr{B}}}{\lambda_{\min}(H^*)^2 N^{3/2}}\right)$ 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 $O\left(\frac{\sigma_*}{\mu\sqrt{N}}\right)$, which is the rate for estimating θ^*. Under Assumption 13, the variance for gradient noise at $\theta \in \mathbb{B}\left(\theta^*, \frac{\sigma_*}{\mu\sqrt{N}}\right)$, pre-conditioned with Hessian H^*, scales as:

$$\begin{aligned}
&\mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta)\right\|_2^2 \\
&\leq \mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta^*)\right\|_2^2 \\
&\quad + 2\sqrt{\mathbb{E}\left\|(H^*)^{-1}\left(\varepsilon_t(\theta) - \varepsilon_t(\theta^*)\right)\right\|_2^2 \cdot \mathbb{E}\left\|(H^*)^{-1}\varepsilon_t(\theta^*)\right\|_2^2} + \mathbb{E}\left\|(H^*)^{-1}\left(\varepsilon_t(\theta) - \varepsilon_t(\theta^*)\right)\right\|_2^2
\end{aligned}$$

$$\leq \operatorname{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big) + 2\frac{\ell_\Xi \sigma_*}{\lambda_{\min}(H^*)^2}\|\theta-\theta^*\|_2 + \frac{\ell_\Xi^2}{\lambda_{\min}(H^*)^2}\|\theta-\theta^*\|_2^2$$
$$= \operatorname{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big) + O\left(\frac{\sigma_*^2\ell_\Xi}{\mu\lambda_{\min}^2(H^*)N^{3/2}}\right)$$

The above derivations is tight in the worst case. Compared to the term $\widetilde{O}\left(\frac{\sigma_*^2\sqrt{\mathscr{B}}}{\lambda_{\min}(H^*)^2N^{3/2}}\right)$ in our bound (3.15b), the difference is that we replace $\frac{\ell_\Xi^2}{\mu^2}$ with $\mathscr{B} = c\left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right)$ and is optimal up to a polylogarithmic factor when $\frac{L}{\mu} \lesssim \frac{\ell_\Xi^2}{\mu^2}$.[6]

- 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:

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

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 3.4 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 3.5. *Under the same setup as in Theorem 3.4, we have the following bounds on the excess risk:*

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \frac{\operatorname{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{2T} + \frac{c\mathscr{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)} + cL_2\widetilde{r}_T^3 + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)} \tag{3.16a}$$

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

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \frac{\operatorname{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{2N} + c\lambda_{\min}(H^*)\cdot\mathscr{H}_N^{(\sigma)} \tag{3.16b}$$

See §3.4.5 for the proof of this theorem.

Note that under the one-point Hessian-Lipschitz Assumption 15′, 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 3.5 differ from those in Theorme 3.4 by a factor of $\lambda_{\min}(H^*)$. This bound replaces the self-concordance assumption in Frostig

[6] In the opposite regime, the non-tight prefactor is considered unavoidable because our method does not account for N in advance.

et al. (2015) with a less structural one-point Hessian-Lipschitz condition. Theorem 3.5 and their results are not comparable in general, as they are based on different assumptions. When taking the optimal trade-off, Theorem 3.5 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 Frostig et al. (2015), and improves the bounds for streaming SVRG in Frostig et al. (2015) in terms of the rate of convergence for the additional term.

3.4 Proof of the non-asymptotic bounds with sharp pre-factor

In this section, we present the proofs for Theorem 3.3, Theorem 3.4 and Theorem 3.5. 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. En route our proof, in §3.4.1 and §3.4.3, we present the proofs of Proposition 3.1 and Proposition 3.2, the non-asymptotic convergence rates for the process $(v_t)_{t\geq\mathscr{B}}$ and $(z_t)_{t\geq\mathscr{B}}$. These results serve as the basic building blocks for the fine-grained asymptotic and non-asymptotic guarantees.

3.4.1 Proof of Proposition 3.1

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 3.6. *Under Assumption 12, 13, 14, when* $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{\ell_\Xi^2}$, *we have:*

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

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

Lemma 3.7. *Under Assumptions 12, 13 and 14, for* $t \geq 1$, *we have:*

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

The proofs of the Lemmas are postponed to Section 3.8.1 and Section 3.8.2 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:

$$\mathbb{E}\left\|\nabla F(\theta_t)\right\|_2^2 \leq 2\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2+2\mathbb{E}\left\|\nabla F(\theta_t)-\nabla F(\theta_{t-1})\right\|_2^2$$

$$\leq 2\mathbb{E}\|v_t - z_t\|_2^2 + 2L^2\eta_t^2\mathbb{E}\|v_t\|_2^2 \leq 6\mathbb{E}\|v_t\|_2^2 + 4\mathbb{E}\|z_t\|_2^2$$

Therefore, by Lemma 3.6, 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:

$$t^2\mathbb{E}\|v_t\|_2^2 \leq (1 - c\eta_{t-1}\mu)(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + \frac{2C}{\mu\eta_{t-1}}\mathbb{E}\|z_{t-1}\|_2^2 + 2\sigma_*^2 \tag{3.17}$$

for a universal constant $C > 0$.

Similarly, by Lemma 3.7, if s satisfies $s\eta_{s-1}\mu > \frac{1}{4C}$ for any $s > \mathscr{B}$, we have:

$$\begin{aligned}\mathbb{E}\|z_t\|_2^2 &\leq \frac{\mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2t^2}\sum_{s=\mathscr{B}}^{t-1}(\mathbb{E}\|z_s\|_2^2 + \mathbb{E}\|v_s\|_2^2) + \frac{\ell_\Xi^2}{t^2}\sum_{s=\mathscr{B}}^{t-1}s^2\eta_s^2\mathbb{E}\|v_s\|_2^2 \\ &\leq \frac{\mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2t^2}\sum_{s=\mathscr{B}}^{t-1}\mathbb{E}\|z_s\|_2^2 + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=\mathscr{B}}^{t-1}s^2\eta_s^2\mathbb{E}\|v_s\|_2^2\end{aligned} \tag{3.18}$$

for a universal constant $C' > 0$.

Note that the bounds (3.17) and (3.18) give recursive upper bounds on the second moments of the processes $(z_t)_{t\geq\mathscr{B}}$ and $(v_t)_{t\geq\mathscr{B}}$, i.e., they bound the quantities $\mathbb{E}\|z_t\|_2^2$ and $\mathbb{E}\|v_t\|_2^2$ based on their history. In the following, we solve the recursive inequalities.

We define the following quantities for $T \geq \mathscr{B}$:

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

First, for any $T > \mathscr{B}$, by taking the supremum in Eq (3.18) over $t \in [\mathscr{B}, T]$, we obtain the following bound:

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

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

$$C\frac{\ell_\Xi^2}{\mu^2}\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t}\frac{1}{s}\le\frac{C\ell_\Xi^2}{\mu^2\mathscr{B}}<\frac{1}{2}$$

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:

$$H_T\le 2H_{\mathscr{B}}+4\sigma_*^2+2C'\ell_\Xi^2\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\eta_{s-1}^2W_s \tag{3.19a}$$

On the other hand, the bound (3.17) implies the bound:

$$W_T\le(1-c\eta_{T-1}\mu)W_{T-1}+\frac{C}{T\mu\eta_{T-1}}H_{T-1}+2\sigma_*^2 \tag{3.19b}$$

for universal constants $c,C>0$.

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

Lemma 3.8. *For a pair of sequences $(H_t)_{t\ge\mathscr{B}}$ and $(W_t)_{t\ge\mathscr{B}}$ satisfying the recursive relation* (3.19a) *with non-increasing stepsize sequence $(\eta_t)_{t\ge\mathscr{B}}$. Assuming that $(H_t)_{t\ge\mathscr{B}}$ is non-decreasing, there exists universal constants $c>0$, such that for $T\ge\mathscr{B}$, we have the bound:*

$$H_T\le c\left(\sigma_*^2+\frac{\ell_\Xi^2\mathscr{B}\eta_{\mathscr{B}}}{\mu}W_{\mathscr{B}}+H_{\mathscr{B}}\right)\qquad\textit{and} \tag{3.20a}$$

$$W_T\le\frac{c}{\eta_T\mu}\sigma_*^2+c\left(\frac{\mathscr{B}}{T\mu^2\eta_{T-1}^2}+e^{-\mu\sum_{t=\mathscr{B}+1}^{T}\eta_t}\mathscr{B}^2\right)W_{\mathscr{B}} \tag{3.20b}$$

See Section 3.8.3 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 (3.20b) satisfies the bound:

$$\mu\sum_{t=\mathscr{B}+1}^{T}\eta_t\ge c_1\sum_{t=\mathscr{B}+1}^{T}\frac{1}{t}\ge c_1\log\frac{T}{\mathscr{B}}$$

For $c_1\ge 2$ and $\eta_T\le\frac{c'}{\mu\mathscr{B}}$, we have that $\frac{\mathscr{B}}{T\mu^2\eta_{T-1}^2}\ge e^{-\mu\sum_{t=\mathscr{B}+1}^{T}\eta_t}\mathscr{B}^2$. So the bound (3.20b) implies that:

$$\mathbb{E}\|v_t\|_2^2\le\frac{c\sigma_*^2}{\mu\eta_t t^2}+\frac{c\mathscr{B}}{t^3\eta_t^2\mu^2}\mathbb{E}\|v_{\mathscr{B}}\|_2^2$$

For the process z_t, by substituting the bounds in Lemma 3.8 into Eq (3.18), for stepsize $\eta_t<\frac{1}{\mu\mathscr{B}}$, we obtain:

$$\mathbb{E}\|z_t\|_2^2\le\frac{\mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2}{t^2}+\frac{2\sigma_*^2}{t}+C\frac{\ell_\Xi^2H_t}{\mu^2t^2}\left(\sum_{s=\mathscr{B}}^{t-1}\frac{1}{s}\right)+C'\frac{\ell_\Xi^2}{t^2}\sum_{s=\mathscr{B}}^{t-1}\eta_s^2W_s$$

$$\leq \frac{\mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2}{t^2}+\frac{2\sigma_*^2}{t}+C\frac{\ell_\Xi^2\log t}{\mu^2t^2}\left(\sigma_*^2+\frac{\ell_\Xi^2\mathscr{B}\eta_{\mathscr{B}}}{\mu}\mathbb{E}\|v_{\mathscr{B}}\|_2^2\right)+C'\frac{\ell_\Xi^2}{t^2}\sum_{s=\mathscr{B}+1}^{t}\eta_s\frac{\sigma_*^2}{\mu}+C'\frac{\ell_\Xi^2}{t^2}\sum_{s=\mathscr{B}}^{t-1}\frac{\mathscr{B}}{s\mu^2}\mathbb{E}\|v_{\mathscr{B}}\|_2^2$$

$$\leq c\frac{\sigma_*^2}{t}+c\frac{\mathscr{B}^2\mathbb{E}\|z_{\mathscr{B}}\|_2^2}{t^2}+c\frac{\ell_\Xi^2\mathscr{B}\log t}{\mu^2t^2}\mathbb{E}\|v_{\mathscr{B}}\|_2^2$$

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

$$\mathbb{E}\|z_{\mathscr{B}}\|_2^2=\mathscr{B}^{-2}\mathbb{E}\left\|\sum_{t=0}^{\mathscr{B}}\varepsilon_t(\theta_0)\right\|_2^2\leq\frac{\sigma_*^2+\ell_\Xi^2\|\theta_0-\theta^*\|_2^2}{\mathscr{B}}$$

$$\mathbb{E}\|v_{\mathscr{B}}\|_2^2\leq 2\|\nabla F(\theta_0)\|_2^2+\mathbb{E}\|z_{\mathscr{B}}\|_2^2\leq 2\|\nabla F(\theta_0)\|_2^2+\frac{2(\sigma_*^2+\ell_\Xi^2\|\theta_0-\theta^*\|_2^2)}{\mathscr{B}}$$

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

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

Putting them together, we have the bounds:

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

which complete the proof of this proposition.

3.4.2 Proof of Theorem 3.3

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 12, we have the bound $\mathbb{E}\|\theta_T-\theta^*\|_2^2\leq r_\theta(t)^2$.

We start by observing the following decomposition:

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2=\mathbb{E}\|z_{T+1}\|_2^2+\mathbb{E}\|v_{T+1}\|_2^2-2\mathbb{E}\langle z_{T+1},v_{T+1}\rangle\tag{3.21}$$

The following lemma provides sharp bounds on the leading-order term $\mathbb{E}\|z_{T+1}\|_2^2$.

Lemma 3.9. *Under above set-up, for $T \geq 2\mathscr{B}\log\mathscr{B}$ and any $G \in \mathbb{R}^{d\times d}$, the following bounds hold true for the process $(z_t)_{t\geq\mathscr{B}}$:*

$$\mathbb{E}\,\|Gz_T\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left(G\Sigma^* G^\top\right) + c\|\!|G|\!\|_{op}^2 \mathscr{H}_T^{(z)} \tag{3.22a}$$

where the high order term $\mathscr{H}_T^{(z)}$ is defined as

$$\mathscr{H}_T^{(z)} := c\left(\sqrt{\frac{\mathscr{B}}{T}} + \frac{\mathscr{B}^\alpha}{T^\alpha}\right)\frac{\sigma_*^2}{T} + c\frac{\mathscr{B}^2\log T}{T^2}\left(1 + \frac{T^{2\alpha-3/2}}{\mathscr{B}^{2\alpha-3/2}}\right)\|\nabla F(\theta_0)\|_2^2 \tag{3.22b}$$

See §3.8.4 for the proof of this lemma.

Invoking Proposition 3.1, we have the bound for v_T:

$$\mathbb{E}\,\|v_T\|_2^2 \leq c\left(\frac{\sigma_*^2}{\mu\eta_T T^2} + \frac{\mathscr{B}}{\mu^2 T^3\eta_T^2}\|\nabla F(\theta_0)\|_2^2\right)$$

For the stepsize choice $\eta_t = \frac{1}{\mu\mathscr{B}^{1-\alpha}t^\alpha}$, we have the bound

$$\mathbb{E}\,\|v_T\|_2^2 \leq c\frac{\mathscr{B}^{1-\alpha}}{T^{1-\alpha}}\cdot\frac{\sigma_*^2}{T} + c\frac{\mathscr{B}^{3-2\alpha}}{T^{3-2\alpha}}\cdot\|\nabla F(\theta_0)\|_2^2 \tag{3.23}$$

Combining the bounds (3.22a) and (3.23) and substituting into the decomposition (3.21), we arrive at the following bound by applying Young's inequality:

$$\begin{aligned}
\mathbb{E}\,\|\nabla F(\theta_T)\|_2^2 &\leq \mathbb{E}\,\|z_{T+1}\|_2^2 + \mathbb{E}\,\|v_{T+1}\|_2^2 + 2\sqrt{\mathbb{E}\,\|z_{T+1}\|_2^2}\cdot\sqrt{\mathbb{E}\,\|v_{T+1}\|_2^2} \\
&\leq \left(1 + \left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}}\right)\cdot\mathbb{E}\,\|z_{T+1}\|_2^2 + \left(1 + \left(\frac{T}{\mathscr{B}}\right)^{\frac{1-\alpha}{2}}\right)\mathbb{E}\,\|v_{T+1}\|_2^2 \\
&\leq \frac{\sigma_*^2}{T} + c\left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T\cdot\|\nabla F(\theta_0)\|_2^2
\end{aligned}$$

which proves the first claim (3.11a).

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 $b = 1, 2, \cdots, B$:

$$\begin{aligned}
\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^2 &\leq \frac{\sigma_*^2}{T^\flat} + c\left(\frac{\mathscr{B}}{T^\flat}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T^\flat} + c\left(\frac{\mathscr{B}}{T^\flat}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T\cdot\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2 \\
&\overset{(i)}{\leq} \frac{2\sigma_*^2}{T^\flat} + \frac{1}{2}\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2
\end{aligned}$$

In step (i), we use the fact that $T^\flat \geq 2c\mathscr{B}\log\mathscr{B}$ and that $2\wedge\frac{5-3\alpha}{2} > 1$ for $\alpha\in(0,1)$.

Solving the recursion, we arrive at the bound:

$$\mathbb{E}\left\|\nabla F\left(\theta_0^{(B+1)}\right)\right\|_2^2 \leq \frac{4\sigma_*^2}{\mathscr{B}} + 2^{-B}\mathbb{E}\|\nabla F(\theta_0)\|_2^2$$

Substituting this initial condition into the bound (3.11a), we obtain the final bound:

$$\mathbb{E}\left\|\nabla F\left(\theta_T^{(B+1)}\right)\right\|_2^2 \leq \frac{\sigma_*^2}{T} + c\left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T\cdot\left(\frac{4\sigma_*^2}{\mathscr{B}} + 2^{-B}\mathbb{E}\|\nabla F(\theta_0)\|_2^2\right)$$

Taking $B \geq \log_2\left(\frac{\mathscr{B}\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2}\right)$ and substituting with $T = N - BT^\flat$, we arrive at the conclusion:

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_N)\right\|_2^2 \leq \left(1 + c\left(\frac{\mathscr{B}}{N}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\log^2 N\right)\frac{\sigma_*^2}{N}$$

which proves the bound (3.11b).

3.4.3 Proof of Proposition 3.2

Throughout the proof, we frequently use the following inequalities for the moments of stochastic gradients, which holds true for any $\theta \in \mathbb{R}^d$:

$$\mathbb{E}\|\nabla f(\theta,\xi_t)\|_2^4 \leq 27\widetilde{\sigma_*}^4 + 27\left(1 + \frac{\widetilde{\ell_\Xi}^4}{\mu^4}\right)\mathbb{E}\|\nabla F(\theta)\|_2^4 \tag{3.24}$$

To see why this is true, we note that:

$$\begin{aligned}
\mathbb{E}\|\nabla f(\theta,\xi_t)\|_2^4 &\leq 27\mathbb{E}\|\nabla F(\theta)\|_2^4 + 27\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^4 + 27\mathbb{E}\|\varepsilon_t(\theta) - \varepsilon_t(\theta^*)\|_2^4 \\
&\leq 27\widetilde{\sigma_*}^4 + 27\mathbb{E}\|\nabla F(\theta)\|_2^4 + 27\widetilde{\ell_\Xi}^4\mathbb{E}\|\theta - \theta^*\|_2^4 \\
&\leq 27\widetilde{\sigma_*}^4 + 27\left(1 + \frac{\widetilde{\ell_\Xi}^4}{\mu^4}\right)\mathbb{E}\|\nabla F(\theta)\|_2^4
\end{aligned}$$

Now we turn to the proof of this proposition. Similar to the proof of Proposition 3.1, we need the following technical lemmas:

Lemma 3.10. *Under Assumption 12, 13′, 14, there exists universal constants $c, c' > 0$, when $\eta_t \leq c\left(\frac{1}{L} \wedge \frac{\mu}{\widetilde{\ell_\Xi}^2}\right)$, we have the bound*

$$\sqrt{\mathbb{E}\|v_t\|_2^4} \leq \left(1 - \frac{1}{t}\right)^2\left(1 - \frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\|v_{t-1}\|_2^4} + \frac{c'}{t^2}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4}\right)$$

Lemma 3.11. *Under Assumption 13′, we have the bound*

$$\sqrt{\mathbb{E}\left\|z_t\right\|_2^4} \leq \frac{c\mathscr{B}^2\left\|z_0\right\|_2^2}{t^2} + \frac{c\widetilde{\sigma_*}^2}{t} + \frac{c\widetilde{\ell_\Xi}^2}{\mu^2 t^2}\sum_{s=\mathscr{B}}^{t-1}\sqrt{\mathbb{E}\left\|\nabla F(\theta_s)\right\|_2^4} + \frac{c\widetilde{\ell_\Xi}^2}{t^2}\sum_{s=\mathscr{B}}^{t-1} s^2\eta_s^2\sqrt{\mathbb{E}\left\|v_s\right\|_2^4}$$

See Section 3.8.5 and 3.8.6 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 3.1. We first note that:

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

Substituting into the bounds in Lemma 3.10 and 3.11, and defining the quantities $H_T := \sup_{\mathscr{B}\leq t\leq T} t\sqrt{\mathbb{E}\left\|z_t\right\|_2^4}$, $W_T := T^2\sqrt{\mathbb{E}\left\|v_T\right\|_2^4}$, we arrive at the following recursive inequalities:

$$H_T \leq 2H_{\mathscr{B}} + 4\widetilde{\sigma_*}^2 + 2C'\widetilde{\ell_\Xi}^2 \sup_{\mathscr{B}\leq t\leq T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\eta_{s-1}^2 W_s \tag{3.25a}$$

$$W_T \leq (1 - c\eta_{T-1}\mu)W_{T-1} + \frac{C}{T\mu\eta_{T-1}}H_{T-1} + 2\widetilde{\sigma_*}^2 \tag{3.25b}$$

Invoking Lemma 3.8 by replacing (ℓ_Ξ, σ_*) with $(\widetilde{\ell_\Xi}, \widetilde{\sigma_*})$, we obtain the following bounds:

$$\begin{aligned}H_T &\leq c\left(\sigma_*^2 + \frac{\ell_\Xi^2\mathscr{B}\eta_{\mathscr{B}}}{\mu}W_{\mathscr{B}} + H_{\mathscr{B}}\right) \qquad \text{and} \\ W_T &\leq \frac{c}{\eta_T\mu}\sigma_*^2 + c\left(\frac{\mathscr{B}}{T\mu^2\eta_{T-1}^2} + e^{-\mu\sum_{t=\mathscr{B}+1}^T\eta_t}\mathscr{B}^2\right)W_{\mathscr{B}}\end{aligned}$$

For the initial conditions, by applying Khintchine's inequality as well as Young's inequality, we note that:

$$\mathbb{E}\left\|z_{\mathscr{B}}\right\|_2^4 = \mathscr{B}^{-4}\mathbb{E}\left\|\sum_{t=1}^{\mathscr{B}}\varepsilon_t(\theta_0)\right\|_2^4 \leq \mathscr{B}^{-4}\mathbb{E}\left(\sum_{t=1}^{\mathscr{B}}\left\|\varepsilon_t(\theta_0)\right\|_2^2\right)^2 \leq 8\mathscr{B}^{-2}\left(\widetilde{\sigma_*}^4 + \widetilde{\ell_\Xi}^4\left\|\theta_0 - \theta^*\right\|_2^4\right)$$

$$\mathbb{E}\left\|v_{\mathscr{B}}\right\|_2^4 \leq 8\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4 + \mathbb{E}\left\|z_{\mathscr{B}}\right\|_2^4 \leq 8\left\|\nabla F(\theta_0)\right\|_2^4 + 8\mathscr{B}^{-2}\left(\widetilde{\sigma_*}^4 + \widetilde{\ell_\Xi}^4\left\|\theta_0 - \theta^*\right\|_2^4\right)$$

Following exactly the same arguments as in the proof of Proposition 3.1, we arrive at the desired bounds.

3.4.4 Proof of Theorem 3.4

We define the quantities $r_\theta(t)$ and $r_v(t)$ the same as in the proof of Theorem 3.3. Furthermore, we denote the following quantities:

$$\widetilde{r}_v(t) := \left(\mathbb{E}\,\|v_t\|_2^4\right)^{1/4} \qquad \text{and} \quad \widetilde{r}_\theta(t) := \frac{1}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_t)\|_2^4\right)^{1/4}$$

Clearly, by the strong convexity Assumption 12, 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 15′:

$$\begin{aligned}\left\|(H^*)^{-1}\nabla F(\theta_T) - (\theta_T - \theta^*)\right\|_2 &\leq \int_0^1 \left\|(H^*)^{-1}\left(\nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\right\|_2 d\gamma \\ &\leq \frac{L_2}{\lambda_{\min}(H^*)}\,\|\theta_T - \theta^*\|_2^2 \leq \frac{L_2}{\lambda_{\min}(H^*)\mu^2}\,\|\nabla F(\theta_T)\|_2^2\end{aligned} \tag{3.26}$$

We can then upper bound the mean-squared error using the processes $(z_t)_{t\geq\mathscr{B}}$ and $(v_t)_{t\geq\mathscr{B}}$:

$$\begin{aligned}\mathbb{E}\,\|\theta_T - \theta^*\|_2^2 &\leq \mathbb{E}\left(\left\|(H^*)^{-1}\nabla F(\theta_T)\right\|_2 + \frac{L_2}{\mu^2\lambda_{\min}(H^*)}\,\|\nabla F(\theta_T)\|_2^2\right)^2 \\ &\leq \mathbb{E}\left\|(H^*)^{-1}\big(v_{T+1} - z_{T+1}\big)\right\|_2^2 + 2\frac{L_2}{\lambda_{\min}(H^*)}\widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)^2}\widetilde{r}_\theta^4(T)\end{aligned} \tag{3.27}$$

The leading-order term in the bound (3.27) admits the following decomposition:

$$\mathbb{E}\left\|(H^*)^{-1}\big(z_{T+1} - v_{T+1}\big)\right\|_2^2 = \mathbb{E}\left\|(H^*)^{-1}z_{T+1}\right\|_2^2 + \mathbb{E}\left\|(H^*)^{-1}v_{T+1}\right\|_2^2 - 2\mathbb{E}\left[\langle (H^*)^{-1}z_T, (H^*)^{-1}.v_T\rangle\right]$$

In the following, we bound the three terms in above equation, respectively. Invoking Lemma 3.9 with $G = (H^*)^{-1}$, we have the bound:

$$\mathbb{E}\left\|(H^*)^{-1}z_{T+1}\right\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{\mathscr{B}}{T}\right)^{\alpha\wedge\frac{1}{2}} + \frac{c\,\|\nabla F(\theta_0)\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \tag{3.28a}$$

For the process v_t, Proposition 3.1 yields the following upper bound:

$$\mathbb{E}\left\|(H^*)^{-1}v_{T+1}\right\|_2^2 \le \frac{1}{\lambda_{\min}(H^*)^2}\mathbb{E}\left\|v_{T+1}\right\|_2^2 \le \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{\mathscr{B}}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\lambda_{\min}(H^*)^2}\left(\frac{\mathscr{B}}{T}\right)^{3-2\alpha} \tag{3.28b}$$

The bound for the cross term is given by the following lemma:

Lemma 3.12. *Under above set-up, for $T \ge c\mathscr{B}\log\mathscr{B}$, for any $d\times d$ deterministic matrix G, the following bound holds true:*

$$\begin{aligned}\left|\mathbb{E}\left[\langle Gz_t, Gv_t\rangle\right]\right| \le c\left|\!\left|\!\left|G\right|\!\right|\!\right|_{op}^2\left(\frac{\mathscr{B}}{t}\right)^{1-\alpha}\left(\frac{\sigma_*^2}{t}+\left(\frac{\mathscr{B}}{t}\right)^{2-\alpha}\left\|\nabla F(\theta_0)\right\|_2^2\right)\log t \\ + c\frac{\left|\!\left|\!\left|G\right|\!\right|\!\right|_{op}^2 L_2}{\mu^2}\left(\frac{\mathscr{B}}{t}\right)^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma}_*^3}{t^{3/2}}+\left(\frac{\mathscr{B}}{t}\right)^{3-3\alpha/2}\log^2 t\left\|\nabla F(\theta_0)\right\|_2^3\right)\end{aligned}$$

See §3.8.7 for the proof of this lemma.

Substituting with $G=(H^*)^{-1}$, we obtain the bound for the cross term:

$$\begin{aligned}\left|\mathbb{E}\left[\langle (H^*)^{-1}z_t, (H^*)^{-1}v_t\rangle\right]\right| \le \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)^2 T}\left(\frac{\mathscr{B}}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{\mathscr{B}}{T}\right)^{3-2\alpha} \\ + \frac{cL_2\widetilde{\sigma}_*^3}{\lambda_{\min}(H^*)^2\mu^2 T^{3/2}}\left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\left\|\nabla F(\theta_0)\right\|_2^3\log^2 T}{\lambda_{\min}(H^*)^2\mu^2}\left(\frac{\mathscr{B}}{T}\right)^{\frac{7}{2}-2\alpha}\end{aligned} \tag{3.28c}$$

For the rest two terms in the expression (3.27), we invoke Proposition 3.2, and obtain the rate:

$$\widetilde{r}_\theta(T) \le \frac{c\widetilde{\sigma}_*}{\mu\sqrt{T}} + \frac{c\sqrt{\log T}}{\mu}\left\|\nabla F(\theta_0)\right\|_2\left(\frac{\mathscr{B}}{T}\right)^{1\wedge 3/2-\alpha} \tag{3.28d}$$

Combining the bounds (3.28a)-(3.28d) and substituting into the decomposition (3.27), we arrive at the bound

$$\begin{aligned}\mathbb{E}\left\|\theta_T-\theta^*\right\|_2^2 \le \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{T} + \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)^2 T}\left(\frac{\mathscr{B}}{T}\right)^{\alpha\wedge 1-\alpha} + \frac{c\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \\ + \frac{cL_2\widetilde{\sigma}_*^3}{\lambda_{\min}(H^*)\mu^3 T^{3/2}} + \frac{cL_2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^3\log^2 T}{\lambda_{\min}(H^*)\mu^3}\left(\frac{\mathscr{B}}{T}\right)^{\frac{7}{2}-2\alpha\wedge 3} \\ + \frac{cL_2^2\widetilde{\sigma}_*^4}{\lambda_{\min}(H^*)^2\mu^4 T^2} + \frac{cL_2^2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\log^2 T}{\lambda_{\min}(H^*)^2\mu^4}\left(\frac{\mathscr{B}}{T}\right)^{6-4\alpha\wedge 4}\end{aligned}$$

Noting that $\frac{7}{2}-2\alpha\wedge 3 \ge \frac{3}{2}$ and $6-4\alpha\wedge 4 \ge 4$, we complete the proof of the bound (3.15a).

Now we turn to the proof of the multi-loop result (3.15b). Invoking Proposition 3.1 and 3.2 and noting that $\left\|\nabla F(\theta_t)\right\|_2 \le \left\|z_{t+1}\right\|_2 + \left\|v_{t+1}\right\|_2$, we obtain the bound for $T^\flat \ge c\mathscr{B}\log\mathscr{B}$:

$$\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^2 \le \frac{1}{2}\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat} \quad \text{and}$$
$$\sqrt{\mathbb{E}\left\|\nabla F\left(\theta_0^{(b+1)}\right)\right\|_2^4} \le \frac{1}{2}\sqrt{\mathbb{E}\left\|\nabla F\left(\theta_0^{(b)}\right)\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat}$$

Solving the recursion, we have that:

$$\mathbb{E}\left\|\nabla F\left(\theta_0^{(B+1)}\right)\right\|_2^2 \le 2^{-B}\mathbb{E}\left\|\nabla F\left(\theta_0\right)\right\|_2^2 + \frac{2c\sigma_*^2}{T^\flat} \quad \text{and}$$
$$\sqrt{\mathbb{E}\left\|\nabla F\left(\theta_0^{(B+1)}\right)\right\|_2^4} \le 2^{-B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{2c\widetilde{\sigma_*}^2}{T^\flat}$$

Taking $B \ge \log_2\left(\frac{\mathscr{B}\mathbb{E}\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2}\right)$ and substituting into the bound (3.15a), we have the following guarantee for the multi-loop estimator:

$$\mathbb{E}\left\|\widehat{\theta}_N - \theta^*\right\|_2^2 \le \frac{\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right)}{N} + \frac{c\sigma_*^2\log^2 N}{\lambda_{\min}(H^*)^2 N}\left(\frac{\mathscr{B}}{N}\right)^{\alpha\wedge 1-\alpha}$$
$$+ \frac{cL_2\widetilde{\sigma_*}^3\log^2 N}{\lambda_{\min}(H^*)\mu^3 N^{3/2}} + \frac{cL_2^2\widetilde{\sigma_*}^4\log^2 N}{\lambda_{\min}(H^*)^2\mu^4 N^2}$$

which completes the proof.

3.4.5 Proof of Theorem 3.5

Applying second-order Taylor expansion with integral remainder, for any $\theta \in \mathbb{R}^d$, we note the following identity.

$$F(\theta) = F(\theta^*) + \langle \theta - \theta^*, \nabla F(\theta^*)\rangle + (\theta-\theta^*)^\top \int_0^1 \nabla^2 F\left(\gamma\theta + (1-\gamma)\theta^*\right) d\gamma \cdot (\theta - \theta^*)$$

Noting that $\nabla F(\theta^*) = 0$ and invoking Assumption 15′, we have that:

$$F(\theta) \le F(\theta^*) + \frac{1}{2}(\theta-\theta^*)^\top H^*(\theta-\theta^*) + \|\theta-\theta^*\|_2 \cdot \int_0^1 \left\|\nabla^2 F\left(\gamma\theta + (1-\gamma)\theta^*\right) - H^*\right\|_{\mathrm{op}} d\gamma \cdot \|\theta-\theta^*\|_2$$
$$\le F(\theta^*) + \frac{1}{2}(\theta-\theta^*)^\top H^*(\theta-\theta^*) + L_2\|\theta-\theta^*\|_2^3 \tag{3.29}$$

Similar to Eq (3.26), we have the bound:

$$\left\|(H^*)^{1/2}(\theta-\theta^*) - (H^*)^{-1/2}\nabla F(\theta)\right\|_2 \le \int_0^1 \left\|(H^*)^{-1/2}\left(\nabla^2 F\left(\gamma\theta^* + (1-\gamma)\theta_T\right) - H^*\right)(\theta_T - \theta^*)\right\|_2 d\gamma$$
$$\le \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}}\|\theta_T - \theta^*\|_2^2 \le \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}\mu^2}\|\nabla F(\theta_T)\|_2^2$$

Denote the residual $q_t := (H^*)^{1/2}(\theta_t - \theta^*) - (H^*)^{-1/2}\nabla F(\theta_t)$. Substituting into the bound (3.29), we have that:

$$\begin{aligned}\mathbb{E}[F(\theta_T)] - F(\theta^*) &\leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}\nabla F(\theta) + q_T\right\|_2^2 + L_2\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^3 \\ &\leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}(z_{T+1} + v_{T+1})\right\|_2^2 + 2L_2\widetilde{r}_\theta^3(T) + \mathbb{E}\left\|q_T\right\|_2^2 \\ &\leq \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}(z_{T+1} + v_{T+1})\right\|_2^2 + 2L_2\widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)}\widetilde{r}_\theta^4(T)\end{aligned}$$

Invoking Proposition 3.1, Lemma 3.9 and 3.14 with $G = (H^*)^{-1/2}$, we have the bounds

$$\begin{aligned}\mathbb{E}\left\|(H^*)^{-1/2}z_{T+1}\right\|_2^2 &\leq \frac{\mathrm{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{\mathscr{B}}{T}\right)^{\alpha\wedge\frac{1}{2}} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \\ \mathbb{E}\left\|(H^*)^{-1/2}v_{T+1}\right\|_2^2 &\leq \frac{\mathbb{E}\left\|v_{T+1}\right\|_2^2}{\lambda_{\min}(H^*)} \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{\mathscr{B}}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\lambda_{\min}(H^*)}\left(\frac{\mathscr{B}}{T}\right)^{3-2\alpha}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\langle(H^*)^{-1/2}z_t, (H^*)^{-1/2}v_t\rangle\right] &\leq \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)T}\left(\frac{\mathscr{B}}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{\mathscr{B}}{T}\right)^{3-2\alpha} \\ &+ \frac{cL_2\widetilde{\sigma}_*^3}{\lambda_{\min}(H^*)\mu^2T^{3/2}}\left(\frac{\mathscr{B}}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\left\|\nabla F(\theta_0)\right\|_2^3\log^2 T}{\lambda_{\min}(H^*)\mu^2}\left(\frac{\mathscr{B}}{T}\right)^{\frac{7}{2}-2\alpha} \qquad (3.30)\end{aligned}$$

Putting them together, we arrive at the bound

$$\begin{aligned}\mathbb{E}[F(\theta_T) - F(\theta^*)] \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*\big)}{2T} &+ \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)T}\left(\frac{\mathscr{B}}{T}\right)^{\alpha\wedge 1-\alpha} \\ &+ \frac{c\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)}\left(\frac{\mathscr{B}}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} + cL_2\widetilde{r}_T^3 + c\frac{L_2^2}{\mu}\widetilde{r}_T^4\end{aligned}$$

for the quantity $\widetilde{r}_T := \frac{\widetilde{\sigma}_*}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{\mathscr{B}}{T}}\cdot\left(\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\right)^{1/4}$.

For the multi-loop algorithm, applying the same argument on the initial gradient norm as in the proof of Theorem 3.4, we arrive at the desired bound.

3.5 Proof of asymptotic results

In this section, we present the proofs for the asymptotic results, Theorem 3.1 and Theorem 3.2. The former guarantees the asymptotic normality of ROOT-SGD un-

der our assumptions, while the latter shows an example that satisfies our assumptions but makes Polyak-Ruppert algorithm fail asymptotically.

3.5.1 Proof of Theorem 3.1

By Proposition 3.1, for $t \geq \mathscr{B}$, taking $\eta_t = \frac{1}{\mu \mathscr{B}^{1-\alpha} t^\alpha}$, 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 $t \geq 2\mathscr{B}\log\mathscr{B}$, we have the bounds:

$$\mathbb{E}\|v_t\|_2^2 \leq a_1\left(\frac{1}{t^2\eta_t} + \frac{1}{t^3\eta_t^2} + \frac{1}{t^2}\right) \leq \frac{3a_1}{t^{2-\alpha}}$$
$$\mathbb{E}\|z_t\|_2^2 \leq \frac{a_2}{t} + \frac{a_2\log t}{t^2} \leq \frac{2a_2}{t}$$

and consequently, we have:

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

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

For the martingale Ψ_t, we note that:

$$\mathbb{E}\|\Psi_t\|_2^2 = \sum_{s=\mathscr{B}}^{t}(s-1)^2\mathbb{E}\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\|_2^2 \leq \sum_{s=\mathscr{B}}^{t}(s-1)^2\ell_\Xi^2\mathbb{E}\|\theta_{s-1} - \theta_{s-2}\|_2^2$$
$$\leq \sum_{s=\mathscr{B}}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\|v_{s-1}\|_2^2 \leq \frac{1}{\mu^2\mathscr{B}^{2-2\alpha}}\sum_{s=0}^{t-1}s^{2-2\alpha}\cdot\frac{3a_1}{s^{2-\alpha}} \leq \frac{3a_1}{(1-\alpha)\mu^2\mathscr{B}^{2-2\alpha}}t^{1-\alpha}$$

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

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

Putting together the pieces, we obtain:

$$t\mathbb{E}\left\|z_t - \frac{1}{t}N_t\right\|_2^2 \leq \frac{3}{t}\|z_0\|_2^2 + \frac{3}{t}\mathbb{E}\|\Psi_t\|_2^2 + \frac{3}{t}\mathbb{E}\|M_t - N_t\|_2^2$$
$$\leq \frac{3}{t}\|z_0\|_2^2 + \frac{3a_1t^{1-\alpha}}{(1-\alpha)\mu^2\mathscr{B}^{2-2\alpha}t} + \frac{3}{t}\cdot\ell_\Xi^2 C\log t \to 0 \qquad (3.31)$$

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

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

The second moment bound (3.31) implies that:

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

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:

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

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 12, 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.

3.5.2 Proof of Theorem 3.2

The proof is by explicit construction of a function (and associated noise) satisfying the Assumptions 12, 13, 14 and 15, 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 13 and 14 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:

$$\forall x > 0 \qquad F_\eta(x) := \eta^{-1} F\left(\sqrt{\eta}x\right) \tag{3.32}$$

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 $\bar{\psi}_T := \bar{\theta}_T/\sqrt{\eta}$. 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 **?** (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 3.3 (Special case of ?, Theorem 5). *Under above setup, we have the following bound for $k = 1, 2 \cdots$*

$$\mathscr{W}_2^2\left(\pi_\eta^{(k)}, \pi_\eta\right) \leq 2e^{-c_1\eta k}\left(\|\psi_0\|_2^2 + 1\right) + c_2\eta \tag{3.33a}$$

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 **?**.

Proposition 3.4 (Special case of ?, Eq (27) and Theorem 15). *Under above setup, given any Lipschitz test function h, let $\bar{h}_{\mathscr{B},T} := \frac{1}{T-\mathscr{B}}\sum_{t=\mathscr{B}}^{T-1} h(\psi_t)$, the following bounds hold true:*

$$\left(\mathbb{E}\left[\bar{h}_{\mathscr{B},T}\right] - \mathbb{E}_{\pi_\eta}\left[h(X)\right]\right)^2 \leq \frac{\|h\|_{\mathrm{Lip}}^2}{T-\mathscr{B}} \sum_{t=\mathscr{B}}^{T} \mathscr{W}_2^2\left(\pi_\eta^{(t)}, \pi_\eta\right) \tag{3.33b}$$

$$var\left(\bar{h}_{\mathscr{B},T}\right) \leq c\frac{\|h\|_{\mathrm{Lip}}^2}{\eta(T-\mathscr{B})} \tag{3.33c}$$

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

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

Substituting into the MSE bound in Proposition 3.4, for any choice of burn-in parameter $\beta \in [0,1)$, we have the bound:

$$\mathbb{E}\left(\bar{\psi}_T - \mathbb{E}_{\pi_\eta}[X]\right)^2 \leq c\left(\eta + \frac{1}{\eta(T-\mathscr{B})}\right) \leq c' T^{-\min(\alpha, 1-\alpha)} \tag{3.34}$$

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 3.13. *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 (3.34) and Lemma 3.13, we arrive at the lower bound:

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

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:

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

Taking the limit, we have:

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

which completes the proof of this theorem.

Proof (Proof of Lemma 3.13). 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:

$$\begin{aligned} e^{-F_\eta(x)} - e^{-F_\eta(-x)} &= e^{-x^2}\left(e^{1/2\int_0^x \delta_\eta(z)dz} - e^{1/4\int_0^x \delta_\eta(z)dz}\right) \geq \frac{e^{-4}}{4}\int_0^x \delta_\eta(z)dz \\ &\geq \frac{e^{-4}}{4}\int_{1/2}^1 \frac{z}{\log\left(e + (z\sqrt{\eta})^{-1}\right)} dz \geq \frac{e^{-4}}{8} \cdot \frac{1}{\log\left(e + \frac{2}{\sqrt{\eta}}\right)} \end{aligned}$$

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 $c > 0$.

3.6 Additional related works

Gradient descent and stochastic gradient descent methods have gained unprecedented popularity in the past decade amidst the era of big data Bubeck (2015); Bottou et al. (2018), driven by the rapid growth of deep learning applications Goodfellow et al. (2016). 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 Le Roux et al. (2012); Johnson & Zhang (2013); Defazio et al. (2014), momentum-accelerated methods Nesterov (1983); Beck & Teboulle (2009), second-order methods Dennis & Moré (1974); Nesterov & Polyak (2006), adaptive gradient methods Duchi et al. (2011); Kingma & Ba (2015), iteration averaging Ruppert (1988); Polyak & Juditsky (1992), and coordinate descent Wright (2015), among others. The Polyak-Ruppert iteration averaging method Polyak & Juditsky (1992); Polyak (1990); Ruppert (1988) and its generalized form **?** have been shown to enhance robustness with respect to step size selection, achieving asymptotic normality with optimal covariance matching local minimax optimality Zhu et al. (2016); Duchi & Ruan (2021). Recent studies have further explored the non-asymptotic behavior of stochastic gradient descent with iteration averaging **??**Bach & Moulines (2013); Bach (2014); Flammarion & Bach (2015); Gadat & Panloup (2017); Dieuleveut et al. (2017, 2020). In the studies of linear regression and stochastic approximation, Zhang (2004); Jain et al. (2017, 2018) have analyzed the "tail-averaging" technique, achieving exponential forgetting and optimal statistical risk simultaneously. Lakshminarayanan & Szepesvari (2018) 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, Mou et al. (2020) establishes Gaussian limit and concentration inequalities for constant stepsize algorithms, with related advancements discussed in **?**.

The weak convergence result from Polyak & Juditsky (1992) has recently been generalized to functional weak convergence by **?** and **?** 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 **???????**. Referenced works beyond this overview delve deeper into topics such as asymptotic normality, statistical inference using gradient-based methods, and variants thereof **??**Li et al. (2018); Liang & Su (2019); **?**); Karimi et al. (2019); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**); **?**.

The asymptotic efficiency of variance-reduced stochastic approximation methods has been relatively underexplored in research. Frostig et al. (2015) introduces an online variant of the SVRG algorithm Johnson & Zhang (2013) and establishes a non-asymptotic upper bound on excess risk, aligning its leading term with opti-

mal asymptotics under specific self-concordant conditions on the objective function. Arnold et al. (2019) proposes *Implicit Gradient Transportation* (IGT) to reduce algorithmic variance. In the context of reinforcement learning for policy evaluation, Khamaru et al. (2020); **?** 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, **?** introduces the ROOT-SGD algorithm that achieves local minimax optimality. This algorithm can be viewed as an online variant of SARAH Nguyen et al. (2017) and connects with extrapolation-smoothing methods like (N)IGT and STORM Arnold et al. (2019); Cutkosky & Orabona (2019); Cutkosky & Mehta (2020). In a different approach, Nesterov (2009); Xiao (2010); Lee et al. (2012) propose dual averaging for the regularized or proximal case.[7] ROOT-SGD distinguishes itself by averaging past stochastic gradients with proper de-bias corrections, achieving both statistical efficiency and non-asymptotic high-order terms.

3.7 Discussion

In this chapter, we conduct a non-asymptotic, two-time-scale analysis of the ROOT-SGD algorithm proposed by **?** with a diminishing stepsize sequence, establishing its fine-grained optimality under different regimes. We demonstrate that the algorithm converges to the optimal normal limit under minimal smoothness assumptions. In contrast, the Polyak-Ruppert averaged SGD is found to be sub-optimal under these assumptions in a presented example. Additionally, we derive non-asymptotic upper bounds on gradient norm, estimation error, and excess risk for ROOT-SGD, achieving a leading term that precisely matches the asymptotic risk under the limiting Gaussian law, alongside high-order terms showing sharp dependencies on problem-specific parameters. Moreover, with a one-point Hessian Lipschitz condition imposed, these additional terms decay at a rate of $O(N^{-3/2})$, achieving optimality without prior knowledge of the sample size N. Our analysis potentially extends to non-strongly convex, non-convex, and stochastic approximation problems with varying geometric properties, indicating critical avenues for future research. Finally, exploring applications to Markovian or distributed data settings remains an important direction for further study.

3.8 Proof of auxiliary lemmas

For the proofs of auxiliary lemmas, we first describe a simple decomposition result for the process $(z_t)_{t\geq\mathscr{B}}$ which plays a central role in our analysis.

[7] See also Duchi & Ruan (2021); Tripuraneni et al. (2018) for manifold first-order optimization methods.

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 (3.5) of the updates, for any iterate $t \geq \mathscr{B}$, 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

$$z_t = \underbrace{\frac{1}{t}\sum_{s=\mathscr{B}}^{t} \varepsilon_s(\theta_{s-1})}_{:=M_t} + \frac{\mathscr{B}}{t} z_{\mathscr{B}} + \underbrace{\frac{1}{t}\sum_{s=\mathscr{B}}^{t} (s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}))}_{:=\Psi_t} \tag{3.35}$$

It can be seen that both both of the sequences $\{M_t\}_{t\geq\mathscr{B}}$ and $\{\Psi_t\}_{t\geq\mathscr{B}}$ are martingales adapted to the filtration $(\mathscr{F}_t)_{t\geq\mathscr{B}}$. We make use of this martingale decomposition throughout our analysis.

3.8.1 Proof of Lemma 3.6

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:

$$\begin{aligned} \mathbb{E}\|v_t\|_2^2 = \left(1 - \frac{1}{t}\right)^2 \underbrace{\mathbb{E}\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2}_{I_1} + \frac{1}{t^2}\underbrace{\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2}_{I_2} \\ \underbrace{+ 2\frac{t-1}{t^2}\mathbb{E}\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t), \nabla f(\theta_{t-1};\xi_t)\rangle}_{I_3} \end{aligned}$$

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:

$$\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right)$$

$$= \|v_{t-1}\|_2^2 + 2\mathbb{E}\left(\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle \mid \mathscr{F}_{t-1}\right) + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right)$$
$$= \|v_{t-1}\|_2^2 - \frac{2}{\eta_{t-1}}\langle \theta_{t-1} - \theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathscr{F}_{t-1}\right)$$

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:

$$\begin{aligned} I_1 &\leq \mathbb{E}\|v_{t-1}\|_2^2 - \frac{2}{\eta_{t-1}}\mathbb{E}\left(\frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}\right) \\ &\quad + 2\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 + 2\mathbb{E}\left(\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)\|_2^2\right) \\ &\leq (1 - \eta_{t-1}\mu + 2\eta_{t-1}^2\ell_\Xi^2)\mathbb{E}\|v_{t-1}\|_2^2 + 2\left(1 - \frac{1}{\eta_{t-1}(\mu + L)}\right)\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\ &\leq \left(1 - \frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\|v_{t-1}\|_2^2 \end{aligned}$$

Now we study the second term, note that

$$\begin{aligned} \mathbb{E}\|\nabla f(\theta_{t-1};\xi_t)\|_2^2 &\leq 2\mathbb{E}\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)\|_2^2 + 2\mathbb{E}\|\nabla f(\theta^*;\xi_t)\|_2^2 \\ &\leq 4\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 4\mathbb{E}\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta^*,\xi_t)\|_2^2 + 2\mathbb{E}\|\nabla f(\theta^*;\xi_t)\|_2^2 \\ &\leq 4\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 4\ell_\Xi^2\mathbb{E}\|\theta_{t-1} - \theta^*\|_2^2 + 2\sigma_*^2 \\ &\leq 4\left(1 + \frac{\ell_\Xi^2}{\mu^2}\right)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2 \end{aligned}$$

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:

$$T_1 \leq \|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2 + \|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2 \leq (1 + \eta_{t-1}L)\|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2$$

For the term T_2, we have:

$$\begin{aligned}
T_2 &\leq \mathbb{E}\left(\|\varepsilon(\theta_{t-1},\xi_t)-\varepsilon(\theta_{t-2},\xi_t)\|_2\cdot\|\varepsilon(\theta_{t-1},\xi_t)\|_2 \mid \mathscr{F}_{t-1}\right)\\
&\leq \sqrt{\mathbb{E}(\|\varepsilon(\theta_{t-1},\xi_t)-\varepsilon(\theta_{t-2},\xi_t)\|_2^2 \mid \mathscr{F}_{t-1})\cdot\mathbb{E}(\|\varepsilon(\theta_{t-1},\xi_t)\|_2^2 \mid \mathscr{F}_{t-1})}\\
&\leq \ell_\Xi^2\eta_{t-1}\|v_{t-1}\|_2\cdot\|\theta_{t-1}-\theta^*\|_2\\
&\leq \frac{\ell_\Xi^2}{\mu}\eta_{t-1}\|v_{t-1}\|_2\cdot\|\nabla F(\theta_{t-1})\|_2
\end{aligned}$$

So we have:

$$\begin{aligned}
I_3 &\leq 3\mathbb{E}\left(\|v_{t-1}\|_2\cdot\|\nabla F(\theta_{t-1})\|_2\right) \leq 3\sqrt{\mathbb{E}\|v_{t-1}\|_2^2\cdot\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2}\\
&\leq \frac{t\eta_{t-1}\mu}{8}\mathbb{E}\|v_{t-1}\|_2^2+\frac{18}{t\mu\eta_{t-1}}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2
\end{aligned}$$

Putting above estimates together, we obtain:

$$\begin{aligned}
\mathbb{E}\|v_t\|_2^2 &\leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\|v_{t-1}\|_2^2+\frac{1}{t^2}\left(2\sigma_*^2+4\big(1+\frac{\ell_\Xi^2}{\mu^2}\big)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right)\\
&\quad+\frac{(t-1)\eta_{t-1}\mu}{8t}\mathbb{E}\|v_{t-1}\|_2^2+\frac{18}{t^2\mu\eta_{t-1}}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\\
&\leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta_{t-1}\mu}{4}\right)\mathbb{E}\|v_{t-1}\|_2^2+\frac{26}{t^2\mu\eta_{t-1}}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2+\frac{2\sigma_*^2}{t^2}
\end{aligned}$$

which finishes the proof.

3.8.2 Proof of Lemma 3.7

Taking the squared norm of z_t in the martingale decomposition (3.35) and applying the triangle inequality yields

$$\mathbb{E}\|z_t\|_2^2 \leq \frac{2}{t^2}\mathbb{E}\|M_t\|_2^2+\frac{\mathscr{B}^2}{t^2}\|z_0\|_2^2+\frac{2}{t^2}\mathbb{E}\|\Psi_t\|_2^2$$

For the martingale M_t, we have:

$$\mathbb{E}\|M_t\|_2^2=\sum_{s=1}^t\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^2\leq 2t\sigma_*^2+2\ell_\Xi^2\sum_{s=1}^t\mathbb{E}\|\theta_{s-1}-\theta^*\|_2^2$$

For the martingale Ψ_t, we have:

$$\mathbb{E}\|\Psi_t\|_2^2=\sum_{s=1}^t(s-1)^2\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2\leq\ell_\Xi^2\sum_{s=1}^t(s-1)^2\eta_{s-1}^2\mathbb{E}\|v_{s-1}\|_2^2$$

Combining the pieces yields

$$\mathbb{E}\|z_t\|_2^2 \le \frac{\mathscr{B}^2\|z_0\|_2^2}{t^2} + \frac{4\sigma_*^2}{t} + \frac{4\ell_\Xi^2}{t^2}\sum_{s=1}^{t}\mathbb{E}\|\theta_{s-1}-\theta^*\|_2^2 + \frac{\ell_\Xi^2}{t^2}\sum_{s=1}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\|v_{s-1}\|_2^2$$

Note that the μ-strong convexity condition (cf. Assumption 12) 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.

3.8.3 Proof of Lemma 3.8

Denote $\ell_t := \sum_{s=\mathscr{B}}^{t}\eta_s$, which is the aggregated step sizes up to time t.

Recursively applying the inequality (3.19b), and noting that H_t is a non-decreasing sequence and that η_t is non-increasing, we obtain:

$$\begin{aligned} W_T &\le 2\sigma_*^2\sum_{t=\mathscr{B}}^{T-1}e^{-\mu(\ell_T-\ell_t)} + 2CH_{T-1}\sum_{t=\mathscr{B}}^{T-1}\frac{e^{-\mu(\ell_T-\ell_t)}}{t\mu\eta_{t-1}} + \sum_{t=\mathscr{B}}^{T-1}e^{-\mu(\ell_T-\ell_{\mathscr{B}})}W_{\mathscr{B}} \\ &\le \frac{2\sigma_*^2}{\eta_T\mu} + \frac{CH_{T-1}}{T(\mu\eta_{T-1})^2} + e^{-\mu(\ell_T-\ell_{\mathscr{B}})}\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2 \end{aligned}$$

Substituting the bound into Eq (3.19a), we obtain:

$$\begin{aligned} H_T \le{}& 4\sigma_*^2 + 2\mathbb{E}\|z_{\mathscr{B}}\|_2^2\mathscr{B} + C'\ell_\Xi^2\frac{2\sigma_*^2}{\mu}\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\eta_s + 2CC'\ell_\Xi^2 H_T\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\frac{1}{s\mu^2} \\ &+ C'\ell_\Xi^2\mathscr{B}^2\mathbb{E}\|v_{\mathscr{B}}\|_2^2\cdot\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}e^{-\mu(\ell_s-\ell_{\mathscr{B}})}\eta_{s-1}^2 \end{aligned}$$

For the quantities involving step size sequences in the inequality above, we have:

$$\begin{aligned} \sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\eta_s &\le \sup_{\mathscr{B}\le t\le T}\frac{1}{t-\mathscr{B}+1}\sum_{s=\mathscr{B}}^{t-1}\eta_s \le \eta_{\mathscr{B}} \\ \sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}e^{-\mu(\ell_s-\ell_{\mathscr{B}})}\eta_{s-1}^2 &\le \frac{1}{\mathscr{B}}\sum_{s=\mathscr{B}}^{T-1}e^{-\mu(\ell_s-\ell_{\mathscr{B}})}\eta_{s-1}^2 \le \frac{\eta_{\mathscr{B}}}{\mathscr{B}\mu} \end{aligned}$$

For $\mathscr{B} > \frac{4CC'\ell_\Xi^2}{\mu^2}$, we have $2CC'\ell_\Xi^2\sup_{\mathscr{B}\le t\le T}\frac{1}{t}\sum_{s=\mathscr{B}}^{t-1}\frac{1}{s\mu^2} \le \frac{1}{2}$, and consequently:

$$H_T \le c\left(\sigma_*^2 + \frac{\ell_\Xi^2\mathscr{B}\eta_{\mathscr{B}}}{\mu}W_{\mathscr{B}} + H_{\mathscr{B}}\right)$$

for universal constants $c > 0$.

Substituting back into the bound (3.19b), for $T \ge \mathscr{B} \ge (\mu\eta_T)^{-1}$, we obtain:

$$W_T = T^2\mathbb{E}\|v_T\|_2^2 \le \frac{c'}{\eta_T\mu}\sigma_*^2 + c'\left(\frac{\mathscr{B}}{T\mu^2\eta_{T-1}^2} + e^{-\mu(\ell_T-\ell_{\mathscr{B}})}\mathscr{B}^2\right)W_{\mathscr{B}}$$

3.8.4 Proof of Lemma 3.9

By the martingale decomposition (3.35), for any $t \geq \mathscr{B}$, we have the identity

$$t^2\mathbb{E}\left\|Gz_t\right\|_2^2 = \mathscr{B}^2\mathbb{E}\left\|Gz_{\mathscr{B}}\right\|_2^2 + \mathbb{E}\left([GM]_t\right) + \mathbb{E}\left([G\Psi]_t\right) + 2\mathbb{E}\left([GM, G\Psi]_t\right) \quad (3.36)$$

For the quadratic variation terms, we note that

$$\begin{aligned}
\mathbb{E}\left([GM]_t\right) &= \sum_{s=\mathscr{B}+1}^{t} \mathbb{E}\left\|G\varepsilon_s(\theta_{s-1})\right\|_2^2 \\
&\leq \sum_{s=\mathscr{B}+1}^{t} \left(\sqrt{\mathbb{E}\left\|G\varepsilon_s(\theta^*)\right\|_2^2} + \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} \sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\right\|_2^2}\right)^2 \\
&\leq \sum_{s=\mathscr{B}+1}^{t} \left(\sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)} + \ell_\Xi \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} \sqrt{\mathbb{E}\left\|\theta_{s-1} - \theta^*\right\|_2^2}\right)^2 \\
&\leq (t-\mathscr{B})\mathrm{Tr}\left(G\Sigma^* G^\top\right) + 2\sum_{s=\mathscr{B}+1}^{t} \sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}\ell_\Xi \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}} r_\theta(s) + \sum_{s=\mathscr{B}+1}^{t} \ell_\Xi^2 \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}}^2 r_\theta^2(s) \\
&\leq (t-\mathscr{B})\mathrm{Tr}\left(G\Sigma^* G^\top\right) + \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}+1}^{t} \left(2\sigma_* \ell_\Xi r_\theta(s) + \ell_\Xi^2 r_\theta^2(s)\right) \quad (3.37)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left([G\Psi]_t\right) &= \sum_{s=\mathscr{B}+1}^{t} (s-1)^2 \mathbb{E}\left\|G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\right\|_2^2 \\
&\leq \ell_\Xi^2 \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}+1}^{t} (s-1)^2 \mathbb{E}\left\|\theta_{s-1} - \theta_{s-2}\right\|_2^2 \\
&\leq \ell_\Xi^2 \left|\!\left|\!\left|G\right|\!\right|\!\right|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}+1}^{t} (s-1)^2 \eta_{s-1}^2 r_v^2(s) \quad (3.38)
\end{aligned}$$

We decompose the cross variation term in two parts, and bound them separately.

$$\begin{aligned}
\mathbb{E}\left([GM, G\Psi]_t\right) &= \sum_{s=\mathscr{B}+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle \\
&= \underbrace{\sum_{s=\mathscr{B}+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_1(t)} \\
&\quad + \underbrace{\sum_{s=\mathscr{B}+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_2(t)}
\end{aligned}$$

For the term Q_1, Cauchy–Schwartz inequality leads to the bound:

$$
\begin{aligned}
Q_1(t) &\leq \sum_{s=\mathscr{B}+1}^{t} (s-1)\|\!|G|\!\|_{\mathrm{op}}^2 \sqrt{\mathbb{E}\,\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta^*)\|_2^2 \cdot \mathbb{E}\,\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2} \\
&\leq \ell_\Xi^2 \|\!|G|\!\|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}+1}^{t} (s-1)\eta_{s-1}\sqrt{\mathbb{E}\,\|\theta_{s-1}-\theta^*\|_2^2 \cdot \mathbb{E}\,\|v_{s-1}\|_2^2} \\
&\leq \ell_\Xi^2 \|\!|G|\!\|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}+1}^{t} (s-1)\eta_{s-1} r_v(s) r_\theta(s) \qquad (3.39)
\end{aligned}
$$

For the term Q_2, we note that

$$
\begin{aligned}
Q_2(t) &= \sum_{s=\mathscr{B}+1}^{t} (s-1)\left(\mathbb{E}\langle G\varepsilon_s(\theta^*), G\varepsilon_s(\theta_{s-1})\rangle - \mathbb{E}\langle G\varepsilon_{s-1}(\theta^*), G\varepsilon_{s-1}(\theta_{s-2})\rangle\right) \\
&\overset{(i)}{=} (\mathscr{B}-1)\mathbb{E}\langle G\varepsilon_t(\theta^*), G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{\mathscr{B}-1})\rangle + \sum_{s=\mathscr{B}}^{t-1} \mathbb{E}\langle G\varepsilon_t(\theta^*), G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{s-1})\rangle \\
&\overset{(ii)}{\leq} (\mathscr{B}-1)\sigma_*\ell_\Xi \|\!|G|\!\|_{\mathrm{op}}^2 \sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta_{\mathscr{B}-1}\|_2^2} + \sigma_*\ell_\Xi \|\!|G|\!\|_{\mathrm{op}}^2 \sum_{s=\mathscr{B}}^{t-1} \sqrt{\mathbb{E}\,\|\theta_{t-1}-\theta_{s-1}\|_2^2} \\
&\leq 2\sigma_*\ell_\Xi \|\!|G|\!\|_{\mathrm{op}}^2 \left(\mathscr{B}\,\|\theta_0-\theta^*\|_2 + t r_\theta(t) + \sum_{s=\mathscr{B}}^{t-1} r_\theta(s)\right) \qquad (3.40)
\end{aligned}
$$

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:

$$
\mathbb{E}\,\|Gz_{\mathscr{B}}\|_2^2 \leq \|\!|G|\!\|_{\mathrm{op}}^2 \cdot \mathbb{E}\,\|z_{\mathscr{B}}\|_2^2 \leq \|\!|G|\!\|_{\mathrm{op}}^2 \cdot \frac{2(\sigma_*^2 + \ell_\Xi^2\,\|\theta_0-\theta^*\|_2^2)}{\mathscr{B}} \qquad (3.41)
$$

Collecting the bounds (3.37)-(3.41) and substituting into the decomposition (3.36), we obtain the inequality:

$$
\begin{aligned}
&\mathbb{E}\,\|Gz_T\|_2^2 \leq \left(1+\frac{\mathscr{B}}{T}\right)\cdot\frac{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}{T} + c\frac{\|\!|G|\!\|_{\mathrm{op}}^2\sigma_*\ell_\Xi}{T^2}\sum_{s=\mathscr{B}}^{T} r_\theta(s) \\
&+ c\frac{\|\!|G|\!\|_{\mathrm{op}}^2\ell_\Xi^2}{T^2}\sum_{s=\mathscr{B}}^{T}\left(r_\theta(s)+(s-1)\eta_{s-1}r_v(s)\right)^2 + c\frac{\mathscr{B}\|\!|G|\!\|_{\mathrm{op}}^2\left(\sigma_*+\ell_\Xi\,\|\theta_0-\theta^*\|_2\right)^2}{T^2}
\end{aligned}
$$

for a universal constant $c>0$.

Invoking Proposition 3.1, we note that:

$$
r_\theta(t) \leq c\frac{\sigma_*}{\mu\sqrt{t}} + \frac{\sqrt{\mathscr{B}\log t}}{\mu^2 t}\left(\ell_\Xi + \frac{1}{\eta_t\sqrt{t}}\right)\|\nabla F(\theta_0)\|_2 \qquad \text{and} \quad r_v(t) \leq c\frac{\sigma_*}{t\sqrt{\mu\eta_t}} + \frac{\sqrt{\mathscr{B}}}{\mu\eta_t t^{3/2}}\|\nabla F(\theta_0)\|_2
$$

Substituting into above upper bound, we obtain:

$$\mathbb{E}\,\|Gz_T\|_2^2 \le \frac{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}{T} + c|||G|||_{\mathrm{op}}^2\left(\frac{\ell_\Xi}{\mu T^{3/2}} + \frac{\ell_\Xi^2\sum_{s=\mathscr{B}}^T \eta_s}{\mu T^2} + \frac{\mathscr{B}}{T^2}\right)\sigma_*^2 + c|||G|||_{\mathrm{op}}^2\frac{\ell_\Xi^2\mathscr{B}\log T}{\mu^2 T^2}\left(1 + \frac{1}{\mu\ell_\Xi}\sum_{s=\mathscr{B}}^T \frac{1}{\eta_s^2 s^{5/2}}\right)\|\nabla F(\theta_0)\|_2^2$$

For the stepsize choice $\eta_t = \frac{1}{\mu\mathscr{B}^{1-\alpha}t^\alpha}$, we have the bound

$$\mathbb{E}\,\|Gz_T\|_2^2 \le \frac{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}{T} + c|||G|||_{\mathrm{op}}^2\left(\frac{\mathscr{B}}{T}\right)^{1/2\wedge\alpha}\frac{\sigma_*^2}{T} + c|||G|||_{\mathrm{op}}^2\frac{\mathscr{B}^2\log T}{T^2}\left(1 + \frac{T^{2\alpha-3/2}}{\mathscr{B}^{2\alpha-3/2}}\right)\|\nabla F(\theta_0)\|_2^2$$

which proves this lemma.

3.8.5 Proof of Lemma 3.10

Similar to the proof of Lemma 3.6, we use the decomposition

$$\begin{aligned}
\mathbb{E}\,\|v_t\|_2^4 &\le \left(1-\frac{1}{t}\right)^4\mathbb{E}\,\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^4 \\
&+ \frac{4}{t}\left(1-\frac{1}{t}\right)^3\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^2\,\langle v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t), \nabla f(\theta_{t-1},\xi_t)\rangle\right) \\
&+ \frac{6}{t^2}\left(1-\frac{1}{t}\right)^2\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right) \\
&+ \frac{4}{t^3}\left(1-\frac{1}{t}\right)\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right) \\
&+ \frac{1}{t^4}\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \qquad (3.42)
\end{aligned}$$

We claim the following bounds on the relevant terms in Eq (3.42), for stepsize choice $\eta_{t-1} \le \frac{1}{8}\left(\frac{1}{L}\wedge\frac{\mu}{\widetilde{\ell_\Xi}^2}\right)$

$$\mathbb{E}\,\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^4 \le (1-\mu\eta_{t-1})\mathbb{E}\,\|v_{t-1}\|_2^4 \qquad (3.43a)$$

and

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^2\,\langle v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t), \nabla f(\theta_{t-1},\xi_t)\rangle\right) \\
&\le \frac{t\mu\eta_{t-1}}{3}\mathbb{E}\,\|v_{t-1}\|_2^4 + \frac{c}{t}\left(\widetilde{\sigma}_*^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \qquad (3.43b)
\end{aligned}$$

Recall that Eq (3.24) implies the bound

$$\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \le 27\widetilde{\sigma_*}^4 + \frac{27}{(\mu\eta_{t-1})^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 \tag{3.43c}$$

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:

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right)\\
&\le\left(\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/2}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{1/2}\\
&\le c\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right)\\
&\le\left(\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/4}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{3/4}\\
&\le ct\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)+\frac{c}{t}\left(\widetilde{\sigma_*}^4+\frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)
\end{aligned}$$

Collecting above bounds, we arrive at the conclusion

$$\begin{aligned}
\mathbb{E}\,\|v_t\|_2^4 &\le\left(1-\frac{1}{t}\right)^4(1-\mu\eta_{t-1})\mathbb{E}\,\|v_{t-1}\|_2^4+\frac{c_1}{t^2}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\\
&\quad+\frac{c_2}{t^4}\left(\widetilde{\sigma_*}^4+\frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)\\
&\le\left[\left(1-\frac{1}{t}\right)^2\left(1-\frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\,\|v_{t-1}\|_2^4}+\frac{c'}{t^2}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4}\right)\right]^2
\end{aligned}$$

for universal constants $c_1,c_2,c'>0$. This completes the proof of this lemma.

Proof of Eq (3.43a):

We note the following expansion:

$$\begin{aligned}
&\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\\
&\le\mathbb{E}\,\|v_{t-1}\|_2^4+4\mathbb{E}\left(\|v_{t-1}\|_2^2\langle v_{t-1},\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2})\rangle\right)+6\mathbb{E}\left(\|v_{t-1}\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2\right)\\
&\quad+4\mathbb{E}\left(\|v_{t-1}\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^3\right)+\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\\
&\le\left(1-4\eta_{t-1}\frac{\mu L}{\mu+L}+6\eta_{t-1}^2\widetilde{\ell_\Xi}^2\right)\mathbb{E}\,\|v_{t-1}\|_2^4+\left(8-\frac{4}{(\mu+L)\eta_{t-1}}\right)\mathbb{E}\left(\|\nabla F(\theta_{t-1})-\nabla F(\theta_{t-2}\|_2^2\cdot\|v_{t-1}\|_2^2\right)
\end{aligned}$$

$$+ 3\mathbb{E}\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^4$$

For the last term, we note that

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^4 \\
&\le 8\mathbb{E}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^4 + 8\mathbb{E}\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)\|_2^4 \\
&\le 8L^2\eta_{t-1}^2\mathbb{E}\left(\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \cdot \|v_{t-1}\|_2^2\right) + 8\widetilde{\ell_\Xi}^4\eta_{t-1}^4\mathbb{E}\|v_{t-1}\|_2^4
\end{aligned}$$

Putting them together, for $\eta_{t-1} \le \frac{1}{8}\left(\frac{1}{L} \wedge \frac{\mu}{\widetilde{\ell_\Xi}^2}\right)$, we arrive at the contraction bound

$$\begin{aligned}
&\mathbb{E}\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\|_2^4 \\
&\le \left(1 - 4\eta_{t-1}\frac{\mu L}{\mu + L} + 6\eta_{t-1}^2\widetilde{\ell_\Xi}^2 + 24\eta_{t-1}^4\widetilde{\ell_\Xi}^4\right)\mathbb{E}\|v_{t-1}\|_2^4 \\
&\qquad + \left(8 - \frac{4}{(L+\mu)\eta_{t-1}} + 24L^2\eta_{t-1}^2\right)\mathbb{E}\left(\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}\|_2^2 \cdot \|v_{t-1}\|_2^2\right) \\
&\le (1 - \mu\eta_{t-1})\mathbb{E}\|v_{t-1}\|_2^4
\end{aligned}$$

which proves this bound.

Proof of Eq (3.43b):

Denote the following random variables for notational convenience

$$\lambda_{t-1} := v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) \qquad \text{and} \quad \zeta_t := \varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)$$

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 13′ that

$$\mathbb{E}\left(\|\zeta_t\|_2^4 \mid \mathscr{F}_{t-1}\right) \le \widetilde{\ell_\Xi}^4\|\theta_{t-1} - \theta_{t-2}\|_2^4 = \widetilde{\ell_\Xi}^4\eta_{t-1}^4\|v_{t-1}\|_2^4$$

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 (3.24) accompanied with Hölder's inequality, we can bound the above terms as follows

$$\mathbb{E}\left(\|\lambda_{t-1}\|_2^2\langle\lambda_{t-1}, \nabla F(\theta_{t-1})\rangle\right) \le \left(\mathbb{E}\|v_{t-1}\|_2^4\right)^{3/4} \cdot \left(\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}$$

$$\mathbb{E}\left(\|\zeta_t\|_2\,\|\lambda_{t-1}\|_2^2\,\|\nabla f(\theta_{t-1},\xi_t)\|_2\right) \le 3\widetilde{\ell_\Xi}\,\eta_{t-1}\mathbb{E}\left(\|v_{t-1}\|_2^3\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\,\|\nabla F(\theta_{t-1})\|_2\right)\right)$$
$$\le 3\widetilde{\ell_\Xi}\,\eta_{t-1}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)$$

and

$$\mathbb{E}\left(\|\zeta_t\|_2^3\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2\right) \le \left(\mathbb{E}\,\|\zeta_t\|_2^4\right)^{3/4}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{1/4}$$
$$\le 3\widetilde{\ell_\Xi}^3\,\eta_{t-1}^3\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)$$

Collecting the three terms, and noting that $\eta_{t-1}\le\left(\frac{1}{L}\wedge\frac{\mu}{\widetilde{\ell_\Xi}^2}\right)\le\frac{1}{\widetilde{\ell_\Xi}}\sqrt{\frac{\mu}{L}}\le\frac{1}{\widetilde{\ell_\Xi}}$, we have

$$\begin{aligned}
&\mathbb{E}\left(\|\lambda_{t-1}+\zeta_t\|_2^2\,\langle\lambda_{t-1}+\zeta_t,\nabla f(\theta_{t-1},\xi_t)\rangle\right)\\
&\le c\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}+\widetilde{\ell_\Xi}\,\eta_{t-1}\widetilde{\sigma_*}\right)\\
&\le\frac{t\mu\eta_{t-1}}{3}\mathbb{E}\,\|v_{t-1}\|_2^4+\frac{c}{t}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}
\end{aligned}$$

which proves this inequality.

3.8.6 Proof of Lemma 3.11

By Eq (3.35) and Minkowski's inequality, we have the bound

$$\mathbb{E}\,\|z_t\|_2^4\le\frac{\mathscr{B}^4}{t^4}\mathbb{E}\,\|z_{\mathscr{B}}\|_2^4+\frac{8}{t^4}\mathbb{E}\,\|M_t\|_2^4+\frac{8}{t^4}\mathbb{E}\,\|\Psi_t\|_2^4$$

Invoking the BDG inequality for Hilbert-space-valued martingales, we have the moment bound

$$\mathbb{E}\,\|M_t\|_2^4\le c\mathbb{E}\left([M]_t^2\right)=c\cdot\mathbb{E}\left(\sum_{s=\mathscr{B}+1}^{t}\|\varepsilon_s(\theta_{s-1})\|_2^2\right)^2\qquad\text{and}$$
$$\mathbb{E}\,\|\Psi_t\|_2^4\le c\mathbb{E}\left([\Psi]_t^2\right)\le c\cdot\mathbb{E}\left(\sum_{s=\mathscr{B}+1}^{t}(s-1)^2\,\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^2\right)^2$$

Invoking Cauchy–Schwartz inequality, we note that

$$\begin{aligned}\mathbb{E}\|M_t\|_2^4 &\le c\sum_{s=\mathscr{B}+1}^{t}\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4 + 2c\sum_{\mathscr{B}+1\le s\le u\le t}\mathbb{E}\left(\|\varepsilon_s(\theta_{s-1})\|_2^2\cdot\mathbb{E}\|\varepsilon_u(\theta_{u-1})\|_2^4\right)\\ &\le c\sum_{s=\mathscr{B}+1}^{t}\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4 + 2c\sum_{\mathscr{B}+1\le s\le u\le t}\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4}\cdot\sqrt{\mathbb{E}\|\varepsilon_u(\theta_{u-1})\|_2^4}\\ &= c\left(\sum_{s=\mathscr{B}+1}^{t}\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4}\right)^2\end{aligned}$$

Similarly, for the martingale $(\Psi_t)_{t\ge\mathscr{B}}$, we have the bound

$$\sqrt{\mathbb{E}\|\Psi_t\|_2^4} \le c\sum_{s=\mathscr{B}+1}^{t}(s-1)^2\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^4}$$

By Eq (3.24), we have the bound

$$\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})\|_2^4} \le c\left(\widetilde{\sigma_*}^2 + \frac{\widetilde{\ell_\Xi}^2}{\mu^2}\sqrt{\mathbb{E}\|\nabla F(\theta_{s-1})\|_2^4}\right)$$

By Assumption 13′, we note that

$$\sqrt{\mathbb{E}\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\|_2^4} \le \widetilde{\ell_\Xi}^2\sqrt{\mathbb{E}\|\theta_{s-1}-\theta_{s-2}\|_2^4} = \widetilde{\ell_\Xi}^2\eta_{t-1}^2\sqrt{\mathbb{E}\|v_{s-1}\|_2^4}$$

Collecting the terms above, we arrive at the conclusion.

3.8.7 Proof of Lemma 3.12

We first note the following decomposition, which holds true for any $\widetilde{T}\in[0,t-\mathscr{B}]$

$$|\mathbb{E}\langle tGz_t, Gv_t\rangle| \le \underbrace{(t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}}, Gv_t\rangle\right|}_{:=Q_3\left(t,\widetilde{T}\right)} + \underbrace{\left|\mathbb{E}\langle G\left(tz_t-(t-\widetilde{T})z_{t-\widetilde{T}}\right), Gv_t\rangle\right|}_{:=Q_4\left(t,\widetilde{T}\right)}.$$

We claim the following upper bounds for the terms $Q_3(t,\widetilde{T})$ and $Q_4(t,\widetilde{T})$, for $\widetilde{T}\in\left[c\mathscr{B}^{1-\alpha}t^\alpha\log t, t/2\right]$:

$$Q_3(t,\widetilde{T}) \le c\frac{\|G\|_{\mathrm{op}}^2 L_2}{\mu^2}t^{\frac{1+\alpha}{2}}\mathscr{B}^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{\mathscr{B}}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right) \tag{3.44a}$$

$$Q_4(t,\widetilde{T}) \le c\|G\|_{\mathrm{op}}^2\sqrt{\widetilde{T}\mathscr{B}^{1-\alpha}t^\alpha}\cdot\left(\frac{\sigma_*^2}{t} + \left(\frac{\mathscr{B}}{t}\right)^{2-\alpha}\|\nabla F(\theta_0)\|_2^2\right) \tag{3.44b}$$

Taking these two bounds as given, we choose the time-lag parameter $\widetilde{T} := c\mathscr{B}^{1-\alpha}t^{\alpha}\log t$, and arrive at the bound:

$$|\mathbb{E}\langle Gz_t, Gv_t\rangle| \leq c|||G|||_{\text{op}}^2\left(\frac{\mathscr{B}}{t}\right)^{1-\alpha}\left(\frac{\sigma_*^2}{t}+(\frac{\mathscr{B}}{t})^{2-\alpha}\|\nabla F(\theta_0)\|_2^2\right)\log t$$
$$+c\frac{|||G|||_{\text{op}}^2 L_2}{\mu^2}\left(\frac{\mathscr{B}}{t}\right)^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}}+\left(\frac{\mathscr{B}}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right)$$

which completes the proof of this lemma.

Proof of the bound (3.44a):

To bound the term Q_3, we use the following lemma

Lemma 3.14. *For $t > \mathscr{B}$ and $s > 0$, the following bound holds true*

$$\mathbb{E}\,\|\mathbb{E}\,(v_{t+s}\mid\mathscr{F}_t)\|_2^2 \leq c\widetilde{r}_v^2(t)e^{-\mu\sum_{k=1}^{s-1}\eta_k}+c\frac{L_2^2}{\mu^2}\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t)$$

See §3.8.8 for the proof of this lemma.

Taking Lemma 3.14 as given, the bound for the term $Q_3(t,\widetilde{T})$ directly follows from Cauchy–Schwartz inequality.

$$Q_3(t,\widetilde{T}) = (t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}}, G\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\widetilde{T}}\right)\rangle\right| \leq t|||G|||_{\text{op}}^2\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\widetilde{T}}\right)\right\|_2^2}$$

For the time-lag $\widetilde{T}\leq\frac{t}{2}$, Proposition 3.1 yields the bound:

$$t\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2} \leq c\sigma_*\sqrt{t}+\mathscr{B}\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2 \tag{3.45a}$$

By Lemma 3.14, for a non-increasing stepsize sequence, when the time-lag $\widetilde{T}$ satisfies $\mu\widetilde{T}\eta_t \geq c\log t$, we have the bound $e^{-\mu\sum_{k=1}^{s-1}\eta_k}\leq\frac{1}{t^3}$. Therefore, given the stepsize choice $\eta_t=\frac{1}{\mu\mathscr{B}^{1-\alpha}t^{\alpha}}$, we have the following bound holding true for $\widetilde{T}\geq c\mathscr{B}^{1-\alpha}t^{\alpha}\log t$:

$$\mathbb{E}\left\|\mathbb{E}\left(v_t\mid\mathscr{F}_{t-\widetilde{T}}\right)\right\|_2^2 \leq cL_2\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t) \tag{3.45b}$$

Combining the bounds (3.45a) and (3.45b), we have the following bound holds true for the time-lag taking values in the interval $\widetilde{T}\in\left[c\mathscr{B}^{1-\alpha}t^{\alpha}\log t, t/2\right]$ (the interval is non-empty for any $t\geq c\mathscr{B}\log\mathscr{B}$):

$$Q_3(t,\widetilde{T}) \leq c\frac{L_2|||G|||_{\text{op}}^2}{\mu}\left(\sigma_*\sqrt{t}+\mathscr{B}\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2\right)\widetilde{r}_v(t)\cdot\widetilde{r}_\theta(t)$$

Noting that $\sigma_* \le \widetilde{\sigma_*}$ and $\ell_\Xi \le \widetilde{\ell_\Xi}$, above bounds lead to the inequality:

$$Q_3(t,\widetilde{T}) \le c\frac{L_2 \|\|G\|\|_{\text{op}}^2}{\mu^2} t^{\frac{1+\alpha}{2}} \mathscr{B}^{\frac{1-\alpha}{2}} \left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{\mathscr{B}}{t}\right)^{3-3\alpha/2} \log^2 t \, \|\nabla F(\theta_0)\|_2^3 \right)$$

which proves the desired result.

Proof of the bound (3.44b):

For the term Q_4, we also apply Cauchy-Schwartz inequality, and obtain the following bound:

$$Q_4(t,\widetilde{T}) \le \|\|G\|\|_{\text{op}}^2 \sqrt{2\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + 2\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2} \cdot \sqrt{\mathbb{E}\|v_t\|_2^2}$$

The mean-squared norms of martingales are just their expected quadratic variation:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([M]_t - [M]_{t-\widetilde{T}}\right) \le 2\widetilde{T}\sigma_*^2 + 2\sum_{s=t-\widetilde{T}+1}^{t} \ell_\Xi^2 r_\theta^2(s)$$

$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([\Psi]_t - [\Psi]_{t-\widetilde{T}}\right) \le \ell_\Xi^2 \sum_{s=t-\widetilde{T}+1}^{t} (s-1)^2 \eta_{s-1}^2 r_v^2(s)$$

Substituting with the rates in Proposition 3.1, we have the bounds:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 \le c\widetilde{T}\left(\sigma_*^2 + \frac{\ell_\Xi^2 \mathscr{B}\log t}{\mu^4 t^2}\left(\ell_\Xi^2 + \frac{1}{\eta_t^2 t}\right)\|\nabla F(\theta_0)\|_2^2\right) \qquad \text{and} \tag{3.46a}$$

$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \le c\widetilde{T}\left(\frac{\ell_\Xi^2 \sigma_*^2 \eta_t}{\mu} + \frac{\mathscr{B}^2}{t^3}\|\nabla F(\theta_0)\|_2^2\right) \tag{3.46b}$$

For the stepsize choice $\eta_t = \frac{1}{\mu \mathscr{B}^{1-\alpha} t^\alpha}$, we have the bound:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + \mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \le c\widetilde{T}\left(\sigma_*^2 + \mathscr{B}\left(\frac{\mathscr{B}}{t}\right)^{\min(2,3-2\alpha)} \|\nabla F(\theta_0)\|_2^2\right)$$

Invoking Proposition 3.1, we can bound the moment of v_t as:

$$\mathbb{E}\|v_t\|_2^2 \le c\frac{\sigma_*^2 \mathscr{B}^{1-\alpha}}{t^{2-\alpha}} + \left(\frac{\mathscr{B}}{t}\right)^{3-2\alpha} \|\nabla F(\theta_0)\|_2^2$$

Combining above bounds, we conclude that

$$Q_4(t,\widetilde{T}) \le c\|\|G\|\|_{\text{op}}^2 \sqrt{\widetilde{T}\mathscr{B}^{1-\alpha}t^\alpha} \cdot \left(\frac{\sigma_*^2}{t} + \left(\frac{\mathscr{B}}{t}\right)^{2-\alpha} \|\nabla F(\theta_0)\|_2^2\right)$$

3.8.8 Proof of Lemma 3.14

Given $t > \mathscr{B}$ fixed, denote $\Delta_s := \mathbb{E}(v_{t+s} \mid \mathscr{F}_t)$ for any $s > 0$.

Taking conditional expectations on both sides of Eq (3.5a), for $s > 0$, we have that

$$\mathbb{E}[v_{t+s} \mid \mathscr{F}_t] = \frac{t+s-1}{t+s}\mathbb{E}[v_{t+s-1} + \nabla F(\theta_{t+s-1}) - \nabla F(\theta_{t+s-2}) \mid \mathscr{F}_t] + \frac{1}{t+s}\mathbb{E}[\nabla F(\theta_{t+s-1}) \mid \mathscr{F}_t] \tag{3.47}$$

By the decomposition $\nabla F(\theta_{t+s-1}) = v_{t+s} - z_{t+s}$ and the fact that $(z_t)_{t \geq \mathscr{B}}$ is a martingale, we note that

$$\mathbb{E}[\nabla F(\theta_{t+s-1}) \mid \mathscr{F}_t] = \mathbb{E}[v_{t+s} \mid \mathscr{F}_t]$$

By the one-point Hessian Lipschitz condition, we note that

$$\begin{aligned}
&\|\nabla F(\theta_{t+s-1}) - \nabla F(\theta_{t+s-2}) + \eta_{t+s-1}H^* v_{t+s-1}\|_2 \\
&= \eta_{t+s-1}\left\| \int_0^1 \left(\nabla^2 F\big(\gamma\theta_{t+s-1} + (1-\gamma)\theta_{t+s-2}\big) - \nabla^2 F(\theta^*)\right) v_{t+s-1}\, d\gamma \right\|_2 \\
&\leq \eta_{t+s-1}L_2 \|v_{t+s-1}\|_2 \cdot \int_0^1 \|\gamma\theta_{t+s-1} + (1-\gamma)\theta_{t+s-2} - \theta^*\|_2 \, d\gamma \\
&\leq \eta_{t+s-1}L_2 \|v_{t+s-1}\|_2 \cdot (\|\theta_{t+s-1} - \theta^*\|_2 + \|\theta_{t+s-2} - \theta^*\|_2)
\end{aligned}$$

Substituting into the identity (3.47), 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.

References

Agarwal, A., Bartlett, P. L., Ravikumar, P., & Wainwright, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58), 3235–3249.

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1), 8194–8244.

Allen-Zhu, Z. (2018). How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems* (pp. 1157–1167).

Arnold, S., Manzagol, P.-A., Harikandeh, R. B., Mitliagkas, I., & Le Roux, N. (2019). Reducing the variance in online optimization by transporting past gradients. In *Advances in Neural Information Processing Systems* (pp. 5391–5402).

Asi, H. & Duchi, J. C. (2019). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3), 2257–2290.

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1), 595–627.

Bach, F. & Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems* (pp. 451–459).

Bach, F. & Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems* (pp. 773–781).

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1), 183–202.

Benveniste, A., Métivier, M., & Priouret, P. (2012). *Adaptive Algorithms and Stochastic Approximations*. Springer.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2), 163–195.

Bertsekas, D. P. & Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223–311.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4), 231–357.

Cauchy, A.-L. (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences*, 25, 536–538.

Chen, X., Lee, J. D., Tong, X. T., & Zhang, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1), 251–273.

Cutkosky, A. & Mehta, H. (2020). Momentum improves normalized SGD. In *International Conference on Machine Learning* (pp. 2260–2268).: PMLR.

Cutkosky, A. & Orabona, F. (2019). Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems* (pp. 15236–15245).

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems* (pp. 1646–1654).

Dennis, J. E. & Moré, J. J. (1974). A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of Computation*, 28(126), 549–560.

Devolder, O., Glineur, F., & Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2), 37–75.

Dieuleveut, A., Durmus, A., & Bach, F. (2020). Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3), 1348–1382.

Dieuleveut, A., Flammarion, N., & Bach, F. (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1), 3520–3570.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2121–2159.

Duchi, J. C. & Ruan, F. (2021). Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49(1), 21–48.

Fang, C., Li, C. J., Lin, Z., & Zhang, T. (2018). SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Advances in Neural Information Processing Systems* (pp. 686–696).

Flammarion, N. & Bach, F. (2015). From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory* (pp. 658–695).

Frostig, R., Ge, R., Kakade, S. M., & Sidford, A. (2015). Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory* (pp. 728–763).

Gadat, S. & Panloup, F. (2017). Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*.

Ghadimi, S. & Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4), 1469–1492.

Ghadimi, S. & Lan, G. (2013a). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4), 2061–2089.

Ghadimi, S. & Lan, G. (2013b). Stochastic first-and zeroth-order methods for non-convex stochastic programming. *SIAM Journal on Optimization*, 23(4), 2341–2368.

Ghadimi, S. & Lan, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2), 59–99.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Hazan, E. & Kale, S. (2014). Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1), 2489–2512.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., & Sidford, A. (2018). Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory* (pp. 545–604).

Jain, P., Netrapalli, P., Kakade, S. M., Kidambi, R., & Sidford, A. (2017). Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(1), 8258–8299.

Johnson, R. & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* (pp. 315–323).

Juditsky, A. & Nesterov, Y. (2014). Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1), 44–80.

Karimi, B., Miasojedow, B., Moulines, É., & Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. *arXiv preprint arXiv:1902.00629*.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., & Jordan, M. I. (2020). Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*.

Kingma, D. & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (pp. 2845–2853).

Kulunchakov, A. & Mairal, J. (2020). Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21(155), 1–52.

Kushner, H. & Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer.

Lakshminarayanan, C. & Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics* (pp. 1347–1355).

Lan, G., Li, Z., & Zhou, Y. (2019). A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems* (pp. 10462–10472).

Lan, G. & Zhou, Y. (2018). An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1-2), 167–215.

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3), 802–828.

Le Roux, N., Schmidt, M., & Bach, F. R. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems* (pp. 2663–2671).

Lee, S., Wright, S. J., & Bottou, L. (2012). Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(6).

Lei, L. & Jordan, M. (2017). Less than a single pass: Stochastically controlled stochastic gradient. In *International Conference on Artificial Intelligence and Statistics* (pp. 148–156).

Lei, L. & Jordan, M. I. (2020). On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2), 1473–1500.

Li, C. J., Mou, W., Wainwright, M. J., & Jordan, M. I. (2020). ROOT-SGD: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. *arXiv preprint arXiv:2008.12690*.

Li, T., Liu, L., Kyrillidis, A., & Caramanis, C. (2018). Statistical inference using SGD. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liang, T. & Su, W. J. (2019). 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.

Lin, H., Mairal, J., & Harchaoui, Z. (2015). A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems* (pp. 3384–3392).

Liu, D., Nguyen, L. M., & Tran-Dinh, Q. (2020). An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., & Jordan, M. I. (2020). On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory* (pp. 2947–2997).

Nedic, A. & Bertsekas, D. P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1), 109–138.

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4), 1574–1609.

Nemirovskii, A. & Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269 (pp. 543–547).

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1), 221–259.

Nesterov, Y. (2018). *Lectures on Convex Optimization*, volume 137. Springer.

Nesterov, Y. & Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1), 177–205.

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning* (pp. 2613–2621).

Nguyen, L. M., Nguyen, P. H., Richtárik, P., Scheinberg, K., Takáč, M., & van Dijk, M. (2019). New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176), 1–49.

Nguyen, L. M., Scheinberg, K., & Takáč, M. (2021). Inexact SARAH algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1), 237–258.

Pham, N. H., Nguyen, L. M., Phan, D. T., & Tran-Dinh, Q. (2020). ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110), 1–48.

Polyak, B. T. (1990). A new method of stochastic approximation type. *Automat. i Telemekh*, 51(7 pt. 2), 937–946.

Polyak, B. T. & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4), 838–855.

Rakhlin, A., Shamir, O., & Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 449–456).

Robbins, H., Monro, S., et al. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3), 400–407.

Ruppert, D. (1988). *Efficient estimations from a slowly convergent Robbins-Monro process*. Technical report, Cornell University Operations Research and Industrial Engineering.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 107–194.

Shalev-Shwartz, S. (2016). SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning* (pp. 747–754).

Shalev-Shwartz, S. & Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 567–599.

Shamir, O. & Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning* (pp. 71–79).

Sra, S., Nowozin, S., & Wright, S. J. (2012). *Optimization for Machine Learning*. MIT Press.

Su, W. J. & Zhu, Y. (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., & Nguyen, L. M. (2021). A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, (pp. 1–67).

Tripuraneni, N., Flammarion, N., Bach, F., & Jordan, M. I. (2018). Averaging stochastic gradient descent on Riemannian manifolds. In *Conference on Learning Theory* (pp. 650–687).

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Application to Statistics*. Springer.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., & Tarokh, V. (2019). SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems* (pp. 2406–2416).

Woodworth, B. E. & Srebro, N. (2016). Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems* (pp. 3639–3647).

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1), 3–34.

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88), 2543–2596.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine learning* (pp. 919–926).

Zhou, D., Xu, P., & Gu, Q. (2020). Stochastic nested variance reduction for nonconvex optimization. *Journal of Machine Learning Research*, 21(103), 1–63.

Zhu, Y., Chatterjee, S., Duchi, J., & Lafferty, J. (2016). Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems*, volume 29 (pp. 3431–3439).