ROOT-SGD: Sharp Nonasymptotics and Asymptotic Efficiency in a Single Algorithm

Chris Junchi Li* JUNCHILI@BERKELEY.EDU

Wenlong Mou* WMOU@BERKELEY.EDU

Department Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720

Martin J. Wainwright WAINWRIG@BERKELEY.EDU

Michael I. Jordan JORDAN@CS.BERKELEY.EDU

Department Electrical Engineering and Computer Sciences & Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

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

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

1. Introduction

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

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{where } F(\theta) := \mathbb{E}\big[f(\theta; \xi)\big], \tag{1}$$

and where the expectation is taken over a random vector $\xi \in \Xi$ with distribution $\mathbb{P}$. Our goal is to approximately solve this minimization problem based on samples $(\xi_i)_{i=1,2,\cdots} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and moreover to

* CJL and WM contributed equally to this work.

do so in a way that is computationally efficient and statistically optimal. When the samples arrive as an online stream, it is desirable to compute the approximate solution in a single pass, without storing the data, and this paper focuses on this online setting.

Stochastic optimization problems of this type underpin a variety of methods in large-scale machine learning and statistical inference. One of the simplest methods is *stochastic gradient descent* (SGD), which recursively updates a parameter vector θ_t by taking a step in the direction of a single stochastic gradient, with a (possibly) time-varying step-size η_t (Robbins and Monro, 1951). This simple strategy has been surprisingly successful in modern large-scale statistical machine learning problems (Nemirovski et al., 2009; Bottou et al., 2018; Nguyen et al., 2019); however, it can be substantially improved, both in theory and in practice, by algorithms that make use of more than a single stochastic gradient. Such algorithms belong to the general family of *stochastic first-order methods*. Various procedures have been studied, involving different weightings of past stochastic gradients, and also a range of analysis techniques. The diversity of approaches is reflected by the wide range of terminology, including *momentum*, *averaging*, *acceleration*, and *variance reduction*. All of these ideas center around two main underlying goals—that of proceeding quickly to a minimum, and that of arriving at a final state that achieves the optimal statistical efficiency and also provides a calibrated assessment of the uncertainty associated with the solution.

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

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

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

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

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{ERM}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right). \tag{2}$$

Furthermore, the asymptotic distribution (2) is known to be locally asymptotic minimax, i.e. given a bowl-shaped loss function, the asymptotic risk of *any* estimator is lower bounded by the expectation under such a Gaussian distribution, in a suitably defined sequence of local neighborhoods. See, for example, Duchi and Ruan (2021) for a precise statement.

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

(1) The Polyak-Ruppert-Juditsky (PRJ) procedure (Polyak and Juditsky, 1992; Polyak, 1990; Ruppert, 1988) incorporates slowly diminishing step-sizes into SGD, thereby achieving asymptotic normality with an optimal covariance matrix (and unity pre-factor). This meets the goal of calibrated uncertainty. However, the PRJ procedure is *not* optimal from a nonasymptotic point of view: rather, it suffers from large high-order nonasymptotic terms and fails to achieve the optimal sample complexity in general (Bach and Moulines, 2011).

(2) On the other hand, variance-reduced stochastic optimization methods have been designed to achieve reduced sample complexity that is the sum of a *statistical error* and an *optimization error* (Le Roux et al., 2012; Shalev-Shwartz and Zhang, 2013; Johnson and Zhang, 2013; Lei and Jordan, 2017; Defazio et al., 2014). These methods yield control on the optimization error, with sharp nonasymptotic rates of convergence, but the guarantees for the statistical error term yield an asymptotic behavior involving constant pre-factors that are strictly greater than unity, and is hence sub-optimal.

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

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

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

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

Note that the leading-order term of the bound (3) is exactly the squared norm of the Gaussian random vector in the local asymptotic minimax limit or Cramér-Rao lower bound, with unity pre-factor. Moreover, our bound is entirely nonasymptotic, valid for all finite n. We also prove that high-order term $O\big(n^{-3/2}\big)$ is unavoidable under a natural setup, and it improves upon existing $O(n^{-7/6})$ and $O(n^{-5/4})$ rates for the PRJ procedure (Bach and Moulines, 2011; Xu, 2011; Gadat and Panloup, 2017). We also derive similar bounds for the objective gap $F\big(\widehat{\theta}_n^{\mathrm{ROOT}}\big) - F(\theta^*)$ and the gradient norm $\|\nabla F\big(\widehat{\theta}_n^{\mathrm{ROOT}}\big)\|^2$.

(2) Furthermore, the nonasymptotic bound (3) holds true under a mild sample-size requirement. Indeed, given a L-smooth and μ-strongly convex population-level function F, and assuming that the noise $\varepsilon(\cdot;\xi) := \nabla f(\cdot;\xi) - \nabla F(\cdot)$ satisfies a stochastic Lipschitz condition with parameter ℓ_Ξ, the finite-sample bounds are viable as long as $n \gtrsim \frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}$. The first term $O\big(\frac{L}{\mu}\big)$ matches the iteration complexity of gradient descent, and the $O\big(\frac{\ell_\Xi^2}{\mu^2}\big)$ term is the sample complexity needed for distinguishing a μ-strongly convex function from a constant function. The high-order terms in Eq. (3) also depend on the parameters (μ, L, ℓ_Ξ) in a similar way. This exhibits the fast nonasymptotic convergence of our algorithm, matching state-of-the-art variance reduction algorithms.

(3) We also establish asymptotic guarantees for the ROOT-SGD algorithm. Assuming insetad a mild one-point Hessian continuity condition at the minimizer, for a broad range of step-size choices the last iterate $\widehat{\theta}_n^{\mathrm{ROOT}}$ converges in distribution to the optimal Gaussian law (2) whenever the Hessian matrix $\nabla^2 F$ is continuous at θ^*, a much weaker condition manifesting the difference between ROOT-SGD and the Polyak-Ruppert averaging procedure (Polyak and Juditsky, 1992).

Notably, both the MSE bound of the form (3) and the asymptotic normality are fine-grained guarantees that are satisfied by the empirical risk minimizer, under comparable assumption posed on the continuity of Hessian matrix. To the best of our knowledge, such guarantees have not been available heretofore in the literature on stochastic optimization. The ROOT-SGD algorithm proposed in this paper achieves these guarantees not only simultaneously, but also with sharp nonasymptotic sample complexity.

The rest of the paper is organized as follows. We present the ROOT-SGD algorithm in §2, and delineate the asymptotic normality and nonasymptotic upper bounds in §3. We present our conclusions in §4. Full proofs and discussions are provided in the appendix.

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

We finally introduce some martingale-related notations. Given vector-valued martingales $(X_t)_{t\geq T_0}, (Y_t)_{t\geq T_0}$ adapted to the filtration $(\mathcal{F}_t)_{t\geq T_0}$, we use the following notation for cross variation for $t \geq T_0$:

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

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

2. Constructing the ROOT-SGD algorithm

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

2.1. Motivation and gradient estimator

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

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

$$\mathsf{Estimator}_t - \nabla F(\theta_{t-1}) = \frac{1}{t}\sum_{s=1}^{t}(\nabla f(\theta_{s-1};\xi_s) - \nabla F(\theta_{s-1})). \tag{4}$$

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

The recursive aspect of the algorithm arises as follows. We set $\mathsf{Estimator}_1 = \nabla f(\theta_0;\xi_1)$ and express (4) 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. Intuitively, the replacement should not affect much as long as the stochastic function admits some smoothness condition. Letting v_t denote $\mathsf{Estimator}_t$ with this replacement, we obtain the following recursive update:

$$\begin{aligned} v_t &= \frac{1}{t}\nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}\left(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right) + \frac{t-1}{t}v_{t-1} \\ &= \underbrace{\nabla f(\theta_{t-1};\xi_t)}_{\text{stochastic gradient}} + \underbrace{\frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2};\xi_t)\right)}_{\text{correction term}}, \end{aligned} \tag{5}$$

consisting of both a stochastic gradient and a correction term.

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

$$v_t = \nabla f(\theta_{t-1};\xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2};\xi_t)\right) \tag{6a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t, \tag{6b}$$

where $\{\eta_t\}_{t\geq 1}$ is a suitably chosen sequence of positive step-sizes. Note that v_t defined in Eq. (5) is a recursive estimate of $\nabla F(\theta_{t-1})$ that is *unconditionally* unbiased in the sense that $\mathbb{E}[v_t] = \mathbb{E}[\nabla F(\theta_{t-1})]$. So the θ-update is an approximate gradient-descent step that moves along the negative direction $-v_t$.[1]

We initialize $\theta_0 \in \mathbb{R}^d$, and, to avoid ambiguity, we define the update (6) at $t = 1$ to use only $v_1 = \nabla f(\theta_0;\xi_1)$. Overall, given the initialization $(\theta_0, v_0, \theta_{-1}) = (\theta_0, 0, \text{arbitrary})$, at each step $t \geq 1$ we take as input $\xi_t \sim P$, and perform an update of $(\theta_t, v_t, \theta_{t-1})$. This update depends only on $(\theta_{t-1}, v_{t-1}, \theta_{t-2})$ and ξ_t, and is first-order and Markovian.

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

2.2. Two-time-scale structure and burn-in period

For the purposes of both intuition and the proof itself, it is useful to observe that the iterates (6) evolve in a two-time-scale manner. Define the process $z_t := v_t - \nabla F(\theta_{t-1})$ for $t = 1, 2, \cdots$, which characterizes the *tracking error* of v_t as an estimator for the gradient. For each $\theta \in \mathbb{R}^d$ and $\xi \sim P$ we define the noise term $\varepsilon(\theta; \xi) = \nabla_\theta f(\theta; \xi) - \nabla F(\theta)$, and use the shorthand notation $\varepsilon_s(\cdot) = \varepsilon(\cdot; \xi_s)$. Some algebra yields the decomposition

$$t \cdot z_t = \sum_{s=1}^{t} \varepsilon_s(\theta_{s-1}) + \sum_{s=1}^{t} (s-1)\big(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\big), \quad \text{valid for } t = 1, 2, \ldots. \tag{7a}$$

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

$$tv_t = (t-1)\big\{v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\big\} + \nabla F(\theta_{t-1}) + \varepsilon_t(\theta_{t-1}) + (t-1)\big(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\big). \tag{7b}$$

Given an appropriate step-size η_t, the first term on the RHS of Eq. (7b) exhibits a contractive behavior. Consequently, the process $(tv_t)_{t \geq 1}$ plays the role of a *fast process*, driving the motion of iterates $(\theta_t)_{t \geq 0}$, and the noise-collecting process z_t is a *slow process*, collecting the noise along the path and contributing to the asymptotic efficiency of θ_t. Note that the fast process moves with a step-size η_t, making $\eta_t \mu$ progress when F is μ-strongly convex, while the slow process works with a step-size $\frac{1}{t}$. In order to make the iterates stable, we need the fast process to be fast in a relative sense, requiring that $\eta_t \mu \geq \frac{1}{t}$. This motivates a burn-in period in the algorithm, namely, in the first T_0 iterations, we run the recursion (6) with step-size zero and simply average the noise at θ_0; we then start the algorithm with an appropriate choice of step-size. Concretely, given some initial vector $\theta_0 \in \mathbb{R}^d$, we set $\theta_t = \theta_0$ for all $t = 1, \ldots, T_0 - 1$, and compute

$$v_t = \frac{1}{t} \sum_{s=1}^{t} \nabla f(\theta_0; \xi_s), \qquad \text{for all } t = 1, \ldots, T_0. \tag{8}$$

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

$$\eta_t = \begin{cases} \eta, & \text{for } t \geq T_0, \\ 0, & \text{for } t = 1, \ldots, T_0 - 1, \end{cases} \tag{9}$$

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

briefed as $\eta_t = \eta \cdot 1[t \geq T_0]$, and, accordingly, the update rule from Eqs. (6a) and (6b) splits into two phases:

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

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

Such an algorithmic design has the length of the burn-in period for our algorithm is identical to the number of processed samples, so it features that the iteration number is identical to the sample complexity. The ROOT-SGD scheme is presented formally as Algorithm 1; for the remainder of this paper, when referring to ROOT-SGD, we mean Algorithm 1 unless specified otherwise.

3. Main results

In this section, we present our main nonasymptotic and asymptotic results. We first establish a preliminary nonasymptotic result in §3.1. With augmented smoothness and moment assumptions, we then introduce in §3.2 sharp nonasymptotic upper bounds with unit pre-factor on the term characterizing the optimal statistical risk. Finally, in §3.3, we establish the asymptotic efficiency of ROOT-SGD.

3.1. Preliminary nonasymptotic results

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

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

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

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

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

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

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

Note that we only assume a finite variance on the stochastic gradient at the global minimizer θ^*. This is significantly weaker than the standard assumption of a globally bounded noise variance. See Nguyen et al. (2019) and Lei and Jordan (2020) for a detailed discussion of this assumption on the noise.

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

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

$$\mathbb{E}\left\|\varepsilon(\theta;\xi) - \varepsilon(\theta';\xi)\right\|_2^2 \leq \ell_\Xi^2 \left\|\theta - \theta'\right\|_2^2, \qquad \text{for all pairs } \theta;\theta' \in \mathbb{R}^d. \tag{11}$$

We note that in making Assumption 3, we separate the stochastic smoothness (in the L^2 sense) of the noise, $\varepsilon(\theta;\xi) = \nabla f(\theta;\xi) - \nabla F(\theta)$, from the smoothness of the population-level objective. The magnitude of ℓ_Ξ and L are not comparable in general. This flexibility permits, for example, mini-batch algorithms where the population-level Lipschitz constant L remains fixed but the parameter ℓ_Ξ decreases with batch size. Such a separation has been adopted in nonconvex stochastic optimization literature (Arjevani et al., 2020).[2]

Finally, we remark that all of these assumptions are standard in the stochastic optimization and statistical literature; and specific instantations of these assumptions are satisfied by a broad class of statistical models and estimators. We should note, however, that the strong convexity and smoothness (Assumption 1) is a global condition stronger than those typically used in the asymptotic analysis of M-estimators in the statistical literature. These conditions are needed for the fast convergence of the algorithm as an optimization algorithm, making it possible to establish nonasymptotic bounds. Assumptions 2 and 3 are standard for proving asymptotic normality of M-estimators and Z-estimators (see, e.g., van der Vaart, 2000, Theorem 5.21). In contrast to some prior work (e.g., Ghadimi and Lan, 2012, 2013), we *do not* assume uniform upper bounds on the variance of the stochastic gradient noise; this assumption fails to hold for various statistical models of interest, and theoretical results that dispense with it are of practical interest.

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

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

$$T_0 := \frac{24}{\eta\mu} \qquad \text{and} \quad \eta \in (0, \eta_{max}], \qquad \text{where} \quad \eta_{max} := \frac{1}{4L} \wedge \frac{\mu}{8\ell_\Xi^2}. \tag{12}$$

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

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \frac{28\ \sigma_*^2}{T} + \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 T^2}. \tag{13}$$

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

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

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

We provide a complete analysis of Theorem 1 in §C.1. In order to interpret the result, we make few remarks in order.

(i) Note when stating the upper bound (13) we adopt the convergence metrics in expected squared gradient norm. The guarantee in (13) consists of the sum of two terms which differs in magnitude as $T \to \infty$. The leading-order first term is contributed by the *optimal statistical risk*, and is determined by the noise variance σ_*^2 at the minimizer. The higher-order second term, on the other hand, exhibits an $O(\frac{1}{T^2})$-dependency on the initial condition, which is suboptimal and can be improved to an exponential dependency by properly restarting the algorithm. When the step-size η is fixed, a comparison of the two summand terms (13) yields that the optimal asymptotic risk $\frac{\sigma_*^2}{T}$ for the squared gradient holds up to an absolute constant whenever $T \gtrsim \frac{1}{\eta\mu} \vee \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2\sigma_*^2}$.

(ii) Suppose that we use the maximal step-size η_{max} permitted by condition (12). By converting the convergence rate bound (13) into a sample complexity bound, we then find that it suffices to take

$$C_1(\varepsilon) = \max\left\{\frac{74}{\eta_{\max}\mu}\cdot\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{56\sigma_*^2}{\varepsilon^2}\right\} \asymp \max\left\{\left(\frac{L}{\mu}+\frac{\ell_\Xi^2}{\mu^2}\right)\cdot\frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2}\right\} \tag{14}$$

samples in order to obtain an estimate of θ^* with gradient norm bounded as $O(\varepsilon)$. When the asymptotics holds as ε tends to zero with other problem-dependent constants being bounded away from zero, the leading-order term $\asymp \frac{\sigma_*^2}{\varepsilon^2}$ in $C_1(\varepsilon)$ matches the optimal statistical risk up to a constant pre-factor. To the best of our knowledge, such a bound on sample complexity is achieved for the first time by a stochastic first-order algorithm in the setting where only first-order smoothness condition holds, i.e. no continuity condition on the Hessians are posed. The only prior results which reported a near-optimal statistical risk under comparable settings in the leading-order stochastic optimization are due to Nguyen et al. (2021) and Allen-Zhu (2018), achieving the optimal risk up to a polylogarithmic factor. In Appendix §B, we compare our non-asymptotic results with these existing works in detail.

3.2. Improved nonasymptotic upper bounds

The convergence rate bound of Theorem 1 matches the optimal risk by a constant pre-factor c—to be precise, $c = 28$ in the provided analysis. In addition to this non-optimal pre-factor, this result does not match the efficiency of M-estimators in its higher-order dependency. So as to overcome these limitations, we now show how to apply Theorem 1 as the building block to seek to obtain a sharp fine-grained convergence rate via two-time-scale characterization, under additional smoothness and moment assumptions.

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

Assumption 4 (Lipschitz continuous Hessians) *There exists a Lipschitz constant $L_1 > 0$ such that*

$$\left\|\nabla^2 F(\theta) - \nabla^2 F(\theta^*)\right\|_{op} \le L_1 \left\|\theta - \theta^*\right\|_2. \tag{15}$$

Algorithm 2 ROOT-SGD, multi-epoch version

1: **Input:** initialization θ_0; fixed step-size η; burn-in time T_0; short epochs length $T^\flat \geq T_0$; short epochs number B
2: Set initialization for first epoch $\theta_0^{(1)} = \theta_0$
3: **for** $b = 1, 2, \cdots, B$ **do**
4: Run ROOT-SGD (Algorithm 1) for $T^\flat$ iterates with burn-in time T_0 (i.e. step-size sequence $(\eta_t)_{t\geq 1}$ defined as in Eq. (9))
5: Set the initialization $\theta_0^{(b+1)} := \theta_{T^\flat}^{(b)}$ for the next epoch
6: **end for**
7: Run ROOT-SGD (Algorithm 1) for $T := n - T^\flat B$ iterates with burn-in time T_0
8: **Output:** The final iterate estimator $\theta_n^{\text{final}} := \theta_T^{(B+1)}$

We also need fourth-moment analogue of Assumptions 2 and 3, stated as follows. Note that these conditions are also exploited in prior work on nonasymptotic analyses of PRJ averaging procedure (Bach and Moulines, 2011; Xu, 2011; Gadat and Panloup, 2017) and Streaming SVRG (Frostig et al., 2015).

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

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

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

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

$$\sqrt{\mathbb{E}\left\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\right\|_2^4} \leq \ell_\Xi^2 \left\|\theta_1 - \theta_2\right\|_2^2, \qquad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d. \tag{16}$$

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

Formally, we present a multi-epoch version of the ROOT-SGD algorithm in Algorithm 2. The algorithm runs B short epochs and one long epoch. The goal of each short epoch is to "halve" the dependency on the initial condition $\left\|\nabla F(\theta_0)\right\|_2$, and it suffices to take $T^\flat = cT_0$ for some universal constant $c > 1$. We further impose the mild condition that the quantity $\left\|\nabla F(\theta_0)\right\|_2 / \sigma_*$ scales as a polynomial function of n.[3] In the following Theorem 2, we present the gradient norm bounds satisfied by the multi-epoch ROOT-SGD algorithm:

Theorem 2 (Improved nonasymptotic upper bound, multi-epoch ROOT-SGD) *Under Assumptions 1, 4, 5, 6, suppose that we run Algorithm 2 with the number of short epochs* $B = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$,

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

the burn-in time $T_0 = \frac{24}{\eta\mu}$*, and the small epoch length* $T^\flat = \frac{7340}{\eta\mu}$*. Then for any step-size* $\eta \in \left(0, \frac{1}{56L} \wedge \frac{\mu}{64\ell_\Xi^2}\right]$ *and* $n \geq T^\flat B + 1$*, it returns an estimate* θ_n^{final} *such that*

$$\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\log T}{\eta\mu T} + \frac{\ell_\Xi^2 \log T}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \tag{17}$$

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

See §C.4 for the proof of this theorem.

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

$$\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{n} \lesssim \left\{\frac{\ell_\Xi}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}\frac{\sigma_*^2 \log n}{n} + \underbrace{\left\{\frac{\ell_\Xi}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}^{1/2}\frac{L_1}{\mu^2}\left(\frac{\widetilde{\sigma_*}^2}{n}\right)^{3/2}}_{=:\widetilde{\mathcal{H}}_n}, \tag{18}$$

where $\widetilde{\mathcal{H}}_n$ is the linearization-induced term. Given the sufficiently large sample size n satisfying the requirement $n \gtrsim \frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}$, the pre-factors in the second term of (18), as well as the linearization error term $\widetilde{\mathcal{H}}_n$, start to diminish when the sample size n grows. The gradient norm bound (18) consists of three terms. We discuss each of them as follows:

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

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

(iii) The remaining high-order term $\widetilde{\mathcal{H}}_n$ in the bound (18) scales as $O(n^{-7/4})$, a faster rate of decay than the previous term. This term is induced by a linearization argument in our proof, and therefore depends on the Lipschitz constant L_1 of the Hessian matrix.

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

Besides the expected gradient norm squared metric, we also establish guarantees in alternative metrics including the estimation error $\|\theta_n - \theta^*\|_2$ and the objective gap $F(\theta_n) - F(\theta^*)$ with expectation

taken. In order to state the theorem, we define the following linearization-induced error terms that appears in the bound

$$\widetilde{\mathcal{H}}_n^{(\mathrm{MSE})} := \frac{c}{\lambda_{\min}(H^*)^2} \cdot \left\{ \frac{L_1}{\mu^2} \cdot \left(\frac{\widetilde{\sigma_*^2}}{n} \right)^{3/2} + \frac{L_1^2}{\mu^4} \cdot \left(\frac{\widetilde{\sigma_*^2}}{n} \right)^2 \right\}, \quad \text{and} \tag{19a}$$

$$\widetilde{\mathcal{H}}_n^{(\mathrm{OBJ})} := \frac{c}{\mu} \cdot \frac{L_1}{\mu^2} \cdot \left(\frac{\widetilde{\sigma_*^2}}{n} \right)^{3/2} + \frac{c}{\lambda_{\min}(H^*)} \cdot \frac{L_1^2}{\mu^4} \cdot \left(\frac{\widetilde{\sigma_*^2}}{n} \right)^2. \tag{19b}$$

For simplicity we only consider the multi-epoch ROOT-SGD as specified in Theorem 2, where we conclude the following Corollary:

Corollary 3 (Nonasymptotic bounds in alternative metrics, multi-epoch ROOT-SGD) *Under the setup of Theorem 2, the multi-epoch ROOT-SGD algorithm with the optimal step-size choice of* $\eta \asymp \frac{1}{L} \wedge \frac{1}{\ell_\Xi \sqrt{n}}$ *produces an estimator that satisfies the following bound for* $n \geq T^\flat B + 1$*:*

$$\mathbb{E} \left\| \theta_n^{\mathrm{final}} - \theta^* \right\|_2^2 - \frac{1}{n} \mathrm{Tr} \left((H^*)^{-1} \Sigma^* (H^*)^{-1} \right) \leq c \left\{ \frac{\ell_\Xi}{\mu \sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{\lambda_{\min}(H^*)^2 n} + \widetilde{\mathcal{H}}_n^{(\mathrm{MSE})}, \tag{20a}$$

$$\mathbb{E} \left[F\left(\theta_n^{\mathrm{final}}\right) - F(\theta^*) \right] - \frac{1}{2n} \mathrm{Tr} \left((H^*)^{-1} \Sigma^* \right) \leq c \left\{ \frac{\ell_\Xi}{\mu \sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{\lambda_{\min}(H^*) n} + \widetilde{\mathcal{H}}_n^{(\mathrm{OBJ})}. \tag{20b}$$

See §C.5 for the proof of this result, where the key technical addition lies on the adoption of a generic matrix-induced bound on the stochastic processes. Obviously, the optimal step-size choice in alternative metrics is same in magnitude as the one in squared gradient norm metric in Theorem 2. We can compare the bounds in Corollary 3 with the gradient norm bound (18) induced by Theorem 2, discussed term-by-term as follows:

(i) The leading-order terms in the bound (20a) and (20b), specified in the subtracted second terms on the left hands, are both optimal in a local asymptotic minimax sense with unity pre-factor. In particular, they are exactly the asymptotic risk of the limiting Gaussian random variable $\mathcal{N}\left(\theta^*, \frac{1}{n}(H^*)^{-1}\Sigma^*(H^*)^{-1}\right)$ in corresponding metrics. We also note that in the special case of well-specified maximal-likelihood estimation, Fisher's identity $H^* = \Sigma^*$ holds true, and the leading-order terms in Eq. (20a) and (20b) become $\frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\right)$ and $\frac{d}{2n}$, respectively. This is in accordance with the classical asymptotic theory for M-estimators (c.f. van der Vaart (2000, §5.3));

(ii) The dominating term $\left\{ \frac{\ell_\Xi}{\mu\sqrt{n}} + \frac{L}{\mu n} \right\} \frac{\sigma_*^2 \log n}{n}$ in the gradient norm bound is correspondingly multiplied by a factor of $\lambda_{\min}(H^*)^{-2}$ (resp. $\lambda_{\min}(H^*)^{-1}$) in the MSE (resp. objective gap) bound on the right hand (20a) (resp. (20b)), which is intuitively consistent with conversions in metric;

(iii) While the dominating high-order term on the right hand matches the corresponding optimal statistical risk and the higher-order terms altogether scale as $O(n^{-3/2})$, the linearization-induced error terms $\widetilde{\mathcal{H}}_n^{\mathrm{MSE}}$ and $\widetilde{\mathcal{H}}_n^{\mathrm{OBJ}}$ both decay at a rate of $O\left(n^{-3/2}\right)$ as long as L_1 is bounded away from zero i.e. the objective is essentially nonquadratic. This is vastly different from the bound in

gradient norm (17) of Theorem 2 where the linearization-induced terms are all incorporated in $O(n^{-7/4})$, primarily due to that the pre-factor $\left\{\frac{\ell_\Xi}{\mu\sqrt{n}} + \frac{L}{\mu n}\right\}^{1/2}$ in the linearization-induced term $\widetilde{\mathcal{H}}_n$ in (18) is replaced by unity.[4] In consistency with the metric conversion, the linearization-induced terms $\widetilde{\mathcal{H}}_n^{\mathrm{MSE}}$ and $\widetilde{\mathcal{H}}_n^{\mathrm{OBJ}}$ also incur additional factors related to the smallest eigenvalue of H^* on top of the linearization-induced error term $\widetilde{\mathcal{H}}_n$ in (18).

3.3. Asymptotic results

In this subsection, we study the asymptotic behavior of our ROOT-SGD algorithm. We aim to prove the asymptotic efficiency of the multi-epoch estimator of Algorithm 2 under minimal assumptions. In this case, Assumptions 1, 2 and 3 are the standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see e.g. van der Vaart (2000, Theorem 5.21)). We first introduce our one-point Hessian continuity condition as follows as the qualitative counterpart of the continuity Assumption 4:

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

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

Note in Assumption 7 we assume only the continuity of Hessian matrix at θ^* without posing any bounds on its modulus of continuity. This is a much weaker condition than a Lipschitz or Hölder condition posted on the Hessian matrix as required in the analysis of the Polyak-Ruppert averaging procedure (Polyak and Juditsky, 1992).

With this setup, we are ready to state our weak convergence asymptotic efficiency result for θ^{final} in the following theorem,[5] whose proof is provided in §C.6:

Theorem 4 (Asymptotic efficiency, multi-epoch ROOT-SGD) *Under Assumptions 1, 2, 3 and 7, suppose that we run the multi-epoch Algorithm 2 with burn-in time $T_0 = \frac{24}{\eta\mu}$, short-epoch length $T^\flat = \frac{7340}{\eta\mu}$ and number of short epochs $B = \left\lceil \frac{1}{2}\log\left(\frac{e\|\nabla F(\theta_0)\|_2^2}{\eta\mu\sigma_*^2}\right)\right\rceil$. Then as $n \to \infty$, $\eta \to 0$ such that $\eta(n - T^\flat B) \to \infty$ and $T^\flat B/n \to 0$, the estimate $\theta_n^{\mathrm{final},(\eta)}$ satisfies the weak convergence*

$$\sqrt{n}\left(\theta_n^{\mathrm{final},(\eta)} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, [\nabla^2 F(\theta^*)]^{-1}\Sigma^*[\nabla^2 F(\theta^*)]^{-1}\right), \tag{21}$$

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

We remark that Theorem 4 holds under the mere additional assumption of one-point continuity on the Hessian matrix, which is usually the minimal assumption needed for an asymptotic efficiency result

4. This is because the Hessian-Lipschitz assumption plays a key role in relating MSE and objective gap to the underlying noise structure in the stochastic optimization problem, paying for larger linearization error; whereas in the gradient norm bound, the Hessian-Lipschitz assumption is employed only to mitigate the effect correlation that appears at even higher-order terms in the bound.
5. We emphasize our estimator's dependency on the step-size η by explicitly bracketing it in the superscript.

to hold. Here we are adopting the multi-epoch ROOT-SGD with the same algorithmic specifications as in Theorem 2, and we achieve the asymptotic convergence to the Gaussian limit that matches the Cramér-Rao lower bound. The asymptotic covariance matrix in Eq. (21), however, carries significantly more information than the (scalar) optimal asymptotic risk. Our asymptotic result is in a triangular-array format: we let the fixed constant step-size scale down with n where the scaling condition is essentially $\eta \to 0$, $n \to \infty$ with $\frac{\eta n}{\log(\eta^{-1})} \to \infty$, which is satisfied when $\eta \asymp \frac{1}{n^{c_1}}$ for any fixed $c_1 \in (0, 1)$. Although not directly comparable, the range of step-size asymptotics is broader than Polyak and Juditsky (1992) and accordingly hints at potential advantages over PRJ, primarily due to our de-biasing corrections in our algorithm design and is consistent with our improved higher-order term in nonasymptotic result (Theorem 2 and Corollary 3). In Appendix §D, we establish an additional asymptotic normality result for ROOT-SGD with fixed *constant* step-size, which exhibits exactly the same limiting behavior as constant-step-size *linear* stochastic approximation with PRJ averaging procedure under comparable asssumptions (Mou et al., 2020).

We end this subsection by remarking that Theorem 4 only requires strong convexity, smoothness, and a set of noise moment assumptions standard in asymptotic statistics, but not any higher-order smoothness other than the continuity of Hessian matrices at θ^*. This matches the assumptions for asymptotic efficiency results in classical statistics literature (van der Vaart, 2000; van der Vaart and Wellner, 1996).

4. Future directions

We have shown that ROOT-SGD enjoys favorable asymptotic and nonasymptotic behavior for solving the stochastic optimization problem (1) in the smooth, strongly convex case. With this result in hand, several promising future directions arise. First, it is natural to extend the results for ROOT-SGD to non-strongly convex and nonconvex settings, for both nonasymptotic and asymptotic analyses. Second, it would also be of significant interest to investigate both the nonasymptotic bounds and asymptotic efficiency of the variance-reduced estimator of ROOT-SGD in Nesterov's acceleration setting, in the hope of achieving all regime optimality in terms of the sample complexity to the stochastic first-order oracle. Finally, for statistical inference using online samples, the near-unity nonasymptotic and asymptotic results presented in this work can potentially yield confidence intervals and other inferential assertions for the use of ROOT-SGD estimators.

Acknowledgements

We thank Peter Bartlett, Nicolas Flammarion, Koulik Khamaru, and Nicolas Le Roux for helpful discussions. This work was done in part while the authors were participating in the Theory of Reinforcement Learning program at the Simons Institute for the Theory of Computing. This research was supported in part by Office of Naval Research Grant DOD ONR-N00014-18-1-2640 and N00014-21-1-2842, NSF-DMS grant 2015454 and 1612948, as well as NSF-IIS grant 1909365 to MJW, and also by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941 and NSF grant IIS-1901252 to MIJ. This research was also generously supported by NSF grant NSF-FODSI 202350 to MJW and MIJ.

References

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58):3235–3249, 2012.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems*, pages 1157–1167, 2018.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

Sébastien Arnold, Pierre-Antoine Manzagol, Reza Babanezhad Harikandeh, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. In *Advances in Neural Information Processing Systems*, pages 5391–5402, 2019.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Reza Babanezhad, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Konečnỳ, and Scott Sallinen. Stop wasting my gradients: Practical SVRG. In *Advances in Neural Information Processing Systems*, volume 28, pages 2251–2259, 2015.

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

Dimitris P Bertsekas and John N Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1989.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Léon Bottou and Yann Le Cun. Large scale online learning. In *Advances in Neural Information Processing Systems*, volume 16, pages 217–224. MIT Press, 2004.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International Conference on Machine Learning*, pages 2260–2268. PMLR, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *International Conference on Artificial Intelligence and Statistics*, volume 38, pages 205–213. PMLR, 2015.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

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

Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 686–696, 2018.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

Dylan J Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.

Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, pages 728–763, 2015.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Peter Hall and Christopher C Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(1):8258–8299, 2017.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, 2018a.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory*, pages 545–604, 2018b.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

Jakub Konečnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017.

Andrei Kulunchakov and Julien Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21 (155):1–52, 2020.

Harold Kushner and G George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer, 2003.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1-2):167–215, 2018.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

Nicolas Le Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

Sangkyun Lee, Stephen J Wright, and Léon Bottou. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(6), 2012.

Lihua Lei and Michael Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 148–156, 2017.

Lihua Lei and Michael I Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic Approximation and Optimization of Random Systems*. Springer, 1992.

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

Wenlong Mou, Koulik Khamaru, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. Optimal variance-reduced stochastic approximation in banach spaces. *arXiv preprint arXiv:2201.08518*, 2022.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

Lam M Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takáč, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.

Lam M Nguyen, Katya Scheinberg, and Martin Takáč. Inexact SARAH algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Boris T Polyak. A new method of stochastic approximation type. *Automat. i Telemekh*, 51(7 pt. 2): 937–946, 1990.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

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

Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754, 2016.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2):567–599, 2013.

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

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, pages 1–67, 2021.

Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Conference on Learning Theory*, pages 650–687, 2018.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer, 1996.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2406–2416, 2019.

Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3639–3647, 2016.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*, 2011.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine learning*, pages 919–926, 2004.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 3921–3932, 2018.

Appendices

In this appendix, we provide deferred proofs for theorems and lemmas in the main text organized as follows. §A provides additional work related to us. §B provides additional discussion on comparison of our results with concurrent work. §C proves the main results for both the nonasymptotic and the asymptotic convergence properties of our ROOT-SGD algorithm. §D complements our asymptotic efficiency result in §3.3 and establishes an additional asymptotic result for constant-step-size ROOT-SGD. §E presents auxiliary lemmas stated in §C.3, §C.4 and §C.5. Finally, §F proves necessary lemmas for the proof of Proposition 2 (in §D.1).

Appendix A. Additional related work

SGD and Polyak-Ruppert-Juditsky averaging procedure The theory of the stochastic approximation method has a long history since its birth in the 1950s (Robbins and Monro, 1951; Bottou and Le Cun, 2004; Zhang, 2004; Nemirovski et al., 2009; Bottou, 2010; Bubeck, 2015) and recently regains its attention due to its superb performance in real-world application practices featured by deep learning (Goodfellow et al., 2016), primarily due to its exceptional handling of the online samples. Classics on this topic include Bertsekas and Tsitsiklis (1989); Benveniste et al. (1990); Ljung et al. (1992); Borkar (2008) and many more. Specially on the study of asymptotic normality which can trace back to Fabian (1968), the general idea of iteration averaging is based on the analysis of two-time-scale iteration techniques and it achieves asymptotic normality with an optimal covariance (Ruppert, 1988; Polyak, 1990; Polyak and Juditsky, 1992). Recent work along this line includes Bach and Moulines (2011, 2013); Bach (2014); Défossez and Bach (2015); Flammarion and Bach (2015); Dieuleveut and Bach (2016); Duchi and Ruan (2021); Dieuleveut et al. (2017); Allen-Zhu (2018); Dieuleveut et al. (2020); Asi and Duchi (2019), presenting attractive asymptotic and nonasymptotic properties under a variety of settings and assumptions. Agarwal et al. (2012); Woodworth and Srebro (2016) provide minimax lower bounds for stochastic first-order algorithms. Jain et al. (2017, 2018a,b) analyze SGD and its acceleration with *tail-averaging* that simultaneously achieves exponential forgetting and optimal statistical risk up to a constant, nonunity pre-factor. It is also worth mentioning that iteration averaging provides robustness and adaptivity (Lei and Jordan, 2020). Instead of averaging the iterates, our ROOT-SGD algorithm averages the past stochastic gradients with proper de-biasing corrections and achieves competitive asymptotic performance.[6] For statistical inferential purposes, recent work (Chen et al., 2020; Su and Zhu, 2018) presents confidence interval assertions via online stochastic gradient with Polyak-Ruppert-Juditsky averaging procedure; analogous results for the ROOT-SGD algorithm are hence worth exploring, building upon the asymptotic normality that our work has established.

Variance-reduced gradient methods In the field of smooth and convex stochastic optimization, variance-reduced gradient methods represented by, but not limited to, SAG (Le Roux et al., 2012), SDCA (Shalev-Shwartz and Zhang, 2013), SVRG (Johnson and Zhang, 2013; Konečnỳ and Richtárik, 2017; Babanezhad et al., 2015; Lei and Jordan, 2017), SAGA (Defazio et al., 2014), SARAH (Nguyen et al., 2017, 2021) have been proposed to improve the theoretical convergence rate of (stochastic)

6. A related but fundamentally different idea was proposed in Nesterov (2009); Xiao (2010); Lee et al. (2012) called *dual averaging* for optimizing the regularized objectives. In contrast to their method, we focus in this work on the smooth objective setting and augment our estimator with de-biasing corrections. See also Duchi and Ruan (2021); Tripuraneni et al. (2018) for more on first-order optimization methods on Riemannian manifolds.

gradient descent. Accelerated variants of SGD provide further improvements in convergence rate (Lin et al., 2015; Shalev-Shwartz, 2016; Allen-Zhu, 2017; Lan and Zhou, 2018; Kulunchakov and Mairal, 2020; Lan et al., 2019). More recently, a line of work on recursive variance-reduced stochastic first-order algorithms have been studied in the nonconvex stochastic optimization literature (Fang et al., 2018; Zhou et al., 2018; Wang et al., 2019; Nguyen et al., 2021; Pham et al., 2020; Li et al., 2021). These algorithms, as well as their hybrid siblings (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2021), achieve optimal iteration complexities for an appropriate class of nonconvex functions and in particular are faster than SGD under mild additional smoothness assumption on the stochastic gradients and Hessians (Arjevani et al., 2020). Limited by space, we refer interested readers to a recent survey article by Gower et al. (2020), and while our ROOT-SGD algorithm can be viewed as a variant of variance-reduced algorithms, our goal is substantially different: we aim to establish for strongly convex objectives both a sharp, unity pre-factor nonasymptotic bound and asymptotic normality with Cramér-Rao optimal asymptotic covariance that matches the local asymptotic minimax optimality (Duchi and Ruan, 2021).

Sharp nonasymptotics and asymptotic efficiency When the objective admits additional smoothness, nonasymptotic rate analyses for either SGD with iteration averaging or variance-reduced stochastic first-order algorithms have been studied in various settings. Bach and Moulines (2011) presents a nonasymptotic analysis of SGD with PRJ averaging procedure showing that, after processing n samples, the algorithm achieves a nonasymptotic rate that matches the Cramér-Rao lower bound with a pre-factor equal to one with the additional term being $O(n^{-7/6})$ (see the discussions in §B). Xu (2011); Gadat and Panloup (2017) improves the additional term to $O(n^{-5/4})$ under comparable assumptions. Défossez and Bach (2015); Dieuleveut and Bach (2016); Duchi and Ruan (2021); Asi and Duchi (2019) achieves either sharp nonasymptotic bounds (in the quadratic case) or asymptotic efficiency that matches the local asymptotic minimax lower bound. The asymptotic efficiency of variance-reduced stochastic approximation methods, however, has been less studied. More related to this work, Frostig et al. (2015) establishes the nonasymptotic upper bounds on the objective gap for an online variant of the SVRG algorithm (Johnson and Zhang, 2013), where the leading-order nonasymptotic bound on the excess risk matches the optimal asymptotic behavior of the empirical risk minimizer under certain *self-concordant condition* posed on the objective function; the additional higher-order term reported is at least $\Omega(n^{-8/7})$.

Other related work Lakshminarayanan and Szepesvari (2018); Mou et al. (2020) studies fixed-constant-step-size linear stochastic approximation with PRJ averaging procedure beyond an optimization algorithm (Polyak and Juditsky, 1992), which includes many interesting applications in minimax game and reinforcement learning. To be specific, Lakshminarayanan and Szepesvari (2018) provides general nonasymptotic bounds which suffer from a nonunity pre-factor on the optimal statistical risk, and Mou et al. (2020) studies the PRJ averaging procedure for general linear stochastic approximation and precisely characterizes the asymptotic limiting Gaussian distribution, delineating the additional term that adds onto the Cramér-Rao asymptotic covariance and which vanishes as $\eta \to 0$ (Dieuleveut et al., 2020), and further establishes sharp concentration inequalities under stronger moment conditions on the noise. Arnold et al. (2019) proposes an extrapolation-smoothing scheme of *Implicit Gradient Transportation* to reduce the variance of the algorithm and provides convergence rates for quadratic objectives, which is further generalized to nonconvex optimization to improve the convergence rate of

normalized SGD (Cutkosky and Mehta, 2020). For the policy evaluation problem in reinforcement learning, Khamaru et al. (2020) establishes an instance-dependent non-asymptotic upper bound on the ℓ_∞ estimation error, for a variance-reduced stochastic approximation algorithm. Their bound matches the risk of optimal Gaussian limit up to constant or logarithmic factors. Recently, Mou et al. (2022) extends the algorithmic idea in this work and proposes the recursive variance-reduced stochastic approximation in span seminorm, which is applicable for generative models in reinforcement learning.

Appendix B. Comparison to related works

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

Comparison with classical results on SGD and its acceleration SGD is known to be worst-case optimal for optimizing smooth and strongly-convex objectives up to a constant pre-factor. By way of contrast, our convergence metric in use, oracle query model and assumption on stochasticity are fundamentally different. Despite this, a recent work due to Nguyen et al. (2019) building upon earlier analysis surveyed by Bottou et al. (2018) makes a comparable noise assumption that allows the noise variance to grow at most quadratically with the distance to optimality and applies to SGD. Nguyen et al. (2019) shows that for appropriate diminishing step-sizes η_t we can conclude a guarantee of $\mathbb{E}\|\theta_T^{\mathrm{SGD}} - \theta^*\|_2^2 \lesssim \frac{\sigma_*^2}{\mu^2 T}$. With additional smoothness and noise assumptions we aim to achieve fine-grained non-asymptotic and asymptotic local asymptotic minimax optimality *with unity pre-factor*. It is straightforward to observe that the convergence rate bound of SGD under shared assumptions is in no regime better than that of ROOT-SGD presented in (14).

We turn to compare our ROOT-SGD convergence result with the existing arts on stochastic accelerated gradient descent for strongly convex objectives Ghadimi and Lan (2012, 2013). With an appropriate multi-epoch design, their guarantees on the objective gap are worst-case optimal for optimizing smooth and strongly-convex objectives in terms of the dependency on *all* terms of condition number L/μ and a uniform upper bound on the noise variance. Our preliminary nonasymptotic guarantee for single-epoch ROOT-SGD in Theorem 1, in contrast, does not require uniform boundedness on the variance and depends solely on the variance at the minimizer θ^*.[7] That being said, our result cannot not admit an accelerated rate in terms of condition number $\sqrt{L/\mu}$ even with the help of multi-epoch designs. It is an important direction of future research to incorporate acceleration mechanism into our framework so as to achieve all-regime optimality.

Comparison with near-optimal guarantees in gradient norm Allen-Zhu (2018) develops a multi-epoch variant of SGD with averaging (under the name SGD3) via recursive regularization techniques and achieved a near-optimal rate for attaining an estimator of $O(\varepsilon)$-gradient norm. Our assumptions are not comparable in general: on the one hand, we assume a second-moment version of stochastic Lipschitz assumption (assumption 3), which makes it possible to establish guarantees that depends on

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

the noise variance σ_* at the optimum θ^*; on the other hand, Allen-Zhu (2018) makes no assumption on the modulus of continuity for the stochastic gradient, while their bound depends on a uniform upper bound σ on the noise variance. Besides, it is worth noticing that as $\varepsilon \to 0^+$, the leading-order term in their bound scales as $\frac{\sigma^2}{\varepsilon^2}\log^3\left(\frac{L}{\mu}\right)$, which is sub-optimal by a polylogarithmic factor even if the uniform boundedness assumption on the variance is satisfied. In a subsequent work, Foster et al. (2019) applies the idea of recursive regularization to AC-SA (Ghadimi and Lan, 2012) and achieves an accelerated rate to find an approximate minimizer with ε-gradient norm, while a lower bound analysis provided further justifies the necessity of a multiplicative logarithmic factor in the nonstrongly convex, local oracle setting.

It is also worth-mentioning the (Inexact) SARAH algorithm and its analysis developed by Nguyen et al. (2021), which also achieves a near-optimal complexity upper bound of $O\left(\frac{\sigma_*^2}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right) + \frac{L_{\max}}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$ to obtain an approximate minimizer with ε-gradient norm. Note that the setting for that result is slightly different (their setting is dubbed as the **ISC** case in our extended analysis of Theorem 5 in §C.1). The algorithm of Nguyen et al. (2021) requires random output and burn-in batches that is inversely dependent on the desired accuracy ε, yielding a logarithmic pre-factor on top of the statistical error corresponding to the Cramér-Rao lower bound; in comparison, our preliminary single-epoch ROOT-SGD result has a leading-order term in complexity bound that removes a logarithmic factor, attaining an optimal non-asymptotic guarantee up to a nonunity pre-factor.

Nonasymptotic guarantees matching local asymptotic minimax with near-unity pre-factor [8] For SGD with PRJ averaging procedure, Bach and 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 $\eta_t = Ct^{-\bar{\alpha}}$ for $\bar{\alpha} \in (1/2, 1)$, Bach and Moulines (2011) show that the following bound holds true for the averaged iterates $\bar{\theta}_T$ for the PRJ:

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

which corresponds to an $O(n^{-7/6})$ additional term in the squared estimation error metric ((20a) in Corollary 3). Here, the constant c_0 depends on the initial distance to optimum, smoothness and strong convexity parameters of second- and third-order derivatives, as well as higher-order moments of the noise. Xu (2011); Gadat and Panloup (2017) further improves the higher-order term from $O(n^{-7/6})$ to $O(n^{-5/4})$. The convergence rate of (single-loop) ROOT-SGD is similar to SGD with PRJ averaging procedure in the nature of the leading term and the high-order terms, but our convergence rate bound of ROOT-SGD is comparatively cleaner and easier to interpret.

The work by Frostig et al. (2015) proposes the Streaming SVRG algorithm that provides nonasymptotic guarantees in terms of the objective gap. Under a slightly different setting where smoothness and convexity assumptions are imposed on the individual function, their objective gap bound asymptotically matches the optimal risk achieved by the empirical risk minimizer under an additional self-concordance condition, with a multiplicative constant that can be made arbitrarily small. In particular, via our notations their results take the following form:

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

8. For convenience we include all comparable results in Table 1.

Algorithm	Assumption	Additional Term	Reference
PRJ	Hessian Lipschitz	$O\left(\frac{1}{n^{7/6}}\right)$	(Bach and Moulines, 2011)
PRJ	Hessian Lipschitz	$O\left(\frac{1}{n^{5/4}}\right)$	(Xu, 2011; Gadat and Panloup, 2017)
Streaming SVRG	Self-concordant	multiplicative[9]	(Frostig et al., 2015)
ROOT-SGD	Hessian Lipschitz	$O\left(\frac{1}{n^{3/2}}\right)$	(This work)

Table 1. Comparison of our results with comparative work. For the unity pre-factor nonasymptotic result, we only characterize the additional term to the optimal risk.

where they require $n \geq b^{2p+3}$ for some $p \geq 2$. In order to achieve the sharp pre-factor, the additional term in this bound is at least $\Omega(n^{-8/7})$, a worse rate than our Corollary 3. Additionally, to get the corresponding nonasymptotic guarantees under such a setting, their bound requires a scaling condition $T \gtrsim \frac{L_{\max}^2}{\mu^2}$ where $L_{\max}$ denotes the smoothness of the individual function, which is larger than our burn-in sample size. Without the self-concordance condition, the convergence rate bound of Streaming SVRG suffers from an extra multiplicative factor belonging to the interval $\left[1, \frac{L_{\max}}{\mu}\right]$, and its leading-order term thereby admits a dependency on the condition number worse than SGD.

Appendix C. Proofs of nonasymptotic and asymptotic results

We provide the convergence rate analysis and the proofs of our theorems in this section. In our analysis we utilize the central object the *tracking error process* z_t defined as in (29), and we heavily use the fact that the process $(tz_t)_{t \geq T_0}$ is a martingale adapted to the natural filtration.

C.1. Proof of Theorem 1 and extended analysis

This subsection is devoted to an (extended) analysis and proof of Theorem 1. In part of our analysis, as an alternative to our Lipschitz stochastic noise Assumption 3, we can impose the following *individual convexity and smoothness* condition (Le Roux et al., 2012; Johnson and Zhang, 2013; Defazio et al., 2014; Nguyen et al., 2017):

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

$$\left\|\nabla f(\theta; \xi) - \nabla f(\theta'; \xi)\right\|_2 \leq L_{\max} \left\|\theta - \theta'\right\|_2 \ a.s., \quad \textit{for all pairs } \theta; \theta' \in \mathbb{R}^d. \tag{22}$$

All Assumptions 1 and 2 along with either Assumption 3 or 8, are standard in the stochastic optimization literature (cf. Nguyen et al. (2019); Asi and Duchi (2019); Lei and Jordan (2020)). Note that Assumption 8 implies Assumption 3 with constant $L_{\max}$; in many statistical applications, the quantity $L_{\max}$ can be significantly larger than $\sqrt{L^2 + \ell_{\Xi}^2}$ in magnitude.

9. Note that the paper Frostig et al. (2015) achieves a risk bound whose leading-order term is a $1 + O(b^{-1})$-multiplicative approximation to the optimal risk, with some additional terms (See Corollary 5 in their paper). Since this result requires $b^7 \leq n$, the additional term is at least $\Omega(n^{-8/7})$.

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

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

$$\eta_{max} := \frac{1}{4L} \wedge \frac{\mu}{8\ell_{\Xi}^2}. \tag{23}$$

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

$$\eta_{max} := \frac{1}{4L_{\max}}. \tag{24}$$

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

Theorem 5 (Unified nonasymptotic results, single-epoch ROOT-SGD) *Suppose that the conditions in either the **LSN** or **ISC** Case are in force, and let the step sizes be chosen according to the protocol (9) for some $\eta \in (0, \eta_{max}]$, and assume that we use the following burn-in time:*

$$T_0 := \left\lceil \frac{24}{\eta\mu} \right\rceil. \tag{25}$$

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

$$\mathbb{E}\|\nabla F(\theta_T)\|_2^2 \leq \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2(T+1)^2} + \frac{28\ \sigma_*^2}{T+1}. \tag{26}$$

We provide the proof of Theorem 5 in both the **LSN** and **ISC** cases; the **LSN** case corresponds to Theorem 1. In accordance with the discussion in §1, our nonasymptotic convergence rate upper bound (26) for the expected squared gradient norm consists of the addition of two terms. The first term, $\frac{\sigma_*^2}{T}$, corresponds to the *nonimprovable statistical error* depending on the noise variance at the minimizer. The second term, which is equivalent to $\frac{\|\nabla F(\theta_0)\|_2^2 T_0^2}{T^2}$, corresponds to the *bias* or *optimization error* that indicates the polynomial forgetting from the initialization. Theorem 5 copes with a wide range of step sizes η: fixing the number of online samples T, (26) 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 (26), we can achieve a tight upper bound on the sample complexity to achieve a statistical estimator of θ^* with gradient norm bounded by $O(\varepsilon)$:[10]

$$\begin{aligned} C_1(\varepsilon) &= \max\left\{\frac{74}{\eta_{max}\mu} \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{56\sigma_*^2}{\varepsilon^2}\right\} \\ &\asymp \begin{cases} \max\left\{\left(\frac{L}{\mu} + \frac{\ell_{\Xi}^2}{\mu^2}\right) \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the \textbf{LSN} case,} \\ \max\left\{\frac{L_{\max}}{\mu} \cdot \frac{\|\nabla F(\theta_0)\|_2}{\varepsilon}, \frac{\sigma_*^2}{\varepsilon^2}\right\}, & \text{for the \textbf{ISC} case.} \end{cases} \end{aligned} \tag{27}$$

10. Indeed, we choose T in Eq. (13) to be sufficiently large such that it satisfies the inequalities $T \geq T_0 = \lceil \frac{24}{\eta\mu} \rceil$, as well as $\frac{2700\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 T^2} \leq \frac{\varepsilon^2}{2}$ 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 (14) serves as a tight sample complexity upper bound.

In above, the step size $\eta = \eta_{max}$ is optimized as in (23) for the **LSN** and (24) 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 $C_1(\varepsilon)$ in either case is $\asymp \frac{\sigma_*^2}{\varepsilon^2}$ which matches the optimal statistical error up to universal constants, first among comparable literature in both cases.

Detailed proof. The rest of this subsection devotes to prove Theorem 5. It is straightforward to show first (26) automatically holds for $T < T_0$ since for these T, $\theta_T = \theta_0$ and hence $\mathbb{E}\|\nabla F(\theta_T)\|_2^2 = \mathbb{E}\|\nabla F(\theta_0)\|_2^2$, so we only need to prove the result for $T \geq T_0$.

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

$$\omega_{max} := \begin{cases} \frac{2\ell_{\Sigma}^2}{\mu^2}, & \text{for } \textbf{LSN} \text{ case}, \\ \frac{2L_{\max}}{\mu}, & \text{for } \textbf{ISC} \text{ case}. \end{cases} \tag{28}$$

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

$$z_t := v_t - \nabla F(\theta_{t-1}), \qquad \text{for } t \geq T_0. \tag{29}$$

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 (6), 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

$$\begin{aligned} tz_t = t\left(v_t - \nabla F(\theta_{t-1})\right) &= \varepsilon_t(\theta_{t-1}) + (t-1)(v_{t-1} - \nabla F(\theta_{t-2})) + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) \\ &= \varepsilon_t(\theta_{t-1}) + (t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})). \end{aligned} \tag{30}$$

Unwinding this relation recursively yields the decomposition

$$tz_t - T_0 z_{T_0} = \sum_{s=T_0+1}^{t} \varepsilon_s(\theta_{s-1}) + \sum_{s=T_0+1}^{t} (s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})). \tag{31}$$

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 6 (Recursion involving z_t) *Under the conditions of Theorem 5, for all $t \geq T_0 + 1$, we have*

$$t^2\mathbb{E}\|z_t\|_2^2 \leq (t-1)^2\mathbb{E}\|z_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + 2(t-1)^2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2. \tag{32a}$$

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

$$T_0^2\mathbb{E}\|v_{T_0}\|_2^2 - T_0^2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 = T_0^2\mathbb{E}\|z_{T_0}\|_2^2 = T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2. \tag{32b}$$

See §C.1.1 for the proof of this claim. Note we have $z_{T_0} = v_{T_0} - \nabla F(\theta_0)$ which is simply the arithmetic average of T_0 i.i.d. noise terms at θ_0, $\varepsilon_1(\theta_0), \ldots, \varepsilon_{T_0}(\theta_0)$.

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

Lemma 7 (Evolution of v_t) *Under the settings of Theorem 5, for any $\eta \in (0, \eta_{max}]$, we have*

$$t^2\mathbb{E}\|v_t\|_2^2 - 2t\mathbb{E}\langle v_t, \nabla F(\theta_{t-1})\rangle + \mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^2\,, \tag{33a}$$

and

$$\mathbb{E}\,\|tv_t - \nabla F(\theta_{t-1})\|_2^2 \leq (1-\eta\mu)\cdot(t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 - 2(t-1)^2\mathbb{E}\,\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\|_2^2\,, \tag{33b}$$

for all $t \geq T_0 + 1$.

See §C.1.2 for the proof of this claim.

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

Lemma 8 (Second moment of pointwise stochastic noise) *Under the conditions of Theorem 5, we have*

$$\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 \leq \omega_{max}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2, \qquad \text{for all } t \geq T_0 + 1. \tag{34}$$

See §C.2 for the proof of this claim.

Equipped with these three auxiliary results, we are now ready to prove Theorem 5.

Proof [Proof of Theorem 5] Our proof proceeds in two steps.

Step 1. We begin by applying the Cauchy-Schwarz and Young inequalities to the inner product $\langle v_t, \nabla F(\theta_{t-1})\rangle$. Doing so yields the upper bound

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

Taking the expectation of both sides and applying the bound (33a) from Lemma 7 yields

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

Moreover, since we have $\eta \leq \eta_{max} \leq \frac{1}{4\mu}$ under condition (12), 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 (32a) from Lemma 6 gives

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

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

$$T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2 - T_0^2\mathbb{E}\|z_{T_0}\|_2^2 - T_0^2\mathbb{E}\|v_{T_0}\|_2^2 \\ \leq \sum_{t=T_0+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right]. \quad (35)$$

Next, applying the result (32b) from Lemma 6 yields

$$\frac{T^2}{2}\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq T^2\mathbb{E}\|z_T\|_2^2 + T^2\mathbb{E}\|v_T\|_2^2 \\ \leq T_0^2\mathbb{E}\|z_{T_0}\|_2^2 + T_0^2\mathbb{E}\|v_{T_0}\|_2^2 + \sum_{t=T_0+1}^{T}\left[5\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right] \\ = T_0^2\|\nabla F(\theta_0)\|_2^2 + 2T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2 + 5\sum_{t=T_0+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{1}{\eta\mu}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2.$$

Following some algebra, we find that

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq \frac{2T_0^2\|\nabla F(\theta_0)\|_2^2 + 4T_0\mathbb{E}\|\varepsilon_{T_0}(\theta_0)\|_2^2}{T^2} \\ + \frac{10}{T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 + \frac{2}{\eta\mu T^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2. \quad (36)$$

Combining inequality (36) with the bound (34) from Lemma 8 gives

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

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

$$\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2 \leq \frac{1}{T^2}\left\{\alpha_1\mathbb{E}\|\nabla F(\theta_0)\|_2^2 + \alpha_2\sum_{t=T_0+1}^{T}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right\} + \frac{20\sigma_*^2}{T}, \quad (37)$$

where $\alpha_1 := (4\omega_{max} + 2T_0)\ T_0$ and $\alpha_2 := 10\omega_{max} + \frac{2}{\eta\mu}$.

Step 2. Based on the estimation bound (37), the proof of Theorem 5 relies on a bootstrapping argument in order to remove the dependence of the right-hand side of Eq. (37) on the quantity $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$.

Let $T^* \geq T_0 + 1$ be arbitrary. Telescoping the bound (37) over the iterates $T = T_0 + 1, \ldots, T^*$ yields

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

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

$$\sum_{T=T_0+1}^{T^*} \frac{1}{T^2} \overset{(i)}{\leq} \int_{T_0}^{T^*} \frac{d\tau}{\tau^2} \leq \frac{1}{T_0}, \qquad \text{and} \qquad \sum_{T=T_0+1}^{T^*} \frac{1}{T} \overset{(ii)}{\leq} \int_{T_0}^{T^*} \frac{d\tau}{\tau} = \log\Big(\frac{T^*}{T_0}\Big). \tag{38}$$

We clearly have

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

Moreover, by using the fact that $T^* \geq T$, interchanging the order of summation, and then using inequality (38)(i) again, we have

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

Finally, turning to the third quantity, we have $Q_3 \leq 20\sigma_*^2 \log\big(\frac{T^*}{T_0}\big)$, where we have used inequality (38)(ii). Putting together the pieces yields the upper bound

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

Eqs. (12) imply that, for either case under consideration, we have the bound $\omega_{max} \leq \frac{1}{\eta\mu}$, and, since $0 < \eta\mu \leq \frac{1}{4} < 1$, we have from (12) that $T_0 = \left\lceil \frac{24}{\eta\mu} \right\rceil \leq \frac{1}{\eta\mu}$, resulting in

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

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

$$\frac{1}{2} \sum_{t=T_0+1}^{T^*} \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \leq \frac{54}{\eta\mu} \mathbb{E}\|\nabla F(\theta_0)\|_2^2 + 20\sigma_*^2 \log\left(\frac{T^*}{T_0}\right). \tag{39}$$

Now substituting the inequality (39) back into the earlier bound (37) with $T^* = T$ allows us to obtain a bound on $\mathbb{E}\|\nabla F(\theta_{T-1})\|_2$. In particular, for any $T \geq T_0 + 1$, we have

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

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

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

Shifting the subscript forward by one yields Theorem 5. ■

C.1.1. PROOF OF LEMMA 6

The claim (32b) follows from the definition along with some basic probability. In order to prove the claim (32a), recall from the ROOT-SGD update rule for v_t in the first line of (6) that for $t \geq T_0 + 1$ we have:

$$tv_t = (t-1)v_{t-1} + t\nabla f(\theta_{t-1}; \xi_t) - (t-1)\nabla f(\theta_{t-2}; \xi_t). \tag{40}$$

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

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

where in the last step follows from Young's inequality. Computing the constants out completes the proof of the claim (32a).

C.1.2. PROOF OF LEMMA 7

Eq. (33a) follows in a straightforward manner by expanding the square and taking an expectation. As for the inequality (33b), from the update rule (6) for v_t, we have

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

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

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

Further rearranging yields

$$\begin{aligned} \mathbb{E}\|tv_t - \nabla F(\theta_{t-1})\|_2^2 &= (t-1)^2\mathbb{E}\|v_{t-1}\|_2^2 + 2(t-1)^2\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\|_2^2 \\ &\quad + 2\mathbb{E}\left\|\varepsilon_t(\theta_{t-1})\right\|_2^2 + 2(t-1)^2\mathbb{E}\left\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\rangle. \end{aligned} \tag{41}$$

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

Analysis in the LSN case: From L-Lipschitz smoothness of F in Assumption 1, we have

$$\begin{aligned} \left\langle v_{t-1},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle &= -\frac{1}{\eta}\left\langle \theta_{t-1} - \theta_{t-2},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle \\ &\leq -\frac{1}{\eta L}\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2. \end{aligned} \tag{42}$$

Now consider the inner product term $\langle v_{t-1},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle$ in Eq. (41). We split it into two terms, and upper bound them using equations (44) and (42) respectively. Doing so yields:

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

From the condition (24), we have $1 - \frac{3}{2}\eta\mu + 4\eta^2\ell_\Xi^2 \leq 1 - \eta\mu$, which completes the proof.

Analysis in the ISC case: We deal with the last summand in the last line of Eq. (41), where we use the iterated law of expectation to achieve

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

The update rule for v_t implies that $v_{t-1} = -\frac{\theta_{t-1}-\theta_{t-2}}{\eta}$ for all $t \geq T_0 + 1$. 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

$$\begin{aligned}\left\langle v_{t-1}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\rangle &= -\frac{1}{\eta}\left\langle \theta_{t-1}-\theta_{t-2}, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\rangle \\ &\leq -\frac{1}{\eta L_{\max}}\left\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\|_2^2,\end{aligned} \tag{43}$$

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

$$\begin{aligned}\left\langle v_{t-1}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle &= -\frac{1}{\eta}\left\langle \theta_{t-1}-\theta_{t-2}, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle \\ &\leq -\frac{\mu}{\eta}\|\theta_{t-1}-\theta_{t-2}\|_2^2 = -\eta\mu\|v_{t-1}\|_2^2.\end{aligned} \tag{44}$$

Plugging the bounds (43) and (44) into Eq. (41) yields

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

where in the last inequality relies on the fact that $\eta \in (0, \frac{1}{4L_{\max}}]$ (see Eq. (24)), leading to the bound (33b).

C.2. Proof of Lemma 8

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

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

$$\begin{aligned}\mathbb{E}\|\varepsilon_t(\theta_{t-1})\|_2^2 &\leq 2\mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\ &\leq 2\ell_\Xi^2\mathbb{E}\|\theta_{t-1}-\theta^*\|_2^2 + 2\mathbb{E}\|\varepsilon_t(\theta^*)\|_2^2 \\ &\leq \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + 2\sigma_*^2,\end{aligned} \tag{45}$$

which establishes the claim.

Analysis in the ISC case: Using Assumption 8 and standard inequalities for $L_{\max}$-smooth and convex functions yields

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

Taking expectations in this inequality and performing some algebra[11] yields

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

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

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

We also note that[12]

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

Finally, applying the argument of (45) yields the claim (34).

C.3. Intermediate result Proposition 1 and its proof

En route our proof of Theorem 2 we state and prove an intermediate upper-bound result for single-epoch version of ROOT-SGD. For our convenience we forgo tracking the universal constants (which can be change at each appearance) due to complications of our derivations.

Proposition 1 (Improved nonasymptotic upper bound, single-epoch ROOT-SGD) *Under Assumptions 1, 4, 5, 6, suppose that we run Algorithm 1 with step-size* $\eta \in \left(0, \frac{1}{56L} \wedge \frac{\mu}{64\ell_\Xi^2}\right]$. *Then for any* $T \geq 1$, *the iterate* θ_T *satisfies the bound*

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 - \frac{\sigma_*^2}{T} \leq C\Big\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\log T}{\eta\mu T} + \frac{\ell_\Xi^2 \log T}{\mu^2 T}\Big\}\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \\
+ \frac{C\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2T^2} + \frac{CL_1\left\|\nabla F(\theta_0)\right\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}
\end{aligned} \tag{46}$$

11. 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.
12. This proof strategy is forklore and appears elsewhere in the variance-reduction literature; see, e.g., the proof of Theorem 1 in Johnson and Zhang (2013), and also adopted by Nguyen et al. (2019, 2021).

A few remarks are in order. When setting $T \to \infty$ the leading-order term $(1 + \frac{C\ell_\Xi^2 \eta}{\mu})\frac{\sigma_*^2}{T}$ of the nonasymptotic bound (46) nearly matches the optimal statistical risk for the gradient norm with unit pre-factor when η is prescribed as positively small, and as will be seen later it matches the asymptotic Proposition 2 under a shared umbrella of assumptions. It can be observed that the dependence on the initial gradient norm $\|\nabla F(\theta_0)\|_2$ decays polynomially, which is generally unavoidable for single-epoch ROOT-SGD, as the gradient noise at the initial point θ_0 is also averaged along the iterates. However, as we will see anon, an improved guarantee can be obtained by appropriately re-starting the algorithm, leading to near-optimal guarantees in terms of the gradient norm. In addition, we note that the high-order terms of Eq. (46) contains terms that depend on the step-size η at opposite directions which demands a trade-off. We forgo optimizing the step-size as is the conduct in our multi-epoch result.

For the rest of §C.3 we prepare to prove Proposition 1. From the discussions in §2.2 we decomposes $\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2$ as the summation of three terms:

$$\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 = \mathbb{E}\|v_t - z_t\|_2^2 = \mathbb{E}\|v_t\|_2^2 + \mathbb{E}\|z_t\|_2^2 - 2\mathbb{E}\langle v_t,\, z_t\rangle. \tag{47}$$

En route our proof, we provide estimations for $\mathbb{E}\|tv_t\|_2^2$, $\mathbb{E}\|tz_t\|_2^2$ and $\mathbb{E}\langle tz_t,\, tv_t\rangle$ separately, where our main focus will be on bounding the cross term. On a very intuitive and high-level viewpoint, when comparing with the Polyak-Ruppert-Juditsky analysis, we can roughly think of the $(\eta t v_t : t \geq 0)$ process acts like a last-iterate SGD (as it is in the quadratic minimization case) and is *fast* and *small*. The tz_t process more resembles random walk at a slower rate driven by the same noise sequence. The two timescale intuitions beneath is that two fast-slow discounted random walks processes driven by the same noise has an inner product that is approximately the second moment of the fast process. In our case this results in the "asymptotically independence" of the two processes in the sense that $\mathbb{E}\langle tz_t,\, tv_t\rangle$ scales as $\mathbb{E}\|tv_t\|_2^2$, so $\nabla F(\theta_{t-1}) = v_t - z_t$ is approximately of the same scale as z_t in its first and second orders.

We first introduce the following lemma which is an essential part of the proof:

Lemma 9 (Sharp bound on v_t) *Under the setting of Theorem 1, there exists a universal constant $c > 0$, such that for $T \geq T_0 + 1$, we have:*

$$\mathbb{E}\|v_T\|_2^2 \leq \frac{c\sigma_*^2}{\eta\mu T^2} + \frac{c}{\eta^4\mu^4 T^4}\|\nabla F(\theta_0)\|_2^2. \tag{48}$$

We defer the proof of Lemma 9 to §C.3.1. This lemma, along with Theorem 1, helps conclude the following bound on z_t that has a leading-order term of near-unity pre-factor, that is, $(1 + o(1))\frac{\sigma_*}{\sqrt{t}}$:

Lemma 10 (Sharp bound on z_t) *Under settings of Theorem 1, the following bounds hold true for $T \geq T_0 + 1$:*

$$\mathbb{E}\|z_T\|_2^2 - \frac{\sigma_*^2}{T} \leq c\Big\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\Big\}\frac{\sigma_*^2}{T} + c\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\|\nabla F(\theta_0)\|_2 + c\frac{T_0^2}{T^2}\|\nabla F(\theta_0)\|_2^2, \tag{49}$$

for some universal constant $c > 0$.

See §C.3.2 for the proof of this lemma.

Finally, we need the following lemma, which bounds the cross term $\mathbb{E}\langle v_t,\, z_t\rangle$. Under the Lipschitz condition on the Hessian matrix and additional moment conditions, this lemma provides significant sharper bound than the naïve bound obtained by applying the Cauchy-Schwartz inequality and invoking the previous two lemmas.

Lemma 11 (Sharp bound on the cross term) *Under settings of Theorem 1, we have the following bound for any $T \geq T_0 + 1$:*

$$\left|\mathbb{E}\left\langle v_T,\, z_T\right\rangle\right| \leq c\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 T^4}\right)\log T + cL_1\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right), \quad (50)$$

for some universal constant $c > 0$.

See §C.3.3 for the proof of this lemma.

Taking the aforementioned lemmas as given, we are ready to prove the sharp bound. In particular, by substituting these three lemmas into the decomposition (47), we have the following bound

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^2 - \frac{\sigma_*^2}{T} = \mathbb{E}\left\|v_T - z_T\right\|_2^2 - \frac{\sigma_*^2}{T} = \left(\mathbb{E}\left\|z_T\right\|_2^2 - \frac{\sigma_*^2}{T}\right) + \mathbb{E}\left\|v_T\right\|_2^2 - 2\mathbb{E}\left\langle v_T,\, z_T\right\rangle \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + C\left(\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\left\|\nabla F(\theta_0)\right\|_2 + \frac{\ell_\Xi^2}{\mu^2}\cdot\frac{T_0}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right) \\
&\quad + C\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4 T^4}\right)\log T + 6C_0L_1\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right) \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \underbrace{\frac{\ell_\Xi}{\mu\sqrt{T}}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \\
&\quad + C\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \underbrace{\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\left\|\nabla F(\theta_0)\right\|_2}\right) + C\frac{L_1\left\|\nabla F(\theta_0)\right\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}.
\end{aligned}$$

Absorbing the bracketed cross terms into corresponding sum of the squares, this gives Eq. (46) and concludes Proposition 1.

C.3.1. PROOF OF LEMMA 9

Our main technical tools is the following Lemma 12, which recursively bound the second moments of v_t:

Lemma 12 *Under the setting of Theorem 1, we have the following bound for $t \geq T_0 + 1$*

$$t^2\mathbb{E}\left\|v_t\right\|_2^2 \leq \left(1 - \frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{10}{\eta\mu}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + 4\sigma_*^2. \quad (51)$$

See §E.3.2 for the proof of this lemma.

On the other hand, invoking Theorem 1, we have that

$$\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \le \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2 t^2} + \frac{28\ \sigma_*^2}{t}, \qquad \text{for } t \ge T_0 + 1.$$

Now, to combine everything together, we conclude from (13) and (51) that

$$\begin{aligned} t^2\mathbb{E}\left\|v_t\right\|_2^2 &\le \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{10}{\eta\mu}\left[\frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2} + \frac{28\ \sigma_*^2}{t}\right] + 4\sigma_*^2 \\ &\le \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + c\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3t^2} + c\sigma_*^2. \end{aligned} \tag{52}$$

Multiplying both sides by t^2, we obtain that

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

for time index t satisfying $t \ge T_0 \ge \frac{6}{\eta\mu}$. This gives, by solving the recursion,

$$\begin{aligned} T^4\mathbb{E}\left\|v_T\right\|_2^2 &\le \left(1-\frac{\eta\mu}{6}\right)^{T-T_0} T_0^4\mathbb{E}\left\|v_{T_0}\right\|_2^2 + c\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^3\mu^3} + \sigma_*^2T^2\right) \\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-T_0} T_0^4\mathbb{E}\left\|v_{T_0}\right\|_2^2 + 6c\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4} + 6c\frac{\sigma_*^2}{\eta\mu}T^2. \end{aligned}$$

It suffices to bound the initial condition $\mathbb{E}\left\|v_{T_0}\right\|_2^2$. Recall that $v_{T_0} = \frac{1}{T_0}\sum_{s=1}^{T_0}\nabla f(\theta_0;\xi_s)$, which is average of i.i.d. random vectors. It immediately follows from Assumptions 2 and 3 that:

$$\mathbb{E}\left\|v_{T_0}\right\|_2^2 \le \|\nabla F(\theta_0)\|_2^2 + \frac{2\sigma_*^2}{T_0} + \frac{2\ell_\Xi^2\,\|\nabla F(\theta_0)\|_2^2}{\mu^2T_0}.$$

Putting them together, we complete the proof of this lemma.

C.3.2. PROOF OF LEMMA 10

Recalling that the recursive update rule of z_t reveals an underlying martingale structure

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

Adding and subtracting the $\varepsilon_t(\theta^*)$ term in the above display we express the noise increment as

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

In words, the increment of tz_t splits into two parts: the additive part $\varepsilon_t(\theta^*)$ and the multiplicative part ζ_t. Taking expectation on the squared norm in above and using the property of square-integrable martingales, we have via further expanding the square on the right hand

$$t^2\mathbb{E}\left\|z_t\right\|_2^2 - (t-1)^2\mathbb{E}\left\|z_{t-1}\right\|_2^2 = \mathbb{E}\left\|\varepsilon_t(\theta^*) + \zeta_t\right\|_2^2 = \mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2 + \mathbb{E}\left\|\zeta_t\right\|_2^2 + 2\mathbb{E}\langle\varepsilon_t(\theta^*),\, \zeta_t\rangle.$$

Telescoping the above equality for $t = T_0 + 1, \ldots, T$ gives

$$T^2\mathbb{E}\left\|z_T\right\|_2^2 - T_0^2\mathbb{E}\left\|z_{T_0}\right\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2 + \sum_{t=T_0+1}^{T}\mathbb{E}\left\|\zeta_t\right\|_2^2 + 2\sum_{t=T_0+1}^{T}\mathbb{E}\langle\varepsilon_t(\theta^*),\, \zeta_t\rangle.$$

By Assumption 5, we have $\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2 = \sigma_*^2$. For the additional noise ζ_t, Young's inequality leads to the bound

$$\begin{aligned}
\mathbb{E}\left\|\zeta_t\right\|_2^2 &\leq \left(\ell_\Xi(t-1)\sqrt{\mathbb{E}\left\|\theta_{t-1} - \theta_{t-2}\right\|_2^2} + \ell_\Xi\sqrt{\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^2}\right)^2 \\
&\leq \left(\eta\ell_\Xi(t-1)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2} + \frac{\ell_\Xi}{\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}\right)^2 \\
&\leq 2\eta^2\ell_\Xi^2(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2. \qquad (53)
\end{aligned}$$

It remains to bound the summation of the cross term. Observing that:

$$\begin{aligned}
\sum_{t=T_0+1}^{T}\mathbb{E}\langle\varepsilon_t(\theta^*),\, \zeta_t\rangle &= \sum_{t=T_0+1}^{T}\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*) + (t-1)\left(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right)\rangle \\
&= \sum_{t=T_0+1}^{T}\Big\{t\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle\Big\}.
\end{aligned}$$

Since the random samples $(\xi_t)_{t\geq 1}$ are i.i.d. and the iterate θ_{t-2} is independent of the sample ξ_{t-1}, we have that

$$\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle = \mathbb{E}\langle\varepsilon_{t-1}(\theta^*),\, \varepsilon_{t-1}(\theta_{t-2}) - \varepsilon_{t-1}(\theta^*)\rangle.$$

Consequently, we can re-write the quantity of interests as a telescope sum, leading to the following identity:

$$\begin{aligned}
\sum_{t=T_0+1}^{T}\mathbb{E}\langle\varepsilon_t(\theta^*),\, \zeta_t\rangle &= \sum_{t=T_0+1}^{T}\Big\{t\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-2}) - \varepsilon_t(\theta^*)\rangle\Big\} \\
&= \sum_{t=T_0+1}^{T}\Big\{t\mathbb{E}\langle\varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle - (t-1)\mathbb{E}\langle\varepsilon_{t-1}(\theta^*),\, \varepsilon_{t-1}(\theta_{t-2}) - \varepsilon_{t-1}(\theta^*)\rangle\Big\} \\
&= T\cdot\mathbb{E}\langle\varepsilon_T(\theta^*),\, \varepsilon_T(\theta_{T-1}) - \varepsilon_T(\theta^*)\rangle - T_0\cdot\mathbb{E}\langle\varepsilon_{T_0}(\theta^*),\, \varepsilon_{T_0}(\theta_{T_0-1}) - \varepsilon_{T_0}(\theta^*)\rangle.
\end{aligned}$$

In order to bound the inner product terms, we invoke the Cauchy-Schwartz inequality and Assumption 3. For each $t \geq T_0$, we have that

$$\begin{aligned}|t \cdot \mathbb{E}\langle \varepsilon_t(\theta^*),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\rangle| &\leq t \cdot \sqrt{\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2} \cdot \sqrt{\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\right\|_2^2} \\ &\leq t\sigma_* \ell_\Xi \sqrt{\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^2} \leq \frac{t\sigma_*\ell_\Xi}{\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}. \quad (54)\end{aligned}$$

Plugging $t = T_0$ and $t = T$ into Eq. (54) separately and combining with Eq. (53), we have that

$$\begin{aligned}T^2\mathbb{E}\left\|z_T\right\|_2^2 \leq T_0^2\mathbb{E}\left\|z_{T_0}\right\|_2^2 + (T-T_0)\sigma_*^2 + 2\eta^2\ell_\Xi^2 \sum_{t=T_0+1}^{T}(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{2\ell_\Xi^2}{\mu^2}\sum_{t=T_0+1}^{T}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 \\ + \frac{2\ell_\Xi\sigma_*}{\mu}\left(T\sqrt{\mathbb{E}\|\nabla F(\theta_{T-1})\|_2^2} + T_0\left\|\nabla F(\theta_0)\right\|_2\right). \quad (55)\end{aligned}$$

The above bound involves the second moments of the vectors v_t and $\nabla F(\theta_t)$. We recall the following bounds from Theorem 1 and Lemma 9, for each $t \geq T_0$:

$$\begin{aligned}\mathbb{E}\left\|v_t\right\|_2^2 &\leq \frac{c\sigma_*^2}{\eta\mu t^2} + \frac{c}{\eta^4\mu^4 t^4}\left\|\nabla F(\theta_0)\right\|_2^2, \quad \text{and} \\ \mathbb{E}\left\|\nabla F(\theta_t)\right\|_2^2 &\leq \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2 t^2} + \frac{c\sigma_*^2}{t}.\end{aligned}$$

Substituting these bounds to Eq. (55), we note that

$$\begin{aligned}\sum_{t=T_0+1}^{T}(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 &\leq \frac{c\sigma_*^2 T}{\eta\mu} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4 T_0}, \quad \text{and} \\ \sum_{t=T_0+1}^{T}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 &\leq c\sigma_*^2\log\left(\frac{T}{T_0}\right) + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2 T_0}.\end{aligned}$$

Finally, for the burn-in period we note that $z_{T_0} = \sum_{s=1}^{T_0}\varepsilon_s(\theta_0)$ is a sum of T_0 i.i.d. random vectors, and hence the following estimation holds

$$\begin{aligned}&T_0^2\mathbb{E}\left\|z_{T_0}\right\|_2^2 = T_0\mathbb{E}\left\|\varepsilon_1(\theta_0)\right\|_2^2 \\ &\leq T_0\mathbb{E}\left\|\varepsilon_1(\theta_0) - \varepsilon_1(\theta^*)\right\|_2^2 + 2T_0\sqrt{\mathbb{E}\left\|\varepsilon_1(\theta^*)\right\|_2^2\mathbb{E}\left\|\varepsilon_1(\theta_0) - \varepsilon_1(\theta^*)\right\|_2^2} + T_0\mathbb{E}\left\|\varepsilon_1(\theta^*)\right\|_2^2 \\ &\leq T_0\ell_\Xi^2\left\|\theta_0 - \theta^*\right\|_2^2 + 2T_0\ell_\Xi\sigma_*\left\|\theta_0 - \theta^*\right\|_2 + T_0\sigma_*^2 \leq \frac{T_0\ell_\Xi^2}{\mu^2}\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{2T_0\ell_\Xi\sigma_*}{\mu}\left\|\nabla F(\theta_0)\right\|_2 + T_0\sigma_*^2.\end{aligned}$$

Some algebra yields (49) and hence the whole Lemma 10.

C.3.3. PROOF OF LEMMA 11

First, by Cauchy-Schwartz inequality, we can easily observe that:

$$|\mathbb{E}\langle v_T,\, z_T\rangle| \leq \sqrt{\mathbb{E}\left\|v_T\right\|_2^2} \cdot \sqrt{\mathbb{E}\left\|z_T\right\|_2^2} \leq c\Big(\frac{\sigma_*}{T\sqrt{\eta\mu}} + \frac{\left\|\nabla F(\theta_0)\right\|_2}{\eta^2\mu^2T^2}\Big) \cdot \Big(\frac{\sigma_*}{\sqrt{T}} + \frac{\left\|\nabla F(\theta_0)\right\|_2}{\eta\mu T}\Big).$$

So for $T \le cT_0 \log T_0$, the conclusion of this lemma is automatically satisfied. For the rest of this section, we assume that $\frac{T}{\log T} > cT_0$ for some universal constant $c > 0$.

The proof requires some bounds on the fourth moment of the stochastic process defined by the algorithm. In particular, we need the following two lemmas. The first lemma is analogous to the bound in Theorem 1:

Lemma 13 (Higher-order-moment bound on $\nabla F(\theta_{t-1})$) *Suppose Assumptions 1, 5 and 6 hold. Let the step-size $\eta \le \frac{1}{56L} \wedge \frac{\mu}{64\ell_\Xi^2}$ and the burn-in time $T_0 = \left\lceil \frac{24}{\eta\mu} \right\rceil$. Then for any $T \ge T_0$, the estimator θ_T produced by the ROOT-SGD algorithm satisfies the bound*

$$\left(\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4\right)^{1/2} \le \frac{140\widetilde{\sigma_*}^2}{T+1} + \frac{60\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2(T+1)^2}. \tag{56}$$

Proof can be found in §E.1.

We also need a lemma on the fourth-moment bound of v_t, analogous to Lemma 9:

Lemma 14 (sharp higher-order-moment bound on v_t) *Under the setting of Proposition 1 we have the following bound for $T \ge T_0 + 1$*

$$\sqrt{\mathbb{E}\left\|v_T\right\|_2^4} \le \frac{4484\widetilde{\sigma_*}^2}{\eta\mu T^2} + \frac{1359375}{\eta^4\mu^4T^4}\left\|\nabla F(\theta_0)\right\|_2^2. \tag{57}$$

Proof can be found in §E.2.

Taking these two lemmas as given, we proceed with the proof. Following the two-time-scale intuition discussed in Section 2.2, the process v_t moves faster than the averaging process z_t. Therefore, it is reasonable to expect the correlation between v_t and $z_{t-\widetilde{T}^*}$ to be small, for sufficiently large time window $\widetilde{T}^* > 0$. For the rest of this section, we choose the window size:

$$\widetilde{T}^* = \frac{c}{\mu\eta}\log T, \quad \text{for some universal constant } c > 0. \tag{58}$$

Since we have assumed without loss of generality that $\frac{T}{\log T} > cT_0 = \frac{24c}{\eta\mu}$, the window size guarantees the relation $T - \widetilde{T}^* > T/2$.

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

$$|\mathbb{E}\langle tz_t,\, v_t\rangle| \le (t-\widetilde{T}^*)\underbrace{\left|\mathbb{E}\langle z_{t-\widetilde{T}^*},\, v_t\rangle\right|}_{=:I_1} + \underbrace{\left|\mathbb{E}\langle tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*},\, v_t\rangle\right|}_{=:I_2}. \tag{59}$$

For bounding the term I_2, we make use of the recursive rule of tz_t to obtain the bound

$$\mathbb{E}\left\|tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2 = \mathbb{E}\left\|\sum_{s=t-\widetilde{T}^*+1}^{t}\Big\{(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})) + \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*) + \varepsilon_s(\theta^*)\Big\}\right\|_2^2$$

$$\le \sum_{s=t-\widetilde{T}^*+1}^{t}\left\{(s-1)^2\eta^2\ell_\Xi^2\mathbb{E}\left\|v_{s-1}\right\|_2^2 + \frac{\ell_\Xi^2}{\mu^2}\mathbb{E}\left\|\nabla F(\theta_{s-1})\right\|_2^2 + \sigma_*^2\right\} \le \widetilde{T}^*\cdot\left(\sigma_*^2 + \frac{\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3t^2}\right).$$

Consequently, we have the bound

$$\begin{aligned}\left|\mathbb{E}\langle tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}, v_t\rangle\right| &\le \sqrt{\mathbb{E}\left\|tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|v_t\right\|_2^2} \\ &\le c\sqrt{\widetilde{T}^*}\left(\frac{\sigma_*}{\sqrt{\eta\mu t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2t^2}\right)\left(\sigma_*+\frac{\|\nabla F(\theta_0)\|_2}{\eta^{3/2}\mu^{3/2}t}\right) \le c\left(\frac{\sigma_*^2}{\eta\mu t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4t^3}\right)\sqrt{\log t}.\end{aligned}$$

The bound for the term I_1 in the decomposition (59) is given by the following analysis: law of iterated expectations gives

$$\left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, v_t\rangle\right| = \left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, \mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\rangle\right| \le \sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2}, \tag{60}$$

where the last inequality comes from applying the Cauchy-Schwarz inequality.

The second moment for $z_{t-\widetilde{T}^*}$ is relatively easy to estimate using Lemma 10. It suffices to study the conditional expectation $\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})$. We claim the following bound:

$$\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} \le \frac{cL_1}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2t}\right). \tag{61}$$

We prove this inequality at the end of this section. Taking this bound as given, we now proceed with the proof for Lemma 11.

Bringing this back to the inequality (60) and by utilizing the z_t bound by Lemma 10, we have

$$\mathbb{E}\left\|z_t\right\|_2^2 \le C\left(\frac{\sigma_*^2}{t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2}\right),$$

and thus

$$\begin{aligned}\left|\mathbb{E}\langle z_{t-\widetilde{T}^*}, v_t\rangle\right| &\le \sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2^2} \\ &\le \frac{cL_1}{\mu}\left(\frac{\sigma_*}{\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu t}\right)\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu t}}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta\mu^2t}\right) \\ &\le \frac{cL_1}{\mu}\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{3/2}t^2}+\frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{9/2}t^{7/2}}\right).\end{aligned}$$

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

$$\begin{aligned}&\left|\mathbb{E}\langle tz_t, v_t\rangle\right| \\ &\le c(t-\widetilde{T}^*)\frac{L_1}{\mu}\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{3/2}t^2}+\frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{9/2}t^{7/2}}\right)+c\left(\frac{\sigma_*^2}{\eta\mu t}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4t^3}\right)\sqrt{\log t}.\end{aligned} \tag{62}$$

We conclude by dividing both sides of Eq. (62) by T and arrive at the following bound:

$$\begin{aligned}&\left|\mathbb{E}\langle v_T, z_T\rangle\right| \\ &\le c\left(\frac{\sigma_*^2}{\eta\mu T^2}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right)\sqrt{\log T}+cL_1\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2}+\frac{\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right).\end{aligned}$$

This finishes our bound on the cross term and conclude Lemma 11.

Proof of Eq (61): We note the following expansion:

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

which leads to the following bound under the Lipschitz continuity condition for the Hessians (Assumption 4):

$$\begin{aligned}
&\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - \nabla^2 F(\theta^*)(\theta_{t-1} - \theta_{t-2})\right\|_2 \\
&= \int_0^1 \left\|\left(\nabla^2 F\left(\lambda\theta_{t-2} + (1-\lambda)\theta_{t-1}\right) - \nabla^2 F(\theta^*)\right)(\theta_{t-1} - \theta_{t-2})\right\|_2 d\lambda \\
&\leq \eta L_1 \left\|v_{t-1}\right\|_2 \int_0^1 \left\|\lambda(\theta_{t-2} - \theta^*) + (1-\lambda)(\theta_{t-1} - \theta^*)\right\|_2 d\lambda \\
&\leq \eta L_1 \left\|v_{t-1}\right\|_2 \cdot \max\left(\left\|\theta_{t-1} - \theta^*\right\|_2, \left\|\theta_{t-2} - \theta^*\right\|_2\right). \qquad (63)
\end{aligned}$$

Since $H^* = \nabla^2 F(\theta^*)$ we have

$$\begin{aligned}
&t\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2 \\
&= \left\|\mathbb{E}\left((t-1)\left(v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right) + \nabla F(\theta_{t-1}) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2 \\
&= \left\|\mathbb{E}\left((t-1)\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2})\right.\right.\right. \\
&\quad \left.\left.\left. + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) + \nabla F(\theta_{t-1}) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2 \\
&\leq \left\|\mathbb{E}\left((t-1)\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2 \\
&\quad + \left\|\mathbb{E}\left((t-1)\left(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2. \qquad (64)
\end{aligned}$$

Further by rearranging the terms, and dividing both sides by $(t-1)$, we obtain

$$\begin{aligned}
&\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2 \\
&\leq \left\|\mathbb{E}\left(v_{t-1} + H^*(\theta_{t-1} - \theta_{t-2}) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \left\|\mathbb{E}\left(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) - H^*(\theta_{t-1} - \theta_{t-2})\right) \mid \mathcal{F}_{t-\widetilde{T}^*}\right\|_2 \\
&\leq (1-\eta\mu)\left\|\mathbb{E}\left(v_{t-1} \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2 + \eta L_1 \mathbb{E}\left(\left\|v_{t-1}\right\|_2 \cdot \max\left(\left\|\theta_{t-1} - \theta^*\right\|_2, \left\|\theta_{t-2} - \theta^*\right\|_2\right) \mid \mathcal{F}_{t-\widetilde{T}^*}\right),
\end{aligned}$$

where in the last inequality we apply the result in Eq. (63). Next by calculating the second moment of both the RHS and the LHS of the above quantity and the Hölder's inequality, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} \\
&\leq (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}(v_{t-1} \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} + \eta L_1 \sqrt{\mathbb{E}\left\|\mathbb{E}\left(\left\|v_{t-1}\right\|_2^2 \cdot \left(\left\|\theta_{t-1} - \theta^*\right\|_2^2 + \left\|\theta_{t-2} - \theta^*\right\|_2^2\right) \mid \mathcal{F}_{t-\widetilde{T}^*}\right)\right\|_2^2} \\
&\leq (1-\eta\mu)\sqrt{\mathbb{E}\left\|\mathbb{E}(v_{t-1} \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} + \eta L_1 \left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/4}\left\{\left(\mathbb{E}\left\|\theta_{t-1} - \theta^*\right\|_2^4\right)^{1/4} + \left(\mathbb{E}\left\|\theta_{t-2} - \theta^*\right\|_2^4\right)^{1/4}\right\}.
\end{aligned}$$

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

$$\begin{aligned}&\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} \\ &\leq (1-\eta\mu)^{\widetilde{T}^*}\mathbb{E}\left\|v_{t-\widetilde{T}^*}\right\|_2^2 + \frac{L_1}{\mu}\max_{t-\widetilde{T}^*\leq s\leq t}\left(\mathbb{E}\left\|v_s\right\|_2^4\right)^{1/4}\cdot\max_{t-\widetilde{T}^*\leq s\leq t}\left(\mathbb{E}\left\|\theta_{t-2}-\theta^*\right\|_2^4\right)^{1/4}.\end{aligned} \tag{65}$$

We recall from Lemmas 13 and 14 the following

$$\begin{aligned}\left(\mathbb{E}\left\|v_T\right\|_2^4\right)^{1/2} &\leq C\left(\frac{\widetilde{\sigma_*^2}}{\eta\mu T^2}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^4\mu^4T^4}\right), \qquad \text{and} \\ \left(\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4\right)^{1/2} &\leq C\left(\frac{\widetilde{\sigma_*^2}}{T}+\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2}\right).\end{aligned}$$

Bringing this into Eq. (65) and we have that

$$\begin{aligned}\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} &\leq (1-\eta\mu)^{\widetilde{T}^*}\mathbb{E}\left\|v_{t-\widetilde{T}^*}\right\|_2^2 \\ &+\frac{cL_1}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu}(t-\widetilde{T}^*)}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2(t-\widetilde{T}^*)^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t-\widetilde{T}^*}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2(t-\widetilde{T}^*)}\right).\end{aligned}$$

Substituting with the window size $\widetilde{T}^*$ defined in Eq (58), the above inequality reduces as follows:

$$\sqrt{\mathbb{E}\left\|\mathbb{E}(v_t \mid \mathcal{F}_{t-\widetilde{T}^*})\right\|_2^2} \leq \frac{cL_1}{\mu}\left(\frac{\widetilde{\sigma_*}}{\sqrt{\eta\mu t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta^2\mu^2t^2}\right)\left(\frac{\widetilde{\sigma_*}}{\mu\sqrt{t}}+\frac{\|\nabla F(\theta_0)\|_2}{\eta\mu^2t}\right).$$

C.4. Proof of Theorem 2

Utilizing the intermediate Proposition 1 in §C.3, we now aim to improve the dependency on initialization and turn to the proof of our multi-epoch nonasymptotic result. Invoking Eq. (56) in Lemma 13, we obtain for $b = 1, 2, \cdots, B$ the bound for $T^\flat \geq cT_0$:

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2 &\leq \frac{1}{e^2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2+\frac{c\sigma_*^2}{T^\flat}, \qquad \text{and} \\ \sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^4} &\leq \frac{1}{e^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^4}+\frac{c\widetilde{\sigma_*^2}}{T^\flat},\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^2 &\leq \frac{c\sigma_*^2}{T^\flat}+e^{-2B}\left\|\nabla F(\theta_0)\right\|_2^2, \qquad \text{and} \\ \sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4} &\leq \frac{c\widetilde{\sigma_*^2}}{T^\flat}+e^{-2B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}.\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^2 \leq e^{-2B}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat} \leq \frac{c'\sigma_*^2}{T^\flat}, && \text{and} \\
&\sqrt{\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4} \leq e^{-2B}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat} \leq \frac{c'\widetilde{\sigma_*}^2}{T^\flat}, && \text{and} \\
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^3 \leq \left(\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2^4\right)^{\frac{3}{4}} \leq \frac{c'\widetilde{\sigma_*}^3}{(T^\flat)^{3/2}}, && \text{and} \\
&\mathbb{E}\left\|\nabla F(\theta_0^{(B+1)})\right\|_2 \leq \frac{c\sigma_*}{(T^\flat)^{1/2}},
\end{aligned}$$

where constants c, c' change from line to line. Substituting this initial condition into the bound (46), we obtain the final bound:

$$\begin{aligned}
&\mathbb{E}\left\|\nabla F(\theta_T^{(B+1)})\right\|_2^2 - \frac{\sigma_*^2}{T} \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} \\
&\quad + C\left(\frac{\|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^2T^2} + \frac{\ell_\Xi\sigma_*\|\nabla F(\theta_0)\|_2}{\eta\mu^2T^2}\right) + \frac{CL_1\|\nabla F(\theta_0)\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}} \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + C\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{L_1\widetilde{\sigma_*}^3}{\eta^2\mu^4T^{7/2}}\right) \\
&\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu T^{1/2}} + \frac{\log T}{\eta\mu T} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right)\frac{\sigma_*^2}{T} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2},
\end{aligned}$$

which proves the bound (17).

Finally, substituting T by the final epoch length $n - BT^\flat$ and adopt similar reasoning as the previous one, we arrive at the conclusion:

$$\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 - \frac{\sigma_*^2}{n} \leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu n^{1/2}} + \frac{\log n}{\eta\mu n} + \frac{\log\left(\frac{n}{T_0}\right)\ell_\Xi^2}{\mu^2 n}\right)\frac{\sigma_*^2}{n} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}n^2},$$

which proves the bound (17). Plugging η as given by $\eta = \frac{c}{\ell_\Xi n^{1/2}} \wedge \frac{1}{4L}$ with $c = 0.49$, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_n^{\text{final}})\right\|_2^2 &\leq C\left(\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu n^{1/2}} + \frac{\log n}{\eta\mu n} + \frac{\log\left(\frac{n}{T_0}\right)\ell_\Xi^2}{\mu^2 n}\right)\frac{\sigma_*^2}{n} + \frac{CL_1\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}n^2} \\
&\leq C\left(\log\left(\frac{en}{T_0}\right)\left(\frac{\ell_\Xi}{\mu\sqrt{n}}\right) + \frac{L}{\mu n}\right)\frac{\sigma_*^2}{n} + \frac{C(\ell_\Xi^{1/2}n^{1/4} + L^{1/2})L_1\widetilde{\sigma_*}^3}{\mu^{5/2}n^2}.
\end{aligned}$$

This concludes (18) and hence Theorem 2.

C.5. Proof of Corollary 3

The proof consists of two parts: bounds on the mean-squared error $\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2$ and bounds on the expected objective gap $\mathbb{E}\big[F(\theta_T) - F(\theta^*)\big]$. Two technical lemmas are needed in the proofs for both cases.

The first lemma is analogous to Lemma 10, which provides a sharp bound on Gz_t for any matrix $G \in \mathbb{R}^{d\times d}$.

Lemma 15 *Under settings of Theorem 1, for any matrix $G \in \mathbb{R}^{d\times d}$, the following bounds hold true for $T \geq T_0 + 1$:*

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left(G\Sigma^* G^\top\right) + c\|\!|G|\!\|_{op}^2\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{\log\left(\frac{T}{T_0}\right)\ell_\Xi^2}{\mu^2 T}\right\}\frac{\sigma_*^2}{T} + c\|\!|G|\!\|_{op}^2\left\{\frac{\ell_\Xi\sigma_*}{\mu}\cdot\frac{T_0}{T^2}\left\|\nabla F(\theta_0)\right\|_2 + \frac{T_0^2}{T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right\}. \quad (66)$$

for some universal constant $c > 0$.

The second lemma is analogous to Lemma 11, and provides sharp bound on the cross term $\mathbb{E}\langle Gz_t,\, Gv_t\rangle$.

Lemma 16 *Under settings of Theorem 1, we have the following bound for any $T \geq T_0 + 1$:*

$$\left|\mathbb{E}\big\langle Gv_T,\, Gz_T\big\rangle\right| \leq c\|\!|G|\!\|_{op}^2\left(\frac{\sigma_*^2}{\eta\mu T^2} + \frac{\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4 T^4}\right)\log T + c\|\!|G|\!\|_{op}^2 L_1\left(\frac{\widetilde{\sigma_*}^3}{\eta^{1/2}\mu^{5/2}T^2} + \frac{\left\|\nabla F(\theta_0)\right\|_2^3}{\eta^{7/2}\mu^{11/2}T^{7/2}}\right), \quad (67)$$

for some universal constant $c > 0$.

See §E.4 for the proof of both lemmas.

Taking these two lemmas as given, we now proceed with the proof of Corollary 3.

C.5.1. PROOF OF THE MSE BOUND (20a)

We start with the following decomposition of the gradient:

$$\nabla F(\theta_T) = \int_0^1 \nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big)(\theta_T - \theta^*)d\rho,$$

which leads to the following bound under Assumption 4:

$$\begin{aligned}\left\|(H^*)^{-1}\nabla F(\theta_T) - (\theta_T - \theta^*)\right\|_2 &\leq \int_0^1 \left\|(H^*)^{-1}\left(\nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\right\|_2 d\rho \\ &\leq \frac{L_1}{\lambda_{\min}(H^*)}\left\|\theta_T - \theta^*\right\|_2^2 \leq \frac{L_1}{\lambda_{\min}(H^*)\mu^2}\left\|\nabla F(\theta_T)\right\|_2^2. \quad (68)\end{aligned}$$

We can then upper bound the mean-squared error using the processes $(z_t)_{t\geq T_0}$ and $(v_t)_{t\geq T_0}$:

$$\begin{aligned}\mathbb{E}\left\|\theta_T-\theta^*\right\|_2^2 &\leq \mathbb{E}\left(\left\|(H^*)^{-1}\nabla F(\theta_T)\right\|_2+\frac{L_1}{\mu^2\lambda_{\min}(H^*)}\left\|\nabla F(\theta_T)\right\|_2^2\right)^2\\ &\leq \mathbb{E}\left\|(H^*)^{-1}\big(v_{T+1}-z_{T+1}\big)\right\|_2^2+\frac{2L_1}{\lambda_{\min}(H^*)^2\mu^2}\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^3\\ &\quad+\frac{L_1^2}{\lambda_{\min}(H^*)^2\mu^4}\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4. \end{aligned}\tag{69}$$

The first term in the bound (69) admits the following decomposition:

$$\begin{aligned}&\mathbb{E}\left\|(H^*)^{-1}\big(z_{T+1}-v_{T+1}\big)\right\|_2^2\\ &=\mathbb{E}\left\|(H^*)^{-1}z_{T+1}\right\|_2^2+\mathbb{E}\left\|(H^*)^{-1}v_{T+1}\right\|_2^2-2\mathbb{E}\langle(H^*)^{-1}z_{T+1},\,(H^*)^{-1}v_{T+1}\rangle.\end{aligned}$$

Note that the re-starting scheme in Algorithm 2 gives the initial conditions:

$$\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2\leq\frac{c\sigma_*^2}{T_0},\qquad\text{and}\qquad\left(\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\right)^{1/2}\leq\frac{c\widetilde{\sigma_*}^2}{T_0}.\tag{70}$$

Using these initial conditions, and invoking the Lemma 15 with test matrix $G=(H^*)^{-1}$, we obtain the bound:

$$\mathbb{E}\left\|(H^*)^{-1}z_T\right\|_2^2\leq\frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right)+\frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu}+\frac{\ell_\Xi}{\mu\sqrt{T}}+\frac{T_0}{T}\right\}\frac{\sigma_*^2\log T}{T}.$$

Similarly, invoking Lemma 16 with test matrix $G=(H^*)^{-1}$, we have that:

$$\left|\mathbb{E}\big\langle(H^*)^{-1}v_T,\,(H^*)^{-1}z_T\big\rangle\right|\leq\frac{c\sigma_*^2\sqrt{\log T}}{\lambda_{\min}(H^*)^2\eta\mu T^2}+\frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2}.$$

For the term $\mathbb{E}\left\|(H^*)^{-1}v_T\right\|_2^2$, Lemma 9 along with the initial condition yields:

$$\mathbb{E}\left\|(H^*)^{-1}v_T\right\|_2^2\leq\frac{1}{\lambda_{\min}(H^*)^2}\mathbb{E}\left\|v_T\right\|_2^2\leq\frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2\mu\eta T^2}.$$

Collecting above bounds, we conclude that

$$\begin{aligned}\mathbb{E}\left\|(H^*)^{-1}\nabla F(\theta_T)\right\|_2^2\leq\frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right)+\frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu}+\frac{\ell_\Xi}{\mu\sqrt{T}}+\frac{1}{\mu\eta T}\right\}\frac{\sigma_*^2\log T}{T}\\ +\frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2}.\end{aligned}\tag{71}$$

In order to bound the last two terms of the decomposition (69), we recall from Lemma 13 and the initial condition (70) that:

$$\left(\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4\right)^{1/2}\leq\frac{c\widetilde{\sigma_*}^2}{T}.$$

Combining with Eq. (71) and substituting into the decomposition (69), we conclude that:

$$\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2 \le \frac{1}{T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-\top}\right) + \frac{c}{\lambda_{\min}(H^*)^2}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\eta\mu T}\right\}\frac{\sigma_*^2\log T}{T} + \frac{cL_1}{\lambda_{\min}(H^*)^2\mu^2}\cdot\left\{\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2} + \frac{\widetilde{\sigma_*}^3}{T^{3/2}} + \frac{L_1}{\mu^2}\frac{\widetilde{\sigma_*}^4}{T^2}\right\}.$$

Note in the last line, the second $O(T^{-3/2})$ term is always no smaller than the previous first term. Taking $T = n - BT^\flat$ with $n \ge 2BT^\flat$, some algebra then completes the proof of the desired bound.

C.5.2. PROOF OF THE OBJECTIVE GAP BOUND (20b)

Applying second-order Taylor expansion with integral remainder, for any $\theta \in \mathbb{R}^d$, we note the following identity.

$$F(\theta) = F(\theta^*) + \langle\theta - \theta^*,\, \nabla F(\theta^*)\rangle + (\theta-\theta^*)^\top\int_0^1 \nabla^2 F\big(\rho\theta + (1-\rho)\theta^*\big)d\rho\cdot(\theta-\theta^*).$$

Noting that $\nabla F(\theta^*) = 0$ and invoking assumption 4, we have that:

$$\begin{aligned} &F(\theta)\\ &\le F(\theta^*) + \frac{1}{2}(\theta-\theta^*)^\top H^*(\theta-\theta^*) + \left\|\theta-\theta^*\right\|_2\cdot\int_0^1 \left\|\nabla^2 F\big(\rho\theta+(1-\rho)\theta^*\big) - H^*\right\|_{\mathrm{op}}d\rho\cdot\left\|\theta-\theta^*\right\|_2\\ &\le F(\theta^*) + \frac{1}{2}(\theta-\theta^*)^\top H^*(\theta-\theta^*) + L_1\left\|\theta-\theta^*\right\|_2^3. \end{aligned} \tag{72}$$

Analogous to Eq. (68), we have the bound:

$$\begin{aligned} &\left\|(H^*)^{1/2}(\theta_T-\theta^*) - (H^*)^{-1/2}\nabla F(\theta_T)\right\|_2\\ &\le \int_0^1 \left\|(H^*)^{-1/2}\left(\nabla^2 F\big(\rho\theta^* + (1-\rho)\theta_T\big) - H^*\right)(\theta_T-\theta^*)\right\|_2 d\rho\\ &\le \frac{L_1}{\sqrt{\lambda_{\min}(H^*)}}\left\|\theta_T-\theta^*\right\|_2^2 \le \frac{L_1}{\mu^2\sqrt{\lambda_{\min}(H^*)}}\left\|\nabla F(\theta_T)\right\|_2^2. \end{aligned}$$

Denote the residual $q_t := (H^*)^{1/2}(\theta_t-\theta^*) - (H^*)^{-1/2}\nabla F(\theta_t)$. Substituting into the bound (72), we have that:

$$\begin{aligned} &\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right]\\ &\le \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}\nabla F(\theta) + q_T\right\|_2^2 + L_1\mathbb{E}\left\|\theta_T-\theta^*\right\|_2^3\\ &\le \frac{1}{2}\mathbb{E}\left\|(H^*)^{-1/2}\nabla F(\theta_T)\right\|_2^2 + \frac{1}{\sqrt{\lambda_{\min}(H^*)}}\mathbb{E}\Big[\left\|q_t\right\|_2\cdot\left\|\nabla F(\theta_T)\right\|_2\Big] + \frac{1}{2}\mathbb{E}\left\|q_t\right\|_2^2 + \frac{L_1}{\mu^3}\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^3. \end{aligned} \tag{73}$$

For the first term, by applying Lemma 16 and 15 with $G = (H^*)^{-1/2}$, we can obtain the following bound analogous to Eq. (71):

$$\mathbb{E}\left\|(H^*)^{-1/2}\nabla F(\theta_T)\right\|_2^2 \leq \frac{1}{2T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right) + \frac{c}{\lambda_{\min}(H^*)}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\mu\eta T}\right\}\frac{\sigma_*^2\log T}{T} + \frac{cL_1}{\lambda_{\min}(H^*)\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{(\eta\mu)^{1/2}T^2}.$$

For the rest of the terms, we recall that Lemma 13 with the initial condition (70) gives the bound $\left(\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^4\right)^{1/2} \leq \frac{c\widetilde{\sigma_*}^2}{T}$. Substituting into the decomposition (72), we obtain that:

$$\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \leq \frac{1}{2T}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*\right) + \frac{c}{\lambda_{\min}(H^*)}\left\{\frac{\ell_\Xi^2\eta}{\mu} + \frac{\ell_\Xi}{\mu\sqrt{T}} + \frac{1}{\mu\eta T}\right\}\frac{\sigma_*^2\log T}{T} + \frac{cL_1}{\mu^2}\cdot\frac{\widetilde{\sigma_*}^3}{\mu T^{3/2}} + \frac{L_1^2}{\mu^4}\cdot\frac{\widetilde{\sigma_*}^4}{\lambda_{\min}(H^*)T^2}.$$

Noting that $T = n - BT^\flat$ with $n \geq 2BT^\flat$, we completes the proof of the desired bound.

C.6. Proof of Theorem 4

Here we provide a two-step proof of Theorem 4. We continue to adopt the v_t—z_t decomposition as earlier used, and we proceed with the proof in two steps:

Step 1: We first claim the following single-epoch result, Eq. (74), that under the setting of Theorem 4 along with $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$, the single-epoch estimator produced by Algorithm 1 with burn-in time $T_0 = \left\lceil \frac{24}{\eta\mu} \right\rceil$, as $T \to \infty$, $\eta \to 0$ such that $\eta T \to \infty$ satisfies the following convergence in probability:

$$\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*) \xrightarrow{p} 0. \tag{74}$$

Taking this as given, we now combine Eq. (74) with our multi-epoch design Algorithm 2 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 2 will be triggered with length $T = n - T^\flat B$, and hence we apply Eq. (48) so for some $C \leq 56$ we have the initial condition holds: $\mathbb{E}\|\nabla F(\theta_0^{(\eta)})\|_2^2 \leq \frac{C\sigma_*^2}{T^\flat} = O(\eta\mu\sigma_*^2)$, so that as $\eta T \to \infty$,

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

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 $n - T^\flat B$ we can apply Slutsky's rule of weak convergence and obtain the desired weak convergence: as $\eta \to 0$, $n \to \infty$ such that $\eta(n - T^\flat B) \to \infty$

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

Due to our additional scaling condition, we can further replace $T = n - T^\flat B$ by n, concluding Theorem 4.

Step 2: We proceed to prove Eq. (74) with the extra initialization condition $\|\nabla F(\theta_0)\| = O(\sqrt{\eta\mu\sigma_*^2})$. By Eqs. (48) and (49), we have for $T \geq T_0$ there exist constants $a_1, a_2, a_3 > 0$ independent of η, T but depends on the problem parameters $(\mu, L, \ell_\Xi, \sigma_*, \theta_0, \alpha)$, such that

$$\mathbb{E}\left\|z_T\right\|_2^2 \leq \frac{2a_2}{T},$$

and consequently, we have from Eq. (48) that

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

and hence

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

Note from the definition in Eq. (30)

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

By setting $A_t = (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)$, the process $Tz_T - \sum_{s=1}^T \varepsilon_s(\theta^*) = \sum_{s=1}^T A_s$ is a martingale. To conclude the bound (74), we only need to show the following relation as $T \to \infty$ and $\eta \to 0$:

$$\mathbb{E}\left\|\sqrt{T}z_T - \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*)\right\|_2^2 = \frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\|A_s\|^2 \to 0. \tag{75}$$

Since we have

$$\begin{aligned}&\mathbb{E}\left\|\sum_{s=T_0+1}^{T}(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}))\right\|_2^2 = \sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\left\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\right\|_2^2 \\ &\leq \ell_\Xi^2\sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\left\|\theta_{s-1} - \theta_{s-2}\right\|_2^2 = \eta^2\ell_\Xi^2\sum_{s=T_0+1}^{T}(s-1)^2\mathbb{E}\left\|v_{s-1}\right\|_2^2 \\ &\leq \eta^2\ell_\Xi^2\sum_{s=T_0+1}^{T}(s-1)^2\frac{2a_1}{\eta^4(s-1)^4} \leq \frac{2a_1\ell_\Xi^2}{\eta^2 T_0}.\end{aligned}$$

We note that

$$\mathbb{E}\left\|\sum_{s=T_0+1}^{T}(\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta^*))\right\|_2^2=\sum_{s=T_0+1}^{T}\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta^*)\right\|_2^2$$
$$\leq \ell_\Xi^2\sum_{s=T_0+1}^{T}\mathbb{E}\left\|\theta_{s-1}-\theta^*\right\|_2^2\leq\frac{\ell_\Xi^2}{\mu^2}\cdot 4a_3\log\left(\frac{T}{T_0}\right).$$

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

$$\begin{aligned}\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\|A_t\|_2^2&\leq\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\left\|(t-1)(\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta_{t-2}))+\varepsilon_t(\theta_{t-1})-\varepsilon_t(\theta^*)\right\|_2^2\\&\leq 2\ell_\Xi^2\eta^2\cdot\frac{1}{T}\sum_{t=T_0+1}^{T}(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2+\frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T}\sum_{t=T_0+1}^{T}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2\\&\leq 2\ell_\Xi^2\eta^2\cdot\frac{1}{T}\sum_{t=T_0+1}^{T}(t-1)^2\frac{2a_1}{\eta(t-1)^2}+\frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{1}{T}\sum_{t=T_0+1}^{T}\frac{4a_3}{t}\\&=4a_1\ell_\Xi^2\eta+\frac{2\ell_\Xi^2}{\mu^2}\cdot\frac{4a_3\log\left(\frac{T}{T_0}\right)}{T},\end{aligned}$$

i.e. the limit (75) holds, which implies $\sqrt{T}z_T-\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\varepsilon_s(\theta^*)\xrightarrow{p}0$, completing our proof of Eq. (74).

Appendix D. Asymptotic results for single-epoch fixed-step-size ROOT-SGD

In this section, we complement Theorem 4 in §3.3 and establish an additional asymptotic normality result for ROOT-SGD with large step-size. Notably, the covariance of such asymptotic distribution is the sum of the optimal Gaussian limit and a correction term depending on the step-size, which exactly corresponds to existing results on fine-grained CLT for linear stochastic approximation with fixed step-size (Mou et al., 2020).

First, in order to obtain asymptotic results for single-epoch constant-step-size ROOT-SGD, we impose the following slightly stronger assumptions on the smoothness of stochastic gradients and Hessians:

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

$$\sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}\left\|(\nabla^2 f(\theta;\xi)-\nabla^2 f(\theta^*;\xi))v\right\|_2^2\leq\beta^2\left\|\theta-\theta^*\right\|_2^2. \tag{76a}$$

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

$$\mathbb{E}\left\|\nabla f(\theta^*;\xi)\right\|_2^4<\infty,\qquad\text{and}\quad\ell_\Xi'':=\sup_{v\in\mathbb{S}^{d-1}}\left(\mathbb{E}\left\|\nabla^2 f(\theta^*;\xi)v\right\|_2^4\right)^{1/4}<\infty. \tag{76b}$$

Note that both conditions are imposed solely at the optimal point θ^*; we do not impose globally uniform bounds in $\mathbb{R}^d$.

Defining the random matrix $\Xi(\theta) := \nabla^2 f(\theta;\xi) - \nabla^2 F(\theta)$ for any $\theta \in \mathbb{R}^d$, we consider the following matrix equation (a.k.a. *modified Lyapunov equation*):

$$\Lambda H^* + H^*\Lambda - \eta\mathbb{E}\big[\Xi(\theta^*)\Lambda\Xi(\theta^*)\big] - \eta H^*\Lambda H^* = \eta\Sigma^*. \tag{77}$$

in the symmetric matrix Λ. It can be shown that under the given assumptions, this equation has a unique solution—denoted Λ_η—which plays a key role in the following theorem.

Proposition 2 (Asymptotic efficiency, single-epoch ROOT-SGD) *Suppose that Assumptions 1, 2, and 3 are satisfied, as well as* (CLT.A) *and* (CLT.B). *Then there exist constants c_1, c_2, given the step-size $\eta \in \left(0, c_1(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L} \wedge \frac{\mu^{1/3}}{\ell_\Xi'^{4/3}})\right)$, and burn-in time $T_0 = \frac{c_2}{\mu\eta}$, we have*

$$\sqrt{T}(\theta_T - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\left(\Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right)(H^*)^{-1}\right). \tag{78}$$

See §D.1 for the proof of this theorem.

A few remarks are in order. First, we observe that the asymptotic covariance in Eq. (78) is the sum of the matrix $(H^*)^{-1}\Sigma^*(H^*)^{-1}$ and an additional correction term defined in Eq. (77). The asymptotic covariance of $(H^*)^{-1}\Sigma^*(H^*)^{-1}$ matches the standard Cramér-Rao lower bound in the asymptotic statistics literature (van der Vaart and Wellner, 1996; van der Vaart, 2000) and matches the optimal rates achieved in the theory of stochastic approximation (Kushner and Yin, 2003; Polyak and Juditsky, 1992; Ruppert, 1988). The correction term is of the same form as that of the constant-step-size linear stochastic approximation of the Polyak-Ruppert-Juditsky averaging prcedure as derived in (Mou et al., 2020), while our Proposition 2 is applicable to more general nonlinear stochastic problems. For instance in our setting, the correction terms tends to zero as the (constant) step-size decreases to zero, which along with a trace bound leads to the following asymptotics as $T \to \infty$ (see (Mou et al., 2020)):

$$T\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 \sim \mathrm{Tr}\left(\Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right) \le \left(1 + \frac{\ell_\Xi^2\eta}{\mu}\right)\sigma_*^2.$$

The message conveyed by the last display is consistent with the leading two terms in our earlier nonasymptotic bound Eq. (46) in Proposition 1, and thanks to our additional smoothness assumptions (CLT.A) and (CLT.B) we are able to characterize this correction term in a more fine-grained fashion as in the asymptotic covariance of Eq. (78). Second, we note that Proposition 2 has an additional requirement on the step-size, needing it to be upper bounded by $\frac{\mu^{1/3}}{\ell_\Xi'^{4/3}}$. This is a mild requirement on the step-size. In particular, for applications where the noises are light-tailed, ℓ_Ξ' and ℓ_Ξ are of the same order, and the additional requirement $\eta < \frac{c\mu^{1/3}}{\ell_\Xi'^{4/3}}$ is usually weaker than the condition $\eta < \frac{c\mu}{\ell_\Xi^2}$ needed in the previous section.

D.1. Proof of Proposition 2

Denote $H_t(\theta) := \nabla^2 f(\theta;\xi_t)$ and $\Xi_t(\theta) := H_t(\theta) - \nabla^2 F(\theta)$. Intuitively, since the sequence θ_t is converging to θ^* at a $1/\sqrt{t}$ rate, replacing θ_{s-1} with θ^* will only lead to a small change in the sum.

For the martingale Ψ_t, each term can be written as:

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

By Assumption (CLT.A), this quantity should approach $\eta\Xi_t(\theta^*) \cdot (tv_{t-1})$. If we can show the convergence of the sequence $\{tv_t\}_{t\geq T_0}$ to a stationary distribution, then the asymptotic result follows from the Birkhoff ergodic theorem and a martingale CLT. While the process $\{tv_t\}_{t\geq T_0}$ is not Markovian, we show that it can be well-approximated by a time-homogeneous Markov process that we construct in the proof.

In particular, consider the auxiliary process $\{y_t\}_{t\geq T_0}$, initialized as $y_{T_0} = T_0 v_{T_0}$ and updated as

$$y_t = y_{t-1} - \eta H_t(\theta^*)y_{t-1} + \varepsilon_t(\theta^*), \quad \text{for all } t \geq T_0 + 1. \tag{79}$$

Note that $\{y_t\}_{t\geq T_0}$ is a time-homogeneous Markov process that is coupled to $\{(\theta_t, v_t, z_t)\}_{t\geq T_0}$. We have the following coupling estimate:

Lemma 17 *Supposing that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold, then for any iteration $t \geq T_0$ and any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{8\ell_\Xi^2})$, we have*

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

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 §F.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 18 *Suppose that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq T_0$, any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{8\ell_\Xi^2})$ and for any random vector $\widetilde{\theta}_{t-1} \in \mathcal{F}_{t-1}$, we have*

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

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

See §F.2 for the proof of this lemma.

Finally, the following lemma characterizes the behavior of the process $\{y_t\}_{t\geq T_0}$ defined in Eq. (79):

Lemma 19 *Suppose that Assumptions 1, 2 and 3, as well as Conditions (CLT.A) and (CLT.B) hold. Then for any iteration $t \geq T_0$ and any step-size $\eta \in (0, \frac{1}{4L} \wedge \frac{\mu}{16\ell_\Xi^2} \wedge \frac{\mu^{1/3}}{6\ell_\Xi^{4/3}})$, we have*

$$\mathbb{E}(y_t) = 0 \quad \textit{for all } t \geq T_0, \qquad \textit{and} \quad \sup_{t\geq T_0} \mathbb{E}\left\|y_t\right\|_2^4 < a',$$

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

$$H^* Q_\eta + Q_\eta H^* - \eta \left[H^* Q_\eta H^* + \mathbb{E}(\Xi(\theta^*) Q_\eta \Xi(\theta^*))\right] = \frac{1}{\eta}\Sigma^*.$$

See §F.3 for the proof of this lemma.

Taking these three lemmas as given, we now proceed with the proof of Proposition 2. We first define two auxiliary processes:

$$N_T := \sum_{t=T_0+1}^{T} \varepsilon_t(\theta^*), \qquad \Upsilon_T := \eta \sum_{t=T_0+1}^{T} \Xi_t(\theta^*) y_{t-1}.$$

Observe that both N_T and Υ_T are martingales adapted to $(\mathcal{F}_t)_{t\geq T_0}$. In the following, we first bound the differences $\|M_T - N_T\|_2$ and $\|\Psi_T - \Upsilon_T\|_2$, respectively, and then show the limiting distribution results for $N_T + \Upsilon_T$.

By Theorem 1, define $a_0 := \frac{28\sigma_*^2}{\mu^2} + \frac{2700}{\eta^2\mu^4 T_0}\|\nabla F(\theta_0)\|_2^2$, we have

$$\mathbb{E}\|\theta_t - \theta^*\|_2^2 \leq \frac{1}{\mu^2}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \frac{2700\ \|\nabla F(\theta_0)\|_2^2}{\eta^2\mu^4(t+1)^2} + \frac{28\ \sigma_*^2}{\mu^2(t+1)} \leq \frac{a_0}{t+1}, \quad \text{for all } t \geq T_0. \tag{80}$$

Applying the bound (80) with Assumption 3, we have

$$\mathbb{E}\|M_T - N_T\|_2^2 = \sum_{t=T_0+1}^{T} \mathbb{E}\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\|_2^2 \leq \ell_\Xi^2 \sum_{t=T_0+1}^{T} \mathbb{E}\|\theta_{t-1} - \theta^*\|_2^2 \leq a_0\ell_\Xi^2 \log T. \tag{81}$$

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

$$\begin{aligned}
\mathbb{E}\|\Psi_T - \Upsilon_T\|_2^2 &= \sum_{t=T_0+1}^{T} \mathbb{E}\|\eta\Xi_t(\theta^*)y_{t-1} - (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2}))\|_2^2 \\
&\leq \eta^2 \sum_{t=T_0+1}^{T} \mathbb{E}\int_0^1 \|\Xi_t(\theta^*)y_{t-1} - \Xi_t(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}; \xi_t)(t-1)v_{t-1}\|_2^2\, d\rho \leq I_I + I_2,
\end{aligned}$$

where we define

$$I_1 := 2\eta^2 \sum_{t=T_0+1}^{T} \mathbb{E}\int_0^1 \|(\Xi_t(\theta^*) - \Xi_t(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}))\, y_{t-1}\|_2^2\, d\rho, \quad \text{and}$$

$$I_2 := 2\eta^2 \sum_{t=T_0+1}^{T} \mathbb{E}\int_0^1 \|\Xi_t(\rho\theta_{t-1} + (1-\rho)\theta_{t-2})(y_{t-1} - (t-1)v_{t-1})\|_2^2\, d\rho.$$

We bound each of these two terms in succession.

Bound on I_1: In order to bound the term I_1, we apply Lemma 18 with the choice

$$\widetilde{\theta}_{t-1} = \rho\theta_{t-1} + (1-\rho)\theta_{t-2} \in \mathcal{F}_{t-1},$$

so as to obtain

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

Applying the Cauchy-Schwartz inequality yields

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

Putting the two bounds together, we obtain:

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

Thus, we find that

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

where in the last step we used the inequality (80). Summing over t from T_0+1 to T yields the bound

$$I_1 \leq 2\eta^2 \sum_{t=T_0+1}^{T} \frac{16c_1\sqrt{a_0}}{\sqrt{t}} \leq 64\eta^2 c_1\sqrt{a_0 T}.$$

Bound on I_2: Turning to the term I_2, by Assumption 3 and Lemma 17, we note that:

$$I_2 \leq 2\eta^2 \sum_{t=T_0+1}^{T} \ell_\Xi^2 \mathbb{E}\left\|y_{t-1} - (t-1)v_{t-1}\right\|_2^2 \leq 2\eta^2\ell_\Xi^2 \sum_{t=T_0+1}^{T} \frac{c_0}{\sqrt{t}} \leq 4\eta^2\ell_\Xi^2 c_0\sqrt{T}.$$

Putting these inequalities together, we conclude that:

$$\mathbb{E}\left\|\Psi_T - \Upsilon_T\right\|_2^2 \leq (64\eta^2 c_1\sqrt{a_0} + 4\eta^2\ell_\Xi^2 c_0)\sqrt{T}. \tag{82}$$

Now we have the estimates for the quantities $\|\Psi_T - \Upsilon_T\|_2$ and $\|M_T - N_T\|_2$. In the following, we first prove the CLT for $N_T + \Upsilon_T$, and then use the error bounds to establish CLT for $M_T + \Psi_T$, which ultimately implies the desired limiting result for $\sqrt{T}(\theta_T - \theta^*)$

Define $\nu_t := \varepsilon_t(\theta^*) + \eta \Xi_t(\theta^*) y_{t-1}$. By definition, $N_T + \Upsilon_T = \sum_{t=T_0}^T \nu_t$, and we have

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

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 19, we note that the time-homogeneous Markov process $\{y_t\}_{t \geq T_0}$ converges asymptotically to a stationary distribution with covariance Q_η. Invoking the Birkhoff ergodic theorem, we have

$$\frac{1}{T}\sum_{t=T_0+1}^T \mathbb{E}\left(\Xi_t(\theta^*) y_{t-1} y_{t-1}^\top \Xi_t(\theta^*)^\top \mid \mathcal{F}_{t-1}\right) = \mathbb{E}(\Xi(\theta^*) \otimes \Xi(\theta^*)) \left[\frac{1}{T}\sum_{t=T_0+1}^T y_{t-1} y_{t-1}^\top\right] \xrightarrow{p} \mathbb{E}\left(\Xi(\theta^*) Q_\eta \Xi(\theta^*)^\top\right).$$

For the cross terms, we note that:

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

Note that by Lemma 19, we have $\mathbb{E}(y_t) = 0$ for any $t \geq T_0$. By the weak law of large numbers, we have $\frac{1}{T}\sum_{t=T_0+1}^T y_t \xrightarrow{p} 0$. Putting together these inequalities, we find that

$$\frac{1}{T}\sum_{t=T_0+1}^T \mathbb{E}\left(\nu_t \nu_t^\top \mid \mathcal{F}_{t-1}\right) = \frac{1}{T}\sum_{t=T_0+1}^T \Big(\Sigma^* + \eta^2 \mathbb{E}(\Xi(\theta^*) \otimes \Xi(\theta^*))[y_t y_t^\top] + \eta \mathbb{E}(\varepsilon(\theta^*) \otimes \Xi(\theta^*))[y_{t-1}] + \eta \mathbb{E}(\Xi(\theta^*) \otimes \varepsilon(\theta^*))[y_{t-1}]\Big),$$

and hence the random matrix $\frac{1}{T}\sum_{t=T_0+1}^T \mathbb{E}\left(\nu_t \nu_t^\top \mid \mathcal{F}_{t-1}\right)$ 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 and Heyde (1980)). It remains to verify the conditional Lindeberg condition. Indeed, for any $\varepsilon > 0$, a straightforward calculation yields:

$$R_T(\varepsilon) := \sum_{t=T_0+1}^T \mathbb{E}\left(\left\|\frac{\nu_t}{\sqrt{T}}\right\|_2^2 \mathbf{1}_{\left\|\frac{\nu_t}{\sqrt{T}}\right\|_2 > \varepsilon} \,\middle|\, \mathcal{F}_{t-1}\right)$$
$$\overset{(i)}{\leq} \frac{1}{T}\sum_{t=T_0+1}^T \sqrt{\mathbb{E}\left(\|\nu_t\|_2^4 \mid \mathcal{F}_{t-1}\right)} \cdot \sqrt{\mathbb{P}\left(\|\nu_t\|_2 > \varepsilon\sqrt{T} \mid \mathcal{F}_{t-1}\right)} \overset{(ii)}{\leq} \frac{1}{T}\sum_{t=T_0+1}^T \frac{1}{(\varepsilon\sqrt{T})^2} \mathbb{E}\left(\|\nu_t\|_2^4 \mid \mathcal{F}_{t-1}\right).$$

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:

$$\mathbb{E}\left(\|\nu_t\|_2^4 \mid \mathcal{F}_{t-1}\right) \leq 8\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + 8\ell_\Xi^4 \|y_{t-1}\|_2^4.$$

Plugging back to the upper bound for $R_T(\varepsilon)$, and applying Lemma 19, as $T \to \infty$, we have

$$\mathbb{E}[R_T(\varepsilon)] \leq \frac{8}{T\varepsilon^2}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + \frac{8\ell_\Xi^4}{T^2\varepsilon^2}\sum_{t=T_0+1}^{T}\mathbb{E}\|y_{t-1}\|_2^4 \leq \frac{8}{T\varepsilon^2}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4 + \frac{8\ell_\Xi^4}{T\varepsilon^2}a' \to 0.$$

Note that $R_T(\varepsilon) \geq 0$ by definition. The limit statement implies that $R_T(\varepsilon) \xrightarrow{p} 0$, for any $\varepsilon > 0$. Therefore, the conditional Lindeberg condition holds true, and we have the CLT:

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

By the second-moment estimates (81) and (82), we have

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

With the burn-in time T_0 fixed, we also have $\frac{T_0}{T}z_{T_0} \xrightarrow{p} 0$. By Slutsky's theorem, we have

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

Note that $\nabla F(\theta_{t-1}) = v_t - z_t$. By Lemma 17 and Lemma 19, we have

$$\mathbb{E}\|v_t\|_2^2 \leq \frac{2}{t^2}\mathbb{E}\|tv_t - y_t\|_2^2 + \frac{2}{t^2}\mathbb{E}\|y_t\|_2^2 \leq \frac{2}{t^2}\left(\sqrt{a'} + \frac{c_0}{\sqrt{t}}\right),$$

which implies that $\sqrt{t}v_t \xrightarrow{p} 0$. Recall that $z_t = v_t - \nabla F(\theta_{t-1})$. By Slutsky's theorem, we obtain:

$$\sqrt{T}\cdot\nabla F(\theta_T) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^* + \mathbb{E}\left[\Xi(\theta^*)\Lambda_\eta\Xi(\theta^*)\right]\right).$$

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

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

By Assumption (CLT.A), we have

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

Consequently, we have the bound:

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

By Eq. (80), 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,

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

which finishes the proof.

Appendix E. Proofs of auxiliary lemmas in §C.3, §C.4 and §C.5

E.1. Proof of Lemma 13

Recall that we have the recursive update rule of z_t as

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

Taking fourth moments on both sides, we have

$$\begin{aligned}
\mathbb{E}\|tz_t\|_2^4 &= \mathbb{E}\|(t-1)z_{t-1} + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&= \mathbb{E}\|(t-1)z_{t-1}\|_2^4 + \mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^4 \\
&\quad + 4\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^3 \|(t-1)z_{t-1}\|_2 \\
&\quad + 6\mathbb{E}\|(t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1})\|_2^2 \|(t-1)z_{t-1}\|_2^2 , \qquad (83)
\end{aligned}$$

where one of the terms is zeroed out. By Hölder's inequality and Young's inequality, we bound the third term and the fourth term of the RHS as

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

and

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

Thus Eq. (83) continues 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

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

Then

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

Combining this with Eq. (90) in Lemma 20 that

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

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

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

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

$$\begin{aligned}&T^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4}\le 2T^2\sqrt{\mathbb{E}\left\|z_T\right\|_2^4}+2T^2\sqrt{\mathbb{E}\left\|v_T\right\|_2^4} \\ &\le 2\sqrt{\mathbb{E}\left\|T_0z_{T_0}\right\|_2^4}+2\sqrt{\mathbb{E}\left\|T_0v_{T_0}\right\|_2^4}+\frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}+70(T-T_0)\widetilde{\sigma_*}^2. \qquad (84)\end{aligned}$$

Further for $T_0z_{T_0}$ and $T_0v_{T_0}$ we note that by applying Khintchine's inequality as well as Young's inequality we have

$$\mathbb{E}\left\|T_0z_{T_0}\right\|_2^4=\mathbb{E}\left\|\sum_{t=1}^{T_0}\varepsilon_t(\theta_0)\right\|_2^4\le\mathbb{E}\left(\sum_{t=1}^{T_0}\left\|\varepsilon_t(\theta_0)\right\|_2^2\right)^2\le T_0\mathbb{E}\sum_{t=1}^{T_0}\left\|\varepsilon_t(\theta_0)\right\|_2^4\le 8T_0^2\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4+\widetilde{\sigma_*}^4\right), \qquad (85)$$

and

$$\begin{aligned}\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4 &= \mathbb{E}\left\|T_0 z_{T_0}\right\|_2^4 + \mathbb{E}\left\|T_0\nabla F(\theta_0)\right\|_2^4 + 4\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^3\left\|T_0\nabla F(\theta_0)\right\|_2 + 6\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^2\left\|T_0\nabla F(\theta_0)\right\|_2^2 \\ &\le 7\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4 + 5\mathbb{E}\left\|T_0\nabla F(\theta_0)\right\|_2^4 \le 56T_0^2\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4 + \widetilde{\sigma_*}^4\right) + 5T_0^4\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4.\end{aligned} \tag{86}$$

Taking squared root on Eq. (85) and (86) and recalling that $\eta \le \frac{\mu}{64\ell_\Xi^2}$, we have

$$\sqrt{\mathbb{E}\left\|T_0 z_{T_0}\right\|_2^4} \le 2\sqrt{2}T_0\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \widetilde{\sigma_*}^2\right), \tag{87}$$

and

$$\sqrt{\mathbb{E}\left\|T_0 v_{T_0}\right\|_2^4} \le (\sqrt{5}+1/8)T_0^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + 8T_0\widetilde{\sigma_*}^2. \tag{88}$$

Bringing Eq. (87) and (88) into Eq. (84), we arrive at the following:

$$\begin{aligned}T^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4} &\le 4\sqrt{2}T_0\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \widetilde{\sigma_*}^2\right) + (2\sqrt{5}+1/4)T_0^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} \\ &\quad + 16T_0\widetilde{\sigma_*}^2 + \frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + 70(T-T_0)\widetilde{\sigma_*}^2 \\ &\le 5T_0^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{12}{\eta\mu}\sum_{t=T_0+1}^{T}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + 70T\widetilde{\sigma_*}^2.\end{aligned} \tag{89}$$

Dividing both sides by T^2, summing up Eq. (89) from $T = T_0+1$ to $T^* \ge T_0+1$ and using the fact that $\eta \le \frac{\mu}{64\ell_\Xi^2}$, $T_0 \ge 2$, we have

$$\sum_{T=T_0+1}^{T^*}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4} \le 5T_0\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{12}{\eta\mu T_0}\sum_{t=T_0+1}^{T^*}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + 70\widetilde{\sigma_*}^2\log\left(\frac{T^*}{T_0}\right).$$

Taking $T_0 = \left\lceil\frac{24}{\eta\mu}\right\rceil$, we have

$$\sum_{T=T_0+1}^{T^*}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4} \le 10T_0\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + 140\widetilde{\sigma_*}^2\log\left(\frac{T^*}{T_0}\right).$$

Again by Eq. (89), we have

$$\begin{aligned}&T^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4} \\ &\le 5T_0^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{12}{\eta\mu}\left(10T_0\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + 140\widetilde{\sigma_*}^2\log\left(\frac{T}{T_0}\right)\right) + 70T\widetilde{\sigma_*}^2 \\ &\le 10T_0^2\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + 70T_0\widetilde{\sigma_*}^2\log\left(\frac{T}{T_0}\right) + 70T\widetilde{\sigma_*}^2.\end{aligned}$$

Dividing both sides by T^2 we conclude that

$$\sqrt{\mathbb{E}\left\|\nabla F(\theta_{T-1})\right\|_2^4} \leq \frac{10T_0^2}{T^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + 70\left(1+\frac{T_0}{T}\log\left(\frac{T}{T_0}\right)\right)\frac{\widetilde{\sigma_*}^2}{T}$$
$$\leq \frac{10T_0^2}{T^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{140\widetilde{\sigma_*}^2}{T}.$$

which finishes our proof of Lemma 13.

E.2. Proof of Lemma 14

Our main technical tools is the following lemma, which bound the fourth moment of the v_t recursion.

Lemma 20 *Under the setting of Proposition 1, when $\eta \leq \frac{1}{56L} \wedge \frac{\mu}{64\ell_\Xi^2}$, we have the following bound for $t \geq T_0 + 1$*

$$\sqrt{\mathbb{E}\left\|tv_t\right\|_2^4} \leq \left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4} + \frac{5}{\eta\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + 14\widetilde{\sigma_*}^2. \tag{90}$$

The detailed proof is relegated to §E.3.1. We are ready for the proof of Lemma 14. Indeed, from (56) and (90)

$$t^2\sqrt{\mathbb{E}\left\|v_t\right\|_2^4} \leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{5}{\eta\mu}\left[\frac{60\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^2\mu^2t^2} + \frac{140\ \widetilde{\sigma_*}}{t}\right] + 14\widetilde{\sigma_*}^2$$
$$\leq \left(1-\frac{\eta\mu}{2}\right)(t-1)^2\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3t^2} + 714\widetilde{\sigma_*}^2. \tag{91}$$

We have from (91)

$$t^4\sqrt{\mathbb{E}\left\|v_t\right\|_2^4} \leq \left(1-\frac{\eta\mu}{2}\right)t^2(t-1)^2\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3} + 714\widetilde{\sigma_*}^2t^2$$
$$\leq \left(1-\frac{\eta\mu}{6}\right)(t-1)^4\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3} + 714\widetilde{\sigma_*}^2t^2,$$

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

$$T^4\sqrt{\mathbb{E}\left\|v_T\right\|_2^4} \leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}T_0^4\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4} + \sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\left[\frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3} + 714\widetilde{\sigma_*}^2t^2\right]$$
$$\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}T_0^4\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4} + \sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}\frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3} + \sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}714\widetilde{\sigma_*}^2t^2$$
$$\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}T_0^4\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4} + \frac{6}{\eta\mu}\cdot\frac{310\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^3\mu^3} + \frac{6}{\eta\mu}T^2\cdot 714\widetilde{\sigma_*}^2$$
$$\leq \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}T_0^4\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4} + \frac{187500\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4} + \frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu}T^2. \tag{92}$$

where the summand is increasing so

$$\sum_{t=T_0+1}^{T}\left(1-\frac{\eta\mu}{6}\right)^{T-t}t^2 \le \frac{6}{\eta\mu}T^2.$$

All in all, this concludes

$$\begin{aligned}\sqrt{\mathbb{E}\left\|v_T\right\|_2^4} &\le \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}\frac{T_0^4}{T^4}\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4}+\frac{187500\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}\\ &\le \left(1-\frac{\eta\mu}{6}\right)^{T-T_0}\frac{T_0^4}{T^4}\sqrt{\mathbb{E}\left\|v_{T_0}\right\|_2^4}+\frac{187500\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}.\end{aligned}$$

Bringing the burn-in upper bounds (88), we arrive at our final result for bounding $\sqrt{\mathbb{E}\left\|v_T\right\|_2^4}$:

$$\begin{aligned}\sqrt{\mathbb{E}\left\|v_T\right\|_2^4} &\le \left(\frac{3T_0^4}{T^4}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4}+\frac{8T_0^3}{T^4}\widetilde{\sigma_*}^2\right)+\frac{187500\ \left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}+\frac{(714)(6)\widetilde{\sigma_*}^2}{\eta\mu T^2}\\ &\le \frac{1359375\left\|\nabla F(\theta_0)\right\|_2^2}{\eta^4\mu^4T^4}+\frac{4484\widetilde{\sigma_*}^2}{\eta\mu T^2}.\end{aligned}$$

E.3. Proofs of recursive bounds on v_t

In this section, we prove Lemmas 20 and 12, the two recursive bounds for $\{v_t\}_{t\ge T_0}$ used in the proof of main theorems.

E.3.1. PROOF OF LEMMA 20

By definition, we note that:

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

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

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

Taking the fourth moments on both sides, we have

$$
\begin{aligned}
&\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4 \\
&= \mathbb{E}\left\|(t-1)v_{t-1} + (t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4 \\
&= (t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + \underbrace{4\mathbb{E}\left[\left\|(t-1)v_{t-1}\right\|_2^2 \left\langle (t-1)v_{t-1},\, (t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\rangle\right]}_{=:T_2} \\
&\quad + \underbrace{6\mathbb{E}\left[\left\|(t-1)v_{t-1}\right\|_2^2 \left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^2\right]}_{=:T_1} \\
&\quad + \underbrace{4\mathbb{E}\left[\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^3 \left\|(t-1)v_{t-1}\right\|_2\right]}_{=:T_3} \\
&\quad + \mathbb{E}\left[\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4\right]. \qquad (93)
\end{aligned}
$$

To bound term T_1, we apply the Hölder's inequality and have

$$
T_1 \leq 6\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}. \qquad (94)
$$

To bound term T_3, we again apply the Hölder's inequality:

$$
\begin{aligned}
T_3 &\leq 4\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{3/4}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/4} \\
&\leq 2\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2} \\
&\quad + 2\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4. \qquad (95)
\end{aligned}
$$

To bound term T_2, we first take expectation with respect to ξ_t and have

$$
T_2 = 4\mathbb{E}\left[\left\|(t-1)v_{t-1}\right\|_2^2 \left\langle (t-1)v_{t-1},\, (t-1)\left(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right)\right\rangle\right],
$$

where

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

and

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

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

$$
\begin{aligned}
T_2 &\leq -(t-1)^4\mathbb{E}\left[\left\|v_{t-1}\right\|_2^2\left(\frac{1}{\eta L}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 + 3\eta\mu\left\|v_{t-1}\right\|_2^2\right)\right] \\
&= -3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 - \frac{(t-1)^4}{\eta L}\mathbb{E}\left\|v_{t-1}\right\|_2^2\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 \\
&\leq -3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 - \frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}. \qquad (96)
\end{aligned}
$$

Combining Eqs. (94), (95) and (96) into Eq. (93) we have

$$\begin{aligned}
&\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4 \\
&\le (t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + 3\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4 \\
&\quad + 8\left(\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4\right)^{1/2} \\
&\quad - 3\eta\mu(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 - \frac{(t-1)^4}{\eta L}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4\right)^{1/2}. \quad (97)
\end{aligned}$$

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:

$$\begin{aligned}
&\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4 \\
&\le \mathbb{E}\left\|(t-1)(\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})) + (t-1)(\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*) + \varepsilon_t(\theta^*)\right\|_2^4 \\
&\le 8(t-1)^4\underbrace{\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) + \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right\|_2^4}_{=:I_1} + 8\underbrace{\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*) + \varepsilon_t(\theta^*)\right\|_2^4}_{=:I_2}.
\end{aligned} \quad (98)$$

We claim that

$$I_1 \le 5\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 7\ell_\Xi^4\eta^4\mathbb{E}\left\|v_{t-1}\right\|_2^4, \quad (99)$$

and

$$I_2 \le 8\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\right\|_2^4 + 8\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^4 \le \frac{8\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 8\widetilde{\sigma_*}^4. \quad (100)$$

Combining Eqs. (98), (99) and (100) we have the bound

$$\begin{aligned}
&\mathbb{E}\left\|(t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \varepsilon_t(\theta_{t-1})\right\|_2^4 \\
&\le 40(t-1)^4\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 56\ell_\Xi^4\eta^4(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + \frac{64\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 64\widetilde{\sigma_*}^4.
\end{aligned} \quad (101)$$

Then, we bring Eq. (101) into Eq. (97) and 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 + 120(t-1)^4\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 168\ell_\Xi^4\eta^4(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 \\
&\quad + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\widetilde{\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} \\
&\quad + 64\ell_\Xi^2\eta^2(t-1)^4\mathbb{E}\left\|v_{t-1}\right\|_2^4 + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\left\|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} \\
&\le \left(1 - 3\eta\mu + 64\ell_\Xi^2\eta^2 + 168\ell_\Xi^4\eta^4\right)\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4 \\
&\quad + \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} \\
&\quad + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2} + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\widetilde{\sigma_*}^4 \\
&\le (1-\eta\mu)^2\,\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4 + 64\left(\frac{\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + \widetilde{\sigma_*}^4\right)^{1/2}\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2} \\
&\quad + \frac{192\ell_\Xi^4}{\mu^4}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4 + 192\widetilde{\sigma_*}^4,
\end{aligned}
$$

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

$$
168\ell_\Xi^4\eta^4 \le \eta^2\mu^2, \ell_\Xi^2\eta^2 \le \eta\mu \qquad \text{and} \quad 8\sqrt{40} - \frac{1}{\eta L} + 120L^2\eta^2 \le 0.
$$

Taking squared root on both sides, we have

$$
\sqrt{\mathbb{E}\left\|tv_t - \nabla F(\theta_{t-1})\right\|_2^4} \le (1-\eta\mu)\sqrt{\mathbb{E}\left\|(t-1)v_{t-1}\right\|_2^4} + 32\left(\frac{\ell_\Xi^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + \widetilde{\sigma_*}^2\right). \tag{102}
$$

Furthermore, Young's inequality gives[13]

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

Combining this we have

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

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

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

Rearranging and taking squared root on both sides we conclude that

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

13. Here, a different coefficient from the analysis as in the proof of Theorem 1 is adopted.

Further rearranging, we have

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

which concludes our proof.

Proof of Eq. (99): We use similar decomposition as in the decomposition in Eq. (93) and have

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

where we note that we used the fact that one of the cross terms in the fourth moment decomposition $\mathbb{E}\left[\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2 \left\langle \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}),\, \varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\right\rangle\right] = 0$. Further utilizing the Hölder's inequality, we have

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

This completes the proof of Eq. (99).

E.3.2. PROOF OF LEMMA 12

By definition, we note that:

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

Taking the second moments for both sides, we have:

$$\begin{aligned}
\mathbb{E}\left\|v_t\right\|_2^2 = &\left(1-\frac{1}{t}\right)^2 \underbrace{\mathbb{E}\left\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\right\|_2^2}_{I_1} + \frac{1}{t^2}\underbrace{\mathbb{E}\left\|\nabla f(\theta_{t-1};\xi_t)\right\|_2^2}_{I_2} \\
&+ 2\frac{t-1}{t^2}\underbrace{\mathbb{E}\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t),\, \nabla f(\theta_{t-1};\xi_t)\rangle}_{I_3}.
\end{aligned}$$

For the first term, using the fact that $\theta_{t-1} - \theta_{t-2} = -\eta v_{t-1}$, we start with the following decomposition:

$$\begin{aligned}
&\mathbb{E}\left(\|v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathcal{F}_{t-1}\right) \\
&= \|v_{t-1}\|_2^2 + 2\mathbb{E}\left(\langle v_{t-1},\, \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\rangle \mid \mathcal{F}_{t-1}\right) + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathcal{F}_{t-1}\right) \\
&= \|v_{t-1}\|_2^2 - \frac{2}{\eta}\langle \theta_{t-1} - \theta_{t-2},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle + \mathbb{E}\left(\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)\|_2^2 \mid \mathcal{F}_{t-1}\right).
\end{aligned}$$

Since F is μ-strongly convex and L-smooth, we have the following standard inequality:

$$\langle \theta_{t-1} - \theta_{t-2},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\rangle \geq \frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2 \,\mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}.$$

Hence, when the step-size satisfies the bound $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{2\ell_\Xi^2}$, there is the bound:

$$\begin{aligned}
I_1 &\leq \mathbb{E}\,\|v_{t-1}\|_2^2 - \frac{2}{\eta}\mathbb{E}\left(\frac{\|\theta_{t-1} - \theta_{t-2}\|_2^2\,\mu L}{\mu + L} + \frac{\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2}{\mu + L}\right) + 2\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\
&\quad + 2\mathbb{E}\left(\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta_{t-2};\xi_t)\|_2^2\right) \\
&\leq (1 - \eta\mu + 2\eta^2\ell_\Xi^2)\mathbb{E}\,\|v_{t-1}\|_2^2 + 2\left(1 - \frac{1}{\eta(\mu + L)}\right)\mathbb{E}\,\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2^2 \\
&\leq (1 - 3\eta\mu/4)\mathbb{E}\,\|v_{t-1}\|_2^2.
\end{aligned}$$

Now we study the second term I_2, note that

$$\begin{aligned}
\mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t)\|_2^2 &\leq \mathbb{E}\,\|\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)\|_2^2 + 4\mathbb{E}\,\|\nabla f(\theta^*;\xi_t)\|_2^2 \\
&\leq 2\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 2\mathbb{E}\,\|\varepsilon(\theta_{t-1};\xi_t) - \varepsilon(\theta^*;\xi_t)\|_2^2 + 4\mathbb{E}\,\|\nabla f(\theta^*;\xi_t)\|_2^2 \\
&\leq 2\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 2\ell_\Xi^2\mathbb{E}\,\|\theta_{t-1} - \theta^*\|_2^2 + 4\sigma_*^2 \\
&\leq 2\left(1 + \frac{\ell_\Xi^2}{\mu^2}\right)\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^2 + 4\sigma_*^2.
\end{aligned}$$

For the cross term I_3, we note that:

$$\begin{aligned}
&\mathbb{E}\left(\langle v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t),\, \nabla f(\theta_{t-1};\xi_t)\rangle \mid \mathcal{F}_{t-1}\right) \\
&= \mathbb{E}\left(\langle v_{t-1},\, \nabla f(\theta_{t-1};\xi_t)\rangle \mid \mathcal{F}_{t-1}\right) + \mathbb{E}\left(\langle \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t),\, \nabla F(\theta_{t-1})\rangle \mid \mathcal{F}_{t-1}\right) \\
&\quad + \mathbb{E}\left(\langle \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t),\, \varepsilon_t(\theta_{t-1})\rangle \mid \mathcal{F}_{t-1}\right) \\
&= \underbrace{\langle v_{t-1},\, \nabla F(\theta_{t-1})\rangle + \langle \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}),\, \nabla F(\theta_{t-1})\rangle}_{:=T_1} \\
&\quad + \underbrace{\mathbb{E}\left(\langle \varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2};\xi_t),\, \varepsilon(\theta_{t-1},\xi_t)\rangle \mid \mathcal{F}_{t-1}\right)}_{:=T_2}.
\end{aligned}$$

For the term T_1, we note that:

$$T_1 \leq \|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2 + \|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2 \leq (1 + \eta L)\,\|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2.$$

For the term T_2, we have:

$$\begin{aligned}
T_2 &\leq \mathbb{E}\left(\|\varepsilon(\theta_{t-1};\xi_t)-\varepsilon(\theta_{t-2};\xi_t)\|_2 \cdot \|\varepsilon(\theta_{t-1};\xi_t)\|_2 \mid \mathcal{F}_{t-1}\right) \\
&\leq \sqrt{\mathbb{E}(\|\varepsilon(\theta_{t-1};\xi_t)-\varepsilon(\theta_{t-2};\xi_t)\|_2^2 \mid \mathcal{F}_{t-1}) \cdot \mathbb{E}(\|\varepsilon(\theta_{t-1},\xi_t)\|_2^2 \mid \mathcal{F}_{t-1})} \\
&\leq \ell_\Xi^2 \eta \|v_{t-1}\|_2 \cdot \|\theta_{t-1}-\theta^*\|_2 \\
&\leq \frac{\ell_\Xi^2}{\mu}\eta \|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2 .
\end{aligned}$$

So we have:

$$\begin{aligned}
I_3 &\leq \frac{3}{2}\mathbb{E}\left(\|v_{t-1}\|_2 \cdot \|\nabla F(\theta_{t-1})\|_2\right) \leq \frac{3}{2}\sqrt{\mathbb{E}\|v_{t-1}\|_2^2 \cdot \mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2} \\
&\leq \frac{t\eta\mu}{8}\mathbb{E}\|v_{t-1}\|_2^2 + \frac{9}{2t\mu\eta}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 .
\end{aligned}$$

Putting above estimates together, we obtain:

$$\begin{aligned}
\mathbb{E}\|v_t\|_2^2 &\leq \left(1-\frac{1}{t}\right)^2 (1-3\eta\mu/4)\mathbb{E}\|v_{t-1}\|_2^2 + \frac{1}{t^2}\left(4\sigma_*^2 + 2\big(1+\frac{\ell_\Xi^2}{\mu^2}\big)\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2\right) \\
&\quad + \frac{(t-1)\eta\mu}{4t}\mathbb{E}\|v_{t-1}\|_2^2 + \frac{9}{t^2\mu\eta}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 \\
&\leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta\mu}{2}\right)\mathbb{E}\|v_{t-1}\|_2^2 + \frac{10}{t^2\mu\eta}\mathbb{E}\|\nabla F(\theta_{t-1})\|_2^2 + \frac{4\sigma_*^2}{t^2},
\end{aligned}$$

which completes the proof of this lemma.

E.4. Proof of Lemmas 15 and 16

In this section, we present the proofs of lemma 15 and 16, the two technical lemmas involving a test matrix $G \in \mathbb{R}^{d\times d}$.

E.4.1. PROOF OF LEMMA 15

The proof is similar to that of Lemma 10, and we follow the notation in such lemma throughout. Indeed, we note the following telescope result:

$$T^2\mathbb{E}\|Gz_T\|_2^2 - T_0^2\mathbb{E}\|Gz_{T_0}\|_2^2 = \sum_{t=T_0+1}^{T}\mathbb{E}\|G\varepsilon_t(\theta^*)\|_2^2 + \sum_{t=T_0+1}^{T}\mathbb{E}\|G\zeta_t\|_2^2 + 2\sum_{t=T_0+1}^{T}\mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\zeta_t\rangle.$$

Clearly, for each t, we have the following identity:

$$\mathbb{E}\|G\varepsilon_t(\theta^*)\|_2^2 = \mathrm{Tr}\big(G\Sigma^*G^\top\big).$$

For the additional terms, we note that $\mathbb{E}\left\|G\zeta_t\right\|_2^2 \le \|\!|G|\!\|_{\rm op}^2 \mathbb{E}\left\|\zeta_t\right\|_2^2$, and following the derivation in the proof of Lemma 10, we have the following identity:

$$\begin{aligned}&\sum_{t=T_0+1}^{T} \mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\zeta_t\rangle \\ &= T\cdot \mathbb{E}\langle G\varepsilon_T(\theta^*),\, G\varepsilon_T(\theta_{T-1}) - G\varepsilon_T(\theta^*)\rangle - T_0\cdot \mathbb{E}\langle G\varepsilon_{T_0}(\theta^*),\, G\varepsilon_{T_0}(\theta_{T_0-1}) - G\varepsilon_{T_0}(\theta^*)\rangle.\end{aligned}$$

Applying the Cauchy-Schwartz inequality, we obtain bounds similar to Eq. (54), for $t \in \{T_0, T\}$:

$$\begin{aligned}\left|t\cdot \mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta^*)\rangle\right| &\le t\|\!|G|\!\|_{\rm op}^2 \cdot \sqrt{\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^2}\cdot \sqrt{\mathbb{E}\left\|\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta^*)\right\|_2^2} \\ &\le \|\!|G|\!\|_{\rm op}^2 \frac{t\sigma_*\ell_\Xi}{\mu}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}.\end{aligned}$$

For the burn-in period, we have that:

$$T_0^2\mathbb{E}\left\|Gz_{T_0}\right\|_2^2 \le 2T_0\mathbb{E}\left\|G\big(\varepsilon_1(\theta_0) - \varepsilon_1(\theta^*)\big)\right\|_2^2 + 2T_0\mathbb{E}\left\|G\varepsilon_1(\theta^*)\right\|_2^2 \le \frac{2T_0\ell_\Xi^2}{\mu^2}\left\|\nabla F(\theta_0)\right\|_2^2 + 2T_0\mathrm{Tr}\big(G\Sigma^*G\big).$$

Putting them together, and following the derivation in Lemma 10, we obtain the conclusion of this lemma.

E.4.2. PROOF OF LEMMA 16

The proof is similar to that of Lemma 11. Following the notation in Lemma 11, we have the decomposition:

$$\left|\mathbb{E}\langle tGz_t,\, Gv_t\rangle\right| \le (t-\widetilde{T}^*)\left|\mathbb{E}\langle Gz_{t-\widetilde{T}^*},\, Gv_t\rangle\right| + \left|\mathbb{E}\langle G(tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}),\, Gv_t\rangle\right|.$$

Noting that

$$\left|\mathbb{E}\langle G(tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}),\, Gv_t\rangle\right| \le \|\!|G|\!\|_{\rm op}^2\sqrt{\mathbb{E}\left\|tz_t - (t-\widetilde{T}^*)z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|v_t\right\|_2^2},$$

and that

$$\left|\mathbb{E}\langle Gz_{t-\widetilde{T}^*},\, Gv_t\rangle\right| \le \|\!|G|\!\|_{\rm op}^2\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}^*}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}[v_t \mid \mathcal{F}_{t-\widetilde{T}^*}]\right\|_2^2}.$$

The rest of the proof simply follows that of Lemma 11, with an additional factor of $\|\!|G|\!\|_{\rm op}^2$ in each term.

Appendix F. Proofs of auxiliary lemmas in §D.1

In this section, we prove the three auxiliary lemmas used in the proof of Proposition 2. Note that the proofs of the lemmas have inter-dependencies. In the following, we first prove Lemma 17 assuming Lemma 18, and then prove Lemma 18 assuming Lemma 19. Finally, we give a self-contained proof for Lemma 19.

F.1. Proof of Lemma 17

We begin by making note of the identities

$$tv_t = (t-1)(v_{t-1} + \nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) + \nabla f(\theta_{t-1};\xi_t), \quad \text{and}$$
$$y_t = y_{t-1} - \eta\nabla^2 f(\theta^*;\xi_t)y_{t-1} + \nabla f(\theta^*;\xi_t).$$

Defining the quantity $e_t := tv_t - y_t$, we see that the two identities above imply that

$$\begin{aligned} e_t &= e_{t-1} + \big((t-1)(\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta_{t-2};\xi_t)) - \eta\nabla^2 f(\theta^*;\xi_t)y_{t-1}\big) + (\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)) \\ &= Q_1(t) + Q_2(t) + Q_3(t), \end{aligned}$$

where we define

$$Q_1(t) := e_{t-1} - \eta\int_0^1 \nabla^2 f(\rho\theta_{t-1} + (1-\rho)\theta_{t-2};\xi_t)e_{t-1}d\rho, \qquad Q_2(t) := (\nabla f(\theta_{t-1};\xi_t) - \nabla f(\theta^*;\xi_t)),$$
$$Q_3(t) := \eta\int_0^1 \left(\nabla^2 f(\rho\theta_{t-1} + (1-\rho)\theta_{t-2};\xi_t) - \nabla^2 f(\theta^*;\xi_t)\right) y_{t-1}d\rho.$$

By the triangle inequality, we have

$$\mathbb{E}\left\|e_t\right\|_2^2 \le \left(\sqrt{\mathbb{E}\left\|Q_1(t)\right\|_2^2} + \sqrt{\mathbb{E}\left\|Q_2(t)\right\|_2^2} + \sqrt{\mathbb{E}\left\|Q_3(t)\right\|_2^2}\right)^2.$$

In the following, we bound each term $\mathbb{E}\left\|Q_i(t)\right\|_2^2$ in succession.

Upper bound on $\mathbb{E}\left\|Q_1(t)\right\|_2^2$: Assumption 1 and Assumption 3 together imply that

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

In step (i), we are using the fact that $0 \preceq \nabla^2 F(\rho\theta_{t-1} + (1-\rho)\theta_{t-2}) \preceq LI_d$, and in step (ii), we use the strong convexity of F.

For $\eta < \frac{1}{2L} \wedge \frac{\mu}{2\ell_\Xi^2}$, we have $\mathbb{E}\left\|Q_1(t)\right\|_2^2 \le (1-\mu\eta)\mathbb{E}\left\|e_{t-1}\right\|_2^2$.

Upper bound on $\mathbb{E}\left\|Q_2(t)\right\|_2^2$: By Assumption 3 and Eq. (80), we have

$$\mathbb{E}\left\|Q_2(t)\right\|_2^2 \le \ell_\Xi^2 \mathbb{E}\left\|\theta_{t-1}-\theta^*\right\|_2^2 \le \frac{a_0\ell_\Xi^2}{t},$$

where the last inequality follows from Theorem 1.

Upper bound on $\mathbb{E}\left\|Q_3(t)\right\|_2^2$: Applying Lemma 18 with $\widetilde{\theta}_{t-1} := \rho\theta_{t-1}+(1-\rho)\theta_{t-2} \in \mathcal{F}_{t-1}$, we have

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

Putting the bounds for (Q_1, Q_2, Q_3) together, we obtain:

$$\sqrt{\mathbb{E}\left\|e_t\right\|_2^2} \le \left(1-\frac{\eta\mu}{2}\right)\sqrt{\mathbb{E}\left\|e_{t-1}\right\|_2^2}+\frac{4c_1^{1/2}a_0^{1/4}}{t^{1/4}}+\frac{\ell_\Xi\sqrt{a_0}}{\sqrt{t}}.$$

Solving the recursion, we have

$$\sqrt{\mathbb{E}\left\|e_T\right\|_2^2} \le (4c_1^{1/2}a_0^{1/4}+\ell_\Xi\sqrt{a_0})\sum_{s=T_0+1}^{t} s^{-\frac{1}{4}}\exp\left(-\frac{\mu\eta}{2}(T-s)\right)+e^{-\frac{\mu\eta(T-T_0)}{2}}\sqrt{\mathbb{E}\left\|e_{T_0}\right\|_2^2}.$$

For the first term, we note that:

$$\begin{aligned}\sum_{s=T_0+1}^{T} s^{-\frac{1}{4}}\exp\left(-\frac{\mu\eta}{2}(T-s)\right) &\le \sum_{s=1}^{T/2}\exp\left(-\frac{\mu\eta}{2}T\right)+\frac{1}{(T/2)^{1/4}}\sum_{s=T/2}^{T} e^{-\frac{\mu\eta(T-s)}{2}}\\ &\le \frac{T}{2}e^{-\frac{\mu\eta T}{2}}+\frac{4}{\mu\eta T^{1/4}}.\end{aligned}$$

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 $(a_0, c_1, a', \eta, \mu, T_0)$ but independent of t, such that

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

which finishes the proof.

F.2. Proof of Lemma 18

Observe that Assumption (CLT.A) guarantees that

$$\mathbb{E}\left(\left\|(H_t(\theta^*)-H_t(\widetilde{\theta}_{t-1}))y_{t-1}\right\|_2^2\,\middle|\,\mathcal{F}_{t-1}\right) \le \beta^2\left\|\widetilde{\theta}_{t-1}-\theta^*\right\|_2^2\cdot\left\|y_{t-1}\right\|_2^2.$$

On the other hand, by Assumption 3, we have

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

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

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

where step (i) follows from the Cauchy-Schwarz inequality. Applying Lemma 19, we are guaranteed the existence of a constant $a' > 0$ such that

$$\sup_{t \geq T_0} \mathbb{E}\left\|y_t\right\|_2^4 \leq a' < \infty.$$

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

F.3. Proof of Lemma 19

Throughout this section, we adopt the shorthand notation $H_t := H_t(\theta^*)$ and $\Xi_t := \Xi_t(\theta^*)$. We also use Ξ to denote a generic random variable have the same law as Ξ_1. Beginning with the proof of the first claim, we take expectations on both sides of Eq. (79), thereby finding that

$$\mathbb{E}(y_t) = \mathbb{E}\left(y_{t-1} - \eta H_t(\theta^*) y_{t-1} + \varepsilon_t(\theta^*)\right) = (I - \eta H^*)\mathbb{E}(y_{t-1}) = (I - \eta H^*)^{t-T_0}\mathbb{E}(y_{T_0}) = 0.$$

Our next step is to control the fourth moment. For $\eta \leq \frac{1}{2L} < \frac{1}{2\mu}$, we observe that:

$$\begin{aligned}\mathbb{E}\left\|y_t\right\|_2^4 &= \mathbb{E}\left\|y_{t-1} - \eta H_t(\theta^*) y_{t-1} + \varepsilon_t\right\|_2^4 \\ &\leq \mathbb{E}\left\|(I - \eta H_t) y_{t-1}\right\|_2^4 + 4\mathbb{E}(\|(I - \eta H_t) y_{t-1}\|_2^3 \cdot \|\varepsilon_t\|_2) + 6\mathbb{E}(\|(I - \eta H_t) y_{t-1}\|_2^2 \cdot \|\varepsilon_t\|_2^2) \\ &\quad + 4\mathbb{E}(\|\varepsilon_t\|_2^3 \cdot \|(I - \eta H_t) y_{t-2}\|_2) + \mathbb{E}\left\|\varepsilon_t\right\|_2^4 \\ &\overset{(i)}{\leq} \left(1 + \frac{\eta\mu}{2}\right)\mathbb{E}\left\|(I - \eta H_t) y_{t-1}\right\|_2^4 + \frac{24}{(\eta\mu)^3}\mathbb{E}\left\|\varepsilon_t\right\|_2^4 + \frac{216}{(\eta\mu)^2}\mathbb{E}\left\|\varepsilon_t\right\|_2^4 + \frac{24}{(\eta\mu)}\mathbb{E}\left\|\varepsilon_t\right\|_2^4 + \mathbb{E}\left\|\varepsilon_t\right\|_2^4 \\ &\leq \left(1 + \frac{\eta\mu}{2}\right)\mathbb{E}\left\|(I - \eta H_t) y_{t-1}\right\|_2^4 + \frac{157}{(\mu\eta)^3}\mathbb{E}\left\|\varepsilon(\theta^*)\right\|_2^4,\end{aligned}$$

where in step (i), we use Young's inequality for the last four terms.

Now we study the term $\mathbb{E}\left\|(I - \eta H_t) y_{t-1}\right\|_2^4$. For $\eta < \frac{1}{L}$, straightforward calculation yields:

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

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 $\mathbb{E}\left(\|(I - \eta H_t)y_{t-1}\|_2^4 \mid \mathcal{F}_{t-1}\right) \leq (1 - 2\mu\eta)\|y_{t-1}\|_2^4$. 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 $\mathbb{E}\|y_{T_0}\|_2^4 = 0$. Solving this recursion leads to the bound

$$\sup_{t \geq T_0} \mathbb{E}\|y_t\|_2^4 \leq \frac{157}{(\mu\eta)^4}\mathbb{E}\|\varepsilon(\theta^*)\|_2^4.$$

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

Finally we study the stationary covariance of the process $\{y_t\}_{t \geq T_0}$. The existence and uniqueness of the stationary distribution was established in (Mou et al., 2020). Let π_η denote the stationary distribution of $(y_t)_{t \geq T_0}$, and let $Q_\eta := \mathbb{E}_{Y \sim \pi_\eta}(YY^\top)$. From the first part of this lemma, we can see that $\mathbb{E}_{Y \sim \pi_\eta}(Y) = 0$. For $y_t \sim \pi_\eta$, we have $y_{t+1} \sim \pi_\eta$, and consequently,

$$\begin{aligned}
Q_\eta &= \mathbb{E}(y_{t+1}y_{t+1}^\top) \\
&= \mathbb{E}\left((I - \eta H_{t+1})y_t y_t^\top (I - \eta H_{t+1}^\top) + \varepsilon_{t+1}\varepsilon_{t+1}^\top\right) + \mathbb{E}\left(\varepsilon_{t+1}y_t^\top (I - \eta H_{t+1}^\top) + (I - \eta H_{t+1})y_t\varepsilon_{t+1}^\top\right) \\
&= Q_\eta - \eta(H^*Q_\eta + Q_\eta H^*) + \eta^2(H^*Q_\eta H^* + \mathbb{E}(\Xi Q_\eta \Xi)) + \Sigma^*.
\end{aligned}$$

In the last equation, we use the fact that $\mathbb{E}(y_t) = 0$ and that y_t is independent of $(H_{t+1}, \varepsilon_{t+1})$, which leads to the following equation:

$$\mathbb{E}\left(\varepsilon_{t+1}y_t^\top (I - \eta H_{t+1}^\top)\right) = \mathbb{E}\left(\varepsilon_{t+1}(\theta^*) \otimes (I - \eta H_{t+1}(\theta^*))\right)[\mathbb{E}(y_t)] = 0.$$

Therefore, the matrix Q_η satisfies the equation

$$H^*Q_\eta + Q_\eta H^* - \eta(H^*Q_\eta H^* + \mathbb{E}(\Xi Q_\eta \Xi)) = \frac{\Sigma^*}{\eta},$$

which completes the proof of the last part of the lemma.