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

author names withheld

Under Review for OPT 2024

Abstract

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

1. Introduction

Stochastic optimization has become a cornerstone in machine learning and statistical learning, particularly for large-scale and high-dimensional data. Among various stochastic optimization techniques, *stochastic gradient descent* (SGD) stands out due to its simplicity and effectiveness [62]. However, the performance of SGD is heavily influenced by the stepsize schedule, which determines the balance between convergence speed and stability.

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

In this paper, we revisit ROOT-SGD, recently studied by [41], a novel optimization framework that enhances both the convergence and stability of stochastic gradient methods. ROOT-SGD is designed to achieve theoretical optimality and practical effectiveness, producing estimators that exhibit the same *optimal statistical properties* as empirical risk minimizers. The estimator produced by the ROOT-SGD algorithm retains these optimal properties in both asymptotic and non-asymptotic settings.

The notion of statistical efficiency allows for a rigorous assessment of optimality. Furthermore, local asymptotic minimax theorems demonstrate that, under any bowl-shaped loss function, the optimal asymptotic distribution is Gaussian [20, 73]. The asymptotic covariance reflects the local complexity, and it is desirable to achieve this optimal bound with a *unity* pre-factor. Under relatively mild conditions, the empirical risk minimizer achieves this.

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

Problem formulation In particular, given a function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ that is differentiable as a function of its first argument, consider the unconstrained minimization problem

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{for a function of the form } F(\theta) := \mathbb{E}\big[f(\theta;\xi)\big]. \tag{1}$$

Here the expectation is taken over a random vector $\xi \in \Xi$ with distribution $\mathbb{P}$. Throughout this paper, we consider the case where F is strongly convex and smooth. Suppose that we have access to an oracle that generates samples $\xi \sim \mathbb{P}$. Let θ^* denote the minimizer of F, we defined the matrices $H^* := \nabla^2 F(\theta^*)$ and $\Sigma^* := \mathbb{E}\left[\nabla f(\theta^*;\xi)\nabla f(\theta^*;\xi)^\top\right]$. Under certain regularity assumptions, given $(\xi_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, the following asymptotic limit holds true for the exact minimizer of empirical risk:

$$\widehat{\theta}_n^{\mathrm{ERM}} := \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n f(\theta, \xi_i) \quad \text{satisfies } \sqrt{n}\left(\widehat{\theta}_n^{\mathrm{ERM}} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, (H^*)^{-1}\Sigma^*(H^*)^{-1}\right). \tag{2}$$

Furthermore, the asymptotic distribution (2) is known to be *locally optimal*—see [73] and [20] for the precise statements about the optimality claim. The question naturally arises: ***can a stochastic optimization algorithm, taking the sample ξ_i as input in its i-th iteration without storing it, achieve the optimal guarantee as in equation (2)?***

An affirmative answer to this question at least qualitatively, is provided by the seminal work by [60, 61, 65]. In particular, they show that by taking the Cesáro-average of the stochastic gradient descent (SGD) iterates, one can obtain an optimal estimator that achieves locally minimax limit (2), as the number of samples grows to infinity. This algorithm lays the foundations of online statistical inference [9, 69] and fine-grained error guarantees for stochastic optimization algorithms [17, 53]. However, the gap still exists between the averaged SGD algorithm and the exact minimizer of empirical risk, both asymptotically and non-asymptotically. The following questions remain unresolved:

- The asymptotic properties of the estimators produced by the Polyak-Ruppert algorithm are derived under the Lipschitz or Hölder condition of the Hessian matrix $\nabla^2 F$, at least with respect to the global optimum θ^* in all existing literature (see, e.g., [20, 61]). However, the asymptotic guarantee (2) for the exact minimizer holds true as long as the matrix-valued function $\nabla^2 F$ is *continuous* at θ^*, along with mild moment assumptions (see, e.g., [73]). On a historical note, the mis-match in the assumptions is particularly undesirable, given a large portion of literature is devoted to identify the optimal smoothness conditions required for the asymptotic normality of M-estimators to admit [38, 73]. ***Is there a (single-loop) stochastic optimization algorithm that achieves the asymptotic guarantee (2) under the mildest smoothness conditions including that the Hessian is continuous but not Hölder continuous at its global optimum?***

- On the non-asymptotic side, one would hope to prove a finite-sample upper bound for the estimator produced by the stochastic optimization algorithm under proper smoothness condition, which matches the exact behavior of the asymptotic Gaussian limit (2) with additional terms that decays faster as $n \to +\infty$. For example, under the one-point Hessian Lipschitz condition, [24, 53, 79] established bounds in the form of

$$\mathbb{E}\left\|\widehat{\theta}_n^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \le \frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + \text{high order terms}, \tag{3}$$

for the Polyak-Ruppert estimator $\widehat{\theta}_n^{\mathrm{PRJ}}$. Under the optimal trade-off, the higher-order terms in their bound scale at the order $O(n^{-7/6})$ and $O(n^{-5/4})$, respectively. Compared to the rates for the M-estimator, these bounds on the additional term do not appear to be sharp or optimal. Under suitable Lipschitz conditions, the natural scaling for the additional term would scale as $O(n^{-3/2})$ (see the discussion following Theorem 6 for details). For quadratic objectives, an argument similar to [41] allows one to achieve an $O(n^{-3/2})$ higher-order term

$$\mathbb{E}\left\|\widehat{\theta}_n^{\mathrm{PRJ}} - \theta^*\right\|_2^2 \le \frac{1}{n}\mathrm{Tr}\left((H^*)^{-1}\Sigma^*(H^*)^{-1}\right) + O\left(\frac{1}{n^{3/2}}\right), \tag{4}$$

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

We answer both questions affirmatively using ROOT-SGD with a diminishing stepsize strategy. In the following, we describe the algorithm and explain the connection and differences between our results and [41].

The ROOT-SGD algorithm with varying stepsizes For the stochastic optimization problem in the strongly-convex and mean-squared smooth setup, [41] recently proposed a stochastic approximation algorithm named *Recursive One-Over-T SGD*, or ROOT-SGD for short. To recap at each iteration $t = 1, 2, \ldots$ ROOT-SGD performs the following steps:

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

$$v_t = \nabla f(\theta_{t-1}; \xi_t) + \frac{t-1}{t}\left(v_{t-1} - \nabla f(\theta_{t-2}; \xi_t)\right), \tag{5a}$$

$$\theta_t = \theta_{t-1} - \eta_t v_t, \tag{5b}$$

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

For the purposes of stabilizing the iterates, Algorithm (5) is initialized with a *burn-in* phase of length $T_0 > 1$, in which only the v variable is updated with the θ variable held fixed. Given some initial vector $\theta_0 \in \mathbb{R}^d$, we set $\theta_t = \theta_0$ for all $t = 1, \ldots, T_0$, and compute

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

1. This concept is also known as the *anytime property* in some literature, such as [24]. We will demonstrate that the standard *doubling strategy* in optimization is inadequate and that a carefully designed multi-loop strategy is necessary.

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

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

Building upon the proof techniques in the non-asymptotic bounds of [41], our work provide fine-grained guarantees for ROOT-SGD, addressing both aforementioned questions immediately before introducing ROOT-SGD with affirmative answers. A key technical novelty is a two-time-scale characterization of the iterates (5) for a diminishing stepsize strategy. This allows us to effectively bound various cross terms in the error decomposition, yielding better bounds than those obtained by naïve application of Young's inequality. In addition, we also propose an improved re-starting schedule for the multi-loop algorithm, achieving exponential forgetting of the initial condition without affecting the statistical efficiency on its leading order term.

Contributions Let us summarize the contributions of this paper:

- On the asymptotic side, we show in Theorem 1 that ROOT-SGD with a wide range of diminishing stepsize sequence converges asymptotically to the optimal Gaussian limit as $n \to +\infty$. Notably, this result only requires strong convexity, smoothness, and a set of noise moment assumptions standard in asymptotic statistics. The result does not require any higher-order smoothness other than the continuity of Hessian matrix at θ^*, another standard condition for asymptotic normality. To our knowledge, this provides a first result for a stochastic approximation algorithm that enjoys asymptotic optimality without additional smoothness conditions and the prior knowledge of n.

- On the contrary, we show that without additional smoothness conditions, a constant stepsize variant of Polyak-Ruppert algorithm fails to converge at a desirable rate, for any feasible scalings of stepsize and burn-in time choices. This manifests the difference in asymptotics between variance-reduced methods and Polyak-Ruppert averaging methods. The result is stated in Theorem 3 serving as complementary to the asymptotic Theorem 1.

- Under the same set of assumptions, in Theorem 4, we establish a non-asymptotic gradient norm upper bound with the optimal leading term that exactly matches the optimal asymptotic risk, plus a higher-order term that scales as $O(n^{-4/3})$. When restarting is employed with an appropriate schedule, the resulting upper bound measured in gradient norm is of unity prefactor (arbitrarily close to 1) of the optimal asymptotic risk, with exponentially-decaying additional terms.

- In addition, when the one-point Hessian Lipschitz at the global optimum θ^* and certain fourth-moment conditions are assumed, in Theorem 6, we show an upper bound on the mean-squared error (MSE) in the form of (3). Taking an optimal trade-off leads to a higher-order term that scales as $O(n^{-3/2})$ as $n \to +\infty$ with a sharp problem-specific prefactor, and such a bound is achieved without the prior knowledge of n. With some efforts, we also establish a similar upper bound on the excess risk in Theorem 7.

Organization This paper is organized as follows. §2 describes the asymptotic normality results of ROOT-SGD and also the sub-optimality of Polyak-Ruppert averaging under the Hessian continuity assumption at the optimum. §3 state the non-asymptotic upper bound results on the gradient norm and also the estimation error. Additional related works, proofs and discussions are delegated to the supplementary materials.

2. Asymptotic results

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

$$\varepsilon(\theta;\xi) = \nabla_\theta f(\theta;\xi) - \nabla F(\theta), \tag{6}$$

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

Assumption 1 *The population objective function F is μ-strongly-convex and L-smooth.*

Assumption 2 *The noise function $\theta \mapsto \nabla_\theta f(\theta,\xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\,\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\|_2^2 \le \ell_\Xi^2\,\|\theta_1 - \theta_2\|_2^2 \qquad \text{for all pairs } \theta_1,\theta_2 \in \mathbb{R}^d. \tag{7}$$

Assumption 3 *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has a positive definite covariance matrix; hence $\sigma_*^2 := \mathbb{E}\,\|\nabla f(\theta^*;\xi)\|_2^2$ is positive and finite.*

Assumption 4 *The Hessian matrix $\nabla^2 F(\theta)$ is continuous at the optimum θ^*, i.e.,*

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

Assumption 2 (sometimes referred to as *mean-squared-smoothness*) as well as Assumptions 3 and 4 are standard ones needed for proving asymptotic normality of M-estimators and Z-estimators (see, e.g., [73], Theorem 5.21). They are satisfied by a broad class of statistical models and estimators. Note that we assume only the continuity of Hessian matrix at θ^*, without assuming any bounds on its modulus of continuity. This requires merely slightly more than second-order smoothness, and is usually considered as the minimal assumption needed in the general setup. The weak condition manifests the difference between ROOT-SGD and Polyak-Ruppert averaging procedure.

The strong convexity and smoothness Assumption 1 is a global condition stronger than those typically used in the asymptotic analysis of M-estimators. They are needed for the fast convergence of the optimization algorithm, and makes it possible to establish non-asymptotic bounds. Finally, we note that in making Assumption 2, we separate the stochastic smoothness of the noise $\varepsilon(\theta,\xi) = \nabla f(\theta,\xi) - \nabla F(\theta)$ with the smoothness of the population-level objective itself. The magnitude of ℓ_Ξ and L is not comparable in general. This flexibility allows, for example, mini-batch algorithms where the population-level Lipschitz constant L remains the same but the parameter ℓ_Ξ decreases with batch-size. This setting is called *Lipschitz stochastic noise* (LSN) in [41], which requires weaker conditions than the *individual smooth and convex* (ISC) setting in their paper.

2.1. Asymptotic normality

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

Theorem 1 *Under Assumptions 1, 2 and 3, there exists universal constants $c, c_1 > 0$, such that for any $\alpha \in (0,1)$, ROOT-SGD with burn-in time $T_0 = c(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2})$ and stepsize sequence $\eta_t = \frac{1}{\mu T_0^{1-\alpha} t^\alpha}$ for $t \geq T_0$ satisfies the asymptotic limit:*

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

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

See §B.2 for the proof of this theorem. En route to the proof of this asymptotic guarantee, we establish non-asymptotic bounds on the second moments of the processes $(\theta_t, v_t, z_t)_{t \geq T_0}$, where a central object in our analysis is the *tracking error process*:

$$z_t := v_t - \nabla F(\theta_{t-1}) \qquad \text{for } t \geq T_0. \tag{8}$$

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

Proposition 2 *Under Assumptions 1, 2, and 3, there exist universal constants $c_1, c_2, C > 0$, using burn-in time $T_0 \geq C(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu})$, if the step sequence is non-increasing, and $\frac{c_1}{\mu t} < \eta_t < c_2(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L})$ when $t > T_0$. We have the following bounds for any $T \geq 2T_0 \log T_0$:*

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

See §B.1 for the proof of this claim.

A few remarks are in order. First, we note that this limiting distribution is locally asymptotically optimal (see, e.g., [20]). This result for diminishing stepsize sequence is complementary to the constant-stepsize result in the paper [41], where the asymptotic covariance is inflated by a stepsize-dependent matrix.[2] Moreover, our method achieves optimal asymptotic covariance in a single loop and is agnostic to the knowledge of n in advance, enhancing its practicality. Theorem 6 allows for flexible choice of stepsize decaying rate $\alpha \in (0,1)$, albeit requiring knowledge about the structural parameters (L, ℓ_Ξ, μ). This requirement, on the other hand, can be relaxed with some efforts: given a stepsize sequence $\eta_t = h_0 t^{-\alpha}$ for some $h_0 > 0$ and arbitrary constant burn-in time, the iterates may suffer from exponential blow-up for constant number of steps, but will eventually decay at the desired rate, leading to the same asymptotic results. We omit this for simplicty. In contrast to the asymptotic guarantees by the Polyak-Ruppert averaging scheme [61, 65], Theorem 1 requires no quantitative Lipschitz or Hölder assumptions on the Hessian matrix ∇F, while requiring a stochastic continuity condition (Assumption 2) on the stochastic gradient. As we will see in the next subsection, in contrast to our guarantees, the Polyak-Ruppert procedure is asymptotically sub-optimal for a function within the given class.

2. In the meantime, the asymptotic normality result for multi-loop ROOT-SGD in [41] admits a triangular array format ($n \to \infty$, $\eta \to 0$ with $\frac{\eta n}{\log(\eta^{-1})} \to \infty$), which can be difficult to interpret and impractical for practitioners, and undesirably necessitates knowledge of n aprior.

2.2. Asymptotic sub-optimality of Polyak-Ruppert averaging

In this section, we explicitly construct a problem instance under above set-up, for which Polyak-Ruppert procedure fails to converge to the optimal asymptotic distribution. In conjunction with Theorem 1, this exhibits an asymptotic separation between Polyak-Ruppert averaging and ROOT-SGD.

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

$$\theta_t = \theta_{t-1} - \eta_t \nabla f(\theta, \xi) \qquad \text{for } t = 1, 2, \ldots, \tag{9a}$$

$$\overline{\theta}_T = \frac{1}{T - T_0} \sum_{t=T_0}^{T-1} \theta_t. \tag{9b}$$

We consider a simple special case where the stepsize sequence is constant and fixed in advance, depending on the number of iterations in the algorithm. For the algorithm with T iterations, we consider stepsize $\eta_t = \eta = \eta_0 T^{-\alpha}$ for some constant $\eta_0 > 0$ and $t = 1, 2, \ldots$. This simplification makes the iterate (9a) a time-homogeneous Markov process, which is amendable to our analysis. Such a simplification has been employed in existing literature [2, 17], and the constant-stepsize algorithm usually behaves qualitatively similar to the one with diminishing stepsize $\eta_t = \eta_0 t^{-\alpha}$.

The following theorem shows the asymptotic sub-optimality of the estimator (9), even if started from the optimum, for any choice of burn-in period and step size.

Theorem 3 *There exists a function $F : \mathbb{R} \to \mathbb{R}$ that satisfies Assumptions 1 and 4 with constants $(\mu = 1, L = 2)$ and noise model $f(\cdot, \xi)$ satisfying Assumptions 2 and 3 with constants $(\ell_\Xi = 0, \sigma_* = 2)$. For any $\alpha \in [0, 1)$, $\beta \in [0, 1)$ and $\eta_0 > 0, S_0 > 0$, the procedure (9) starting from $\theta_0 = \theta^*$, with step size $\eta = \eta_0 T^{-\alpha}$ and burn-in time $T_0 = S_0 T^\beta$ leads to the following limit:*

$$\lim_{T \to +\infty} T \cdot \mathbb{E} \left\| \overline{\theta}_T - \theta^* \right\|_2^2 = +\infty. \tag{10}$$

See §B.3 for the proof of this theorem.

Note that Theorem 3 shows that without the Hessian Lipschitz condition, the Polyak-Ruppert algorithm does not even converge with the desired rate, let alone the optimal asymptotic distribution. The proof is done via an explicit construction of a pathological function. With the Hessian Lipschitz condition removed, one could construct a strongly convex and smooth function, whose second derivative has a *sharp spike* at the optimum θ^*. This will break the local linearization arguments for the proof of Polyak-Ruppert algorithm. By employing recent progress in the analysis of MCMC algorithms [21], we can furthermore show that this leads to large bias that cannot be corrected using averaging. On the other hand, for ROOT-SGD, not only the asymptotic guarantees in Theorem 1 but also the non-asymptotic bounds on the gradient norm in Theorem 4 works. Moreover, note that [61] considered the case where the Hessian matrix is λ-Hölder at θ^*, and allows for stepsize choice $\eta_t \propto t^{-\alpha}$ for $\alpha \in [1 - \lambda, 1)$. Theorem 3 can be extended to show that stepsize outside this range does not yield the correct rate. The construction we exploit, on the other hand, is by driving λ to 0 so that no stepsize choice is allowed.

3. Non-asymptotic results

In this section, we present the non-asymptotic results. We first establish sharp bounds on the gradient norm with near-unity pre-factor on the optimal complexity term, and exponentially decaying additional term. Then, we establish an estimation error bound with the pre-factor being unity and the additional term decaying as $n^{-3/2}$. Note that the former result holds true under exactly the same assumptions as needed in §2, while the latter requires additional conditions, as with existing literature [24, 53].

3.1. Upper bounds on the gradient norm

Recall the decomposition $\nabla F(\theta_t) = v_{t+1} - z_{t+1}$, it is easy to see that Proposition 2 implies the following bound on the gradient norm of the last iterate:

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

When taking largest possible stepsize $\eta = c\big(\frac{1}{L} \wedge \frac{\mu}{\ell_\Xi^2}\big)$, this bound matches the gradient norm bound in the original ROOT-SGD paper [41], up to logarithmic factors in the high-order term. Our bound allows a more flexible choice of diminishing stepsizes. This flexibility allows us to achieve the exact asymptotically optimal limiting covariance, as opposed to the slightly larger covariance in the constant stepsize regime [41]. More importantly, this allows us to tune the stepsize sequence in order to address the optimal trade-off between fast convergence and small variance in the asymptotic limit. Note that the pre-factor in the leading term σ_*^2/T is *not* unity. However, owing to the inherent martingale structure in the process $(z_t)_{t\geq T_0}$,[3] one could extract the main part of the variance and bound the additional parts using Proposition 2. The multiplicative constant in such bounds will only contribute to the high-order terms in the final conclusion. See Theorem 4 and its proofs for details.

Note that the bounds in Proposition 2 depends on the initial condition $\|\nabla F(\theta_0)\|_2^2$ with polynomially-decaying factor T^{-2} and $T^{-3}\eta_T^{-2}$. For the algorithm ROOT-SGD, this cannot be avoided in general, as the stochastic gradients from initial rounds are being counted in the averaging process. On the other hand, this issue can be easily mitigated by *re-starting* the process for a few epochs. In Algorithm 1, we present a cold-start version of the algorithm. The algorithm consists of $\mathscr{E}$ short epochs and one long epoch. Each short epoch only uses constant number of data points, while the long epoch uses the rest of data points.

Theorem 4 *Under above set-up, given $\alpha \in (0,1)$, there exists constants $c_1 > 0$ depending only on α, such that the iterates (5) with any burn-in time $T_0 \geq c\big(\frac{\ell_\Xi^2}{\mu^2} + \frac{L}{\mu}\big)$ and stepsize sequence $\eta_t = \frac{1}{c\mu T_0^{1-\alpha} t^\alpha}$ satisfies the bound:*

$$\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 \leq \left(1 + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\right)\cdot\frac{\sigma_*^2}{T} + c\log T\cdot\left(\frac{T_0}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\left\|\nabla F(\theta_0)\right\|_2^2. \tag{11a}$$

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

$$\mathbb{E}\left\|\nabla F(\widehat{\theta}_n)\right\|_2^2 \leq \left(1 + c\left(\frac{T_0}{n}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\log^2 n\right)\frac{\sigma_*^2}{n}. \tag{11b}$$

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

Algorithm 1 ROOT-SGD with cold start

1: **Input:** Burn-in time T_0, stepsize sequence $(\eta_t)_{t \geq T_0}$, number of restart epochs $\mathscr{E}$, initial point θ_0
2: Set initial point for first epoch $\theta_0^{(1)} = \theta_0$
3: **for** $b = 1, 2, \ldots, \mathscr{E}$ **do**
4: Run ROOT-SGD with burn-in time T_0, initial point $\theta_0^{(b)}$ and stepsize $\eta_t := \frac{c}{\mu T_0}$ for $T^\flat := cT_0 \log T_0$ iterations, and obtain the sequence $\left(\theta_t^{(b)}\right)_{t=T_0+1}^{T^\flat}$
5: Set the initial point $\theta_0^{(b+1)} := \theta_{T^\flat}^{(b)}$ for the next round
6: **end for**
7: Run ROOT-SGD for $T := n - \mathscr{E} T^\flat$ rounds with stepsize sequence $(\eta_t)_{t \geq T_0}$ and burn-in period T_0, and output the last iterate $\widehat{\theta}_n := \theta_T^{(\mathscr{E}+1)}$
8: **Output:** The last-iterate estimator $\widehat{\theta}_n$

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

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

$$\mathbb{E} \left\| \nabla F(\widehat{\theta}_n) \right\|_2^2 \leq (1 + \omega) \frac{\mathrm{Tr}(\Sigma^*)}{n} \qquad \text{for } n \geq \frac{c}{\omega^3} \left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2} \right) \log^3 \frac{T_0}{\omega}. \tag{12}$$

In other words, we obtain a near-optimal bound on the gradient norm with $(1 + \omega)$ pre-factor compared to the asymptotic optimal limit, as long as the sample size is larger than the threshold $O\big(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\big)$, up to log factors. We remark that this threshold is also sharp: the term $O\big(\frac{L}{\mu}\big)$ is the number of iterations needed for gradient descent, while the $O\big(\frac{\ell_\Xi^2}{\mu^2}\big)$ term is the smallest sample size needed to distinguish the quadratic function $\frac{\mu}{2} \|x\|_2^2$ from the constant function 0, under the noise Assumption 2. This establish a gradient-norm result complementary to the sub-optimality gap bound in [23]. The gradient norm bound does not require the self-concordant condition needed in [23], and achieves a sharper convergence rate in terms of both the $(1 + \omega)$ factor and the initial condition.[4]

With a potentially sub-optimal choice of $\alpha \in (0, 1)$, one would get a worse exponent in the dependency of n on ω in the bound (12), while the rest parts of the bound remain unchanged. If ω is taken as a constant, the near-optimal bounds are available for the entire range of parameter $\alpha \in (0, 1)$. Finally, we note that the bound (12) lead to an $\widetilde{O}(n^{-4/3})$ bound on the additional term, achieved by the stepsize choice $\eta_t = \frac{1}{c\mu T_0^{2/3} t^{1/3}}$. This rate and step-size choice, however, is not always optimal. In particular, as we will see in the next section, with the one-point Hessian Lipschitz condition on the objective function F, we can obtain an improved $\widetilde{O}(n^{-3/2})$ bound on the additional term.

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

3.2. Upper bounds on the estimation error

To obtain a precise upper bound for the estimation error $\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2$ that matches the asymptotic limit, we need the following one-point Hessian Lipschitz condition, as a quantitative counterpart of the continuity Assumption 4:

Assumption 4′ *There exists $L_2 > 0$, such that for any $\theta \in \mathbb{R}^d$, we have:*

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

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

We also need the following stronger fourth moment conditions for technical reasons. Note that these conditions are also exploited in prior works [24, 53].

Assumption 2′ *The noise function $\theta \mapsto \nabla_\theta f(\theta, \xi)$ in the stochastic gradient satisfies the bound*

$$\mathbb{E}\left\|\varepsilon(\theta_1;\xi) - \varepsilon(\theta_2;\xi)\right\|_2^4 \leq \widetilde{\ell_\Xi}^4 \left\|\theta_1 - \theta_2\right\|_2^4 \qquad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{R}^d. \tag{13}$$

Assumption 3′ *At the optimum θ^*, the stochastic gradient noise $\varepsilon(\theta^*;\xi)$ has bounded fourth moment: $\widetilde{\sigma_*}^4 := \mathbb{E}\left\|\nabla f(\theta^*;\xi)\right\|_2^4$ is finite.*

By Hölder's inequality, it is clear that the constants in Assumptions 2′ and 3′ are larger than their second-moment counterparts, i.e., $\ell_\Xi \leq \widetilde{\ell_\Xi}$ and $\sigma_* \leq \widetilde{\sigma_*}$.

Under the fourth moment conditions, we can establish the following fourth-moment bounds for the processes z_t and v_t, analogous to the second-moment results in Proposition 2.

Proposition 5 *Under Assumptions 1, 2′, and 3′, there exist universal constants $c_1, c_2, C > 0$, using burn-in time $T_0 \geq C(\frac{\widetilde{\ell_\Xi}^2}{\mu^2} + \frac{L}{\mu})$, if the step sequence is non-increasing, and $\frac{c_1}{\mu t} < \eta_t < c_2(\frac{\mu}{\ell_\Xi^2} \wedge \frac{1}{L})$ when $t > T_0$. We have the following bounds for any $T \geq 2T_0 \log T_0$:*

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

See §C.2 for the proof of this claim.

Compared to Proposition 2, the variance parameters (σ_*, ℓ_Ξ) are replaced with their fourth-moment counterparts $(\widetilde{\sigma_*}, \widetilde{\ell_\Xi})$. These fourth-moment estimates are utilized to control the error induced by approximation the estimation error $\theta_T - \theta^*$ using the pre-conditioned gradient $(H^*)^{-1}\nabla F(\theta_T)$. As with the case of Proposition 2, these terms appear only in the high-order terms of Theorem 6.

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

$$\mathcal{H}_T^{(\nabla)} := \log T \cdot \frac{\sigma_*^2}{T}\left(\frac{T_0}{T}\right)^{\alpha \wedge 1-\alpha} + \log T \cdot \left\|\nabla F(\theta_0)\right\|_2^2 \left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha}, \tag{14a}$$

$$\widetilde{r}_T := \frac{\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{T_0}{T}} \cdot \big(\|\nabla F(\theta_0)\|_2^4\big)^{1/4}, \qquad \text{and} \tag{14b}$$

$$\mathcal{H}_n^{(\sigma)} := \frac{\sigma_*^2 \log^2 n}{\lambda_{\min}(H^*)^2 n}\left(\frac{T_0}{n}\right)^{\alpha\wedge 1-\alpha} + \frac{L_2\widetilde{\sigma_*}^3 \log^2 n}{\lambda_{\min}(H^*)\mu^3 n^{3/2}} + \frac{L_2^2\widetilde{\sigma_*}^4 \log^2 n}{\lambda_{\min}(H^*)^2\mu^4 n^2}. \tag{14c}$$

The term $\mathcal{H}_T^{(\nabla)}$ is part of the high-order term that appears in the bound for the gradient norm. It is indeed the upper bound for the *superfluous* part of the noise in the processes $(z_t)_{t\geq T_0}$ and $(v_t)_{t\geq T_0}$, without taking into account the cross term $\mathbb{E}\langle z_t,\, v_t\rangle$. The quantity $\widetilde{r}_T$ is a coarse upper bound on the convergence rate $\|\theta_t - \theta^*\|_2$ in terms of the fourth moment. In combination with the one-point Hessian Lipschitz Assumption 4′, this quantity controls the additional *linearization error* induced by relating the non-asymptotic behavior of the gradient to the iterates. Finally, the term $\mathcal{H}_n^{(\sigma)}$ is used to characterize the high-order terms for the error in the multi-loop estimator produced by Algorithm 1.

Theorem 6 *Under Assumptions 1, 2′, 3′ and 4′, there exists universal constant* $c, c_1 > 0$*, for burn-in-time* $T_0 = c\left(\frac{\widetilde{\ell_\Xi}^2}{\mu^2} + \frac{L}{\mu}\right)$ *and stepsize* $\eta_t = \frac{c_1}{\mu T_0^{1-\alpha} t^\alpha}$ *for* $t \geq T_0$*, we have the following bounds holding true for* $t \geq 2T_0 \log T_0$*:*

$$\mathbb{E}\,\|\theta_T - \theta^*\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\mathcal{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)^2} + \frac{cL_2\widetilde{r}_T^3}{\lambda_{\min}(H^*)} + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)^2}. \tag{15a}$$

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

$$\mathbb{E}\left\|\widehat{\theta}_n - \theta^*\right\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{n} + c\mathcal{H}_n^{(\sigma)}. \tag{15b}$$

See §C.3 for the proof of this theorem.

A few remarks are in order. First, we note that the asymptotically optimal $\frac{1}{n}\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)$ variance is achieved with the exact pre-factor 1. Taking the optimal stepsize choice with $\alpha = 1/2$, the high order term scales as $O(n^{-3/2})$ in both bounds (15a) and (15b). This is made possible by the stochastic Lipschitz condition for the gradient noise, and strictly improves existing bounds of $O(n^{-7/6})$ in the paper [53] and the $O(n^{-5/4})$ bound in the paper [24, 79]. It is easy to see that the bound (15b) is obtained by removing the terms depending on the initial condition, up to logarithmic factors in the additional term. This is natural because the initial condition is forgotten exponentially fast in the first $\mathscr{E}$ restarting epochs of Algorithm 1. Finally, when taking the optimal parameter $\alpha = 1/2$, the three high-order terms in the expression of $\mathcal{H}_n^{(\sigma)}$ have a clean interpretation.

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

$$\mathbb{E}\,\big\|(H^*)^{-1}\varepsilon_t(\theta)\big\|_2^2$$

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

The above derivations is tight in the worst case. Compared to the term $\widetilde{O}\left(\frac{\sigma_*^2\sqrt{T_0}}{\lambda_{\min}(H^*)^2 n^{3/2}}\right)$ in our bound (15b), the difference is that we replace $\frac{\ell_\Xi^2}{\mu^2}$ with $T_0 = c\left(\frac{L}{\mu} + \frac{\ell_\Xi^2}{\mu^2}\right)$ and is optimal up to a polylogarithmic factor when $\frac{L}{\mu} \lesssim \frac{\ell_\Xi^2}{\mu^2}$.

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

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

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

The arguments in the proof of Theorem 6 indeed applies to any function that is *locally quadratic* around θ^*. Applying it to the function F itself, we arrive at the following theorem:

Theorem 7 *Under the same setup as in Theorem 6, we have the following bounds on the excess risk:*

$$
\mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) \le \frac{\mathrm{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{2T} + \frac{c\mathcal{H}_T^{(\nabla)}}{\lambda_{\min}(H^*)} + cL_2\widetilde{r}_T^3 + \frac{cL_2\widetilde{r}_T^4}{\lambda_{\min}(H^*)}, \tag{16a}
$$

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

$$
\mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) \le \frac{\mathrm{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{2n} + c\lambda_{\min}(H^*)\cdot\mathcal{H}_n^{(\sigma)}. \tag{16b}
$$

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

Note that under the one-point Hessian-Lipschitz Assumption 4′, the leading-order term $\frac{\mathrm{Tr}\left(\Sigma^*(H^*)^{-1}\right)}{2n}$ is the asymptotic risk under the limiting Gaussian distribution. The high-order terms in Theorem 7 differ from those in Theorme 6 by a factor of $\lambda_{\min}(H^*)$. This bound replaces the self-concordance assumption in [23] with a less structural one-point Hessian-Lipschitz condition. Theorem 7 and their results are not comparable in general, as they are based on different assumptions. When taking the optimal trade-off, Theorem 7 leads to an $O(n^{-3/2})$ high-order term in addition to the sharp leading-order one. This result matches the bounds for ERM in [23], and improves the bounds for streaming SVRG in [23] in terms of the rate of convergence for the additional term.

References

[1] Sébastien Arnold, Pierre-Antoine Manzagol, Reza Babanezhad Harikandeh, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. *Advances in Neural Information Processing Systems*, 32:5391–5402, 2019.

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

[3] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 26:773–781, 2013.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

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

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

[8] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

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

[10] Xi Chen, Weidong Liu, and Yichen Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 117(540):1858–1874, 2022.

[11] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods. *Journal of the American Statistical Association*, pages 1–24, 2024.

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

[13] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in Neural Information Processing Systems*, 32:15210–15219, 2019.

[14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27:1646–1654, 2014.

[15] John E Dennis and Jorge J Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28(126):549–560, 1974.

[16] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18 (101):1–51, 2017. URL http://jmlr.org/papers/v18/16-335.html.

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

[18] Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under Markov randomness. *arXiv preprint arXiv:2003.10973*, 2020.

[19] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

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

[21] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

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

[23] 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. PMLR, 2015.

[24] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

[25] Richard D Gill and Boris Y Levit. Applications of the van Trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.

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

[27] Dongyan Lucy Huo, Yudong Chen, and Qiaomin Xie. Effectiveness of constant stepsize in Markovian LSA and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20447–20455, 2024.

[28] 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, 2018.

[29] 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(223):1–42, 2018.

[30] Yanhao Jin, Tesi Xiao, and Krishnakumar Balasubramanian. Statistical inference for Polyak-Ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.

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

[32] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.

[33] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

[34] Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

[37] 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.

[38] Lucien LeCam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.

[39] Sangkyun Lee and Stephen J Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(55):1705–1744, 2012. URL http://jmlr.org/papers/v13/lee12a.html.

[40] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

[41] Chris Junchi Li, Wenlong Mou, Martin J Wainwright, and Michael I Jordan. ROOT-SGD: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. *arXiv e-prints*, 2024.

[42] Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, 70(8):5969–5999, 2024.

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

[44] Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local SGD. In *Conference on Learning Theory*, pages 1613–1661. PMLR, 2022.

[45] Xiang Li, Jiadong Liang, Xinyun Chen, and Zhihua Zhang. Stochastic approximation MCMC, online inference, and applications in optimization of queueing systems. *arXiv preprint arXiv:2309.09545*, 2023.

[46] Xiang Li, Jiadong Liang, and Zhihua Zhang. Online statistical inference for nonlinear stochastic approximation with Markovian data. *arXiv preprint arXiv:2302.07690*, 2023.

[47] Tengyuan Liang and Weijie J Su. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):431–456, 2019.

[48] Yiling Luo, Xiaoming Huo, and Yajun Mei. Covariance estimators for the ROOT-SGD algorithm in online learning. *arXiv preprint arXiv:2212.01259*, 2022.

[49] Sean Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

[50] 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.

[51] 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.

[52] Wenlong Mou, Ashwin Pananjady, and Martin J Wainwright. Optimal oracle inequalities for projected fixed-point equations, with applications to policy evaluation. *Mathematics of Operations Research*, 48(4):2308–2336, 2023.

[53] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24:451–459, 2011.

[54] Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.

[55] Jeffrey Negrea, Jun Yang, Haoyue Feng, Daniel M Roy, and Jonathan H Huggins. Tuning stochastic gradient algorithms for statistical inference via large-sample asymptotics. *arXiv preprint arXiv:2207.12395*, 2022.

[56] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Dokl. Akad. Nauk. SSSR*, 269(3):543–547, 1983.

[57] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.

[58] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

[59] 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. PMLR, 2017.

[60] Boris T Polyak. A new method of stochastic approximation type. *Automat. i Telemekh*, 7 (98-107):2, 1990.

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

[62] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[63] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

[64] Abhishek Roy and Krishnakumar Balasubramanian. Online covariance estimation for stochastic gradient descent under Markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.

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

[66] Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *Conference on Learning Theory*, pages 4511–4547. PMLR, 2024.

[67] Chengchun Shi, Rui Song, Wenbin Lu, and Runze Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116(535):1307–1318, 2021.

[68] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[69] Weijie J Su and Yuancheng Zhu. HiGrad: Uncertainty quantification for online learning and stochastic approximation. *Journal of Machine Learning Research*, 24(124):1–53, 2023.

[70] Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[71] Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1290–1298. PMLR, 2016.

[72] 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. PMLR, 2018.

[73] AW van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

[74] Ziyang Wei, Wanrong Zhu, and Wei Biao Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *arXiv preprint arXiv:2307.06915*, 2023.

[75] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.

[76] Eric Xia, Koulik Khamaru, Martin J Wainwright, and Michael I Jordan. Instance-dependent confidence and early stopping for reinforcement learning. *Journal of Machine Learning Research*, 24(392):1–43, 2023.

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

[78] Chuhan Xie and Zhihua Zhang. A statistical online inference approach in averaged stochastic approximation. *Advances in Neural Information Processing Systems*, 35:8998–9009, 2022.

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

[80] Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *Advances in Neural Information Processing Systems*, 34:4234–4248, 2021.

[81] 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.

[82] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

[83] Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

[84] Yi Zhu and Jing Dong. On constructing confidence region for model parameters in stochastic gradient descent via batch means. In *2021 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2021.

[85] Yuancheng Zhu, Sabyasachi Chatterjee, John Duchi, and John Lafferty. Local minimax complexity of stochastic convex optimization. *Advances in Neural Information Processing Systems*, pages 3431–3439, 2016.

Appendix A. Additional related works

Gradient descent and stochastic gradient descent methods have gained unprecedented popularity in the past decade amidst the era of big data [5–7], driven by the rapid growth of deep learning applications [26]. These methods excel in handling large-scale datasets due to their efficient processing of online samples. A myriad of variants have emerged from both theoretical advancements and practical needs, including variance-reduced methods [14, 31, 63], momentum-accelerated methods [4, 56], second-order methods [15, 58], adaptive gradient methods [19, 35], iteration averaging [61, 65], and coordinate descent [75], among others. The Polyak-Ruppert iteration averaging method [60, 61, 65] and its generalized form [36] have been shown to enhance robustness with respect to step size selection, achieving asymptotic normality with optimal covariance matching local minimax optimality [20, 85]. Recent studies have further explored the non-asymptotic behavior of stochastic gradient descent with iteration averaging [2, 3, 16, 17, 22, 24, 53, 79]. In the studies of linear regression and stochastic approximation, [28, 29, 81] have analyzed the ”tail-averaging” technique, achieving exponential forgetting and optimal statistical risk simultaneously. [37] investigates the Ruppert-Polyak averaging method for general linear stochastic approximation, which extends beyond optimization algorithms to applications in reinforcement learning. Under more stringent noise conditions, [50] establishes Gaussian limit and concentration inequalities for constant stepsize algorithms, with related advancements discussed in [42].

The weak convergence result from [61] has recently been generalized to functional weak convergence by [40] and [44] within the framework of i.i.d. online convex stochastic optimization. However, applying this to nonlinear stochastic approximation with Markovian data introduces several challenges that need addressing [18, 32, 46, 54, 64, 66, 78]. Referenced works beyond this overview delve deeper into topics such as asymptotic normality, statistical inference using gradient-based methods, and variants thereof [8, 10, 11, 27, 30, 33, 43, 45, 47–49, 52, 55, 67, 68, 70, 71, 74, 76, 80, 82–84].

The asymptotic efficiency of variance-reduced stochastic approximation methods has been relatively underexplored in research. [23] introduces an online variant of the SVRG algorithm [31] and establishes a non-asymptotic upper bound on excess risk, aligning its leading term with optimal asymptotics under specific self-concordant conditions on the objective function. [1] proposes *Implicit Gradient Transportation* (IGT) to reduce algorithmic variance. In the context of reinforcement learning for policy evaluation, [34, 51] provides an instance-dependent non-asymptotic upper bound on ℓ_∞ estimation error for variance-reduced stochastic approximation algorithms, matching the risk of the optimal Gaussian limit up to constant or logarithmic factors. Central to our study, [41] introduces the ROOT-SGD algorithm that achieves local minimax optimality. This algorithm can be viewed as an online variant of SARAH [59] and connects with extrapolation-smoothing methods like (N)IGT and STORM [1, 12, 13]. In a different approach, [39, 57, 77] propose dual averaging for the regularized or proximal case.[5] ROOT-SGD distinguishes itself by averaging past stochastic gradients with proper de-bias corrections, achieving both statistical efficiency and non-asymptotic high-order terms.

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

5. See also [20, 72] for manifold first-order optimization methods.

$a \vee b := \max(a, b)$. Throughout the paper, we use the σ-fields $\mathcal{F}_t := \sigma(\xi_1, \xi_2, \cdots, \xi_t)$ for any $t \geq 0$. Due to the burn-in period T_0 introduced before, the stochastic processes are indexed from time $t = T_0$. Given vector-valued martingales $(X_t)_{t \geq T_0}, (Y_t)_{t \geq T_0}$ adapted to the filtration $(\mathcal{F}_t)_{t \geq T_0}$, 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 $(X_t)_{t \geq T_0}$.

Appendix B. Proof of asymptotic results

In this section, we present the proofs for the asymptotic results, Theorem 1 and Theorem 3. The former guarantees the asymptotic normality of ROOT-SGD under our assumptions, while the latter shows an example that satisfies our assumptions but makes Polyak-Ruppert algorithm fail asymptotically. En route our proof, in §B.1 we present the proof of Proposition 2, the non-asymptotic convergence rates for the process $(v_t)_{t \geq T_0}$ and $(z_t)_{t \geq T_0}$. This serves as the basic building block for the fine-grained asymptotic and non-asymptotic guarantees.

B.1. Proof of Proposition 2

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

Lemma 8 *Under Assumption 1, 2, 3, when* $\eta_t \leq \frac{1}{2L} \wedge \frac{\mu}{\ell_\Xi^2}$, *we have:*

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

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

Lemma 9 *Under Assumptions 1, 2 and 3, for* $t \geq 1$, *we have:*

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

The proofs of the Lemmas are postponed to Section D.1 and Section D.2 respectively. Given these lemmas, we now give a proof of this proposition.

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

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

Therefore, by Lemma 8, if t and η_{t-1} satisfies $t\eta_{t-1}\mu > \frac{1}{4C}$, we obtain:

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

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

Consequently, we obtain:

$$t^2\mathbb{E}\left\|v_t\right\|_2^2 \leq (1 - c\eta_{t-1}\mu)(t-1)^2\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{2C}{\mu\eta_{t-1}}\mathbb{E}\left\|z_{t-1}\right\|_2^2 + 2\sigma_*^2 \tag{17}$$

for a universal constant $C > 0$.

Similarly, by Lemma 9, if s satisfies $s\eta_{s-1}\mu > \frac{1}{4C}$ for any $s > T_0$, we have:

$$\begin{aligned}\mathbb{E}\left\|z_t\right\|_2^2 &\leq \frac{T_0^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}(\mathbb{E}\left\|z_s\right\|_2^2 + \mathbb{E}\left\|v_s\right\|_2^2) + \frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1} s^2\eta_s^2\mathbb{E}\left\|v_s\right\|_2^2 \\ &\leq \frac{T_0^2\mathbb{E}\left\|z_{T_0}\right\|_2^2}{t^2} + \frac{2\sigma_*^2}{t} + C\frac{\ell_\Xi^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}\mathbb{E}\left\|z_s\right\|_2^2 + C'\frac{\ell_\Xi^2}{t^2}\sum_{s=T_0}^{t-1} s^2\eta_s^2\mathbb{E}\left\|v_s\right\|_2^2 \end{aligned} \tag{18}$$

for a universal constant $C' > 0$.

Note that the bounds (17) and (18) give recursive upper bounds on the second moments of the processes $(z_t)_{t \geq T_0}$ and $(v_t)_{t \geq T_0}$, i.e., they bound the quantities $\mathbb{E}\left\|z_t\right\|_2^2$ and $\mathbb{E}\left\|v_t\right\|_2^2$ based on their history. In the following, we solve the recursive inequalities.

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

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

First, for any $T > T_0$, by taking the supremum in Eq (18) over $t \in [T_0, T]$, we obtain the following bound:

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

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

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

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

$$H_T \leq 2H_{T_0} + 4\sigma_*^2 + 2C'\ell_\Xi^2\sup_{T_0 \leq t \leq T}\frac{1}{t}\sum_{s=T_0}^{t-1}\eta_{s-1}^2 W_s \tag{19a}$$

On the other hand, the bound (17) implies the bound:

$$W_T \leq (1 - c\eta_{T-1}\mu)W_{T-1} + \frac{C}{T\mu\eta_{T-1}}H_{T-1} + 2\sigma_*^2 \tag{19b}$$

for universal constants $c, C > 0$.

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

Lemma 10 *For a pair of sequences $(H_t)_{t\geq T_0}$ and $(W_t)_{t\geq T_0}$ satisfying the recursive relation (19a) with non-increasing stepsize sequence $(\eta_t)_{t\geq T_0}$. Assuming that $(H_t)_{t\geq T_0}$ is non-decreasing, there exists universal constants $c > 0$, such that for $T \geq T_0$, we have the bound:*

$$H_T \leq c\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \eta_{T_0}}{\mu} W_{T_0} + H_{T_0}\right) \qquad \textit{and} \tag{20a}$$

$$W_T \leq \frac{c}{\eta_T \mu}\sigma_*^2 + c\left(\frac{T_0}{T\mu^2\eta_{T-1}^2} + e^{-\mu\sum_{t=T_0+1}^T \eta_t} T_0^2\right) W_{T_0} \tag{20b}$$

See Section D.3 for the proof of this lemma. Taking this lemma as given, we now proceed with the proof of this proposition.

First, we note that the exponent in the bound (20b) satisfies the bound:

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

For $c_1 \geq 2$ and $\eta_T \leq \frac{c'}{\mu T_0}$, we have that $\frac{T_0}{T\mu^2\eta_{T-1}^2} \geq e^{-\mu\sum_{t=T_0+1}^T \eta_t} T_0^2$. So the bound (20b) implies that:

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

For the process z_t, by substituting the bounds in Lemma 10 into Eq (18), for stepsize $\eta_t < \frac{1}{\mu T_0}$, we obtain:

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

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

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

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

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

Putting them together, we have the bounds:

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

which complete the proof of this proposition.

B.2. Proof of Theorem 1

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

$$\mathbb{E}\left\|v_t\right\|_2^2 \leq a_1\left(\frac{1}{t^2\eta_t}+\frac{1}{t^3\eta_t^2}+\frac{1}{t^2}\right) \leq \frac{3a_1}{t^{2-\alpha}}$$
$$\mathbb{E}\left\|z_t\right\|_2^2 \leq \frac{a_2}{t}+\frac{a_2 \log t}{t^2} \leq \frac{2a_2}{t}$$

and consequently, we have:

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

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

For the martingale Ψ_t, we note that:

$$\mathbb{E}\left\|\Psi_t\right\|_2^2 = \sum_{s=T_0}^{t}(s-1)^2\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\right\|_2^2 \leq \sum_{s=T_0}^{t}(s-1)^2\ell_\Xi^2\mathbb{E}\left\|\theta_{s-1}-\theta_{s-2}\right\|_2^2$$
$$\leq \sum_{s=T_0}^{t}(s-1)^2\eta_{s-1}^2\mathbb{E}\left\|v_{s-1}\right\|_2^2 \leq \frac{1}{\mu^2 T_0^{2-2\alpha}}\sum_{s=0}^{t-1}s^{2-2\alpha}\cdot\frac{3a_1}{s^{2-\alpha}} \leq \frac{3a_1}{(1-\alpha)\mu^2 T_0^{2-2\alpha}}t^{1-\alpha}$$

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

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

Putting together the pieces, we obtain:

$$t\mathbb{E}\left\|z_t-\frac{1}{t}N_t\right\|_2^2 \leq \frac{3}{t}\left\|z_0\right\|_2^2+\frac{3}{t}\mathbb{E}\left\|\Psi_t\right\|_2^2+\frac{3}{t}\mathbb{E}\left\|M_t-N_t\right\|_2^2$$
$$\leq \frac{3}{t}\left\|z_0\right\|_2^2+\frac{3a_1 t^{1-\alpha}}{(1-\alpha)\mu^2 T_0^{2-2\alpha}t}+\frac{3}{t}\cdot\ell_\Xi^2 C\log t \to 0 \tag{21}$$

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

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

The second moment bound (21) implies that:

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

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

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

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

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

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

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

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

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

By Assumption 1, we have $\left\|\nabla F(\theta) - \nabla F(\theta^*)\right\|_2 \geq \mu \left\|\theta - \theta^*\right\|_2$, plugging into above bounds, we obtain $\lim_{\theta \to \theta^*} \frac{\left\|\nabla F(\theta) - H^*(\theta - \theta^*)\right\|_2}{\left\|\nabla F(\theta)\right\|_2} = 0$.

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

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

which finishes the proof.

B.3. Proof of Theorem 3

The proof is by explicit construction of a function (and associated noise) satisfying the Assumptions 1, 2, 3 and 4, for which the Polyak-Ruppert procedure fails.

Consider the following function:

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

Some algebra yields:

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

and

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

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

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

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

We consider an additive Gaussian noise model

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

Clearly, the noise model satisfies Assumption 2 and 3 with constants $\sigma_* = \sqrt{2}$ and $\ell_\Xi = 0$.

Now we consider the SGD update rule on function F:

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

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

$$\forall x > 0 \qquad F_\eta(x) := \eta^{-1} F\big(\sqrt{\eta} x\big) \tag{22}$$

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

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

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

This is an instance of *unadjusted Langevin algorithm* (ULA) on the function F_η, which is known to converge to an approximation to the target density $\pi_\eta \propto e^{-F_\eta}$. More precisely, the following non-asymptotic error bounds are known from the paper [21] (for notational simplicity, we suppress the dependency on the strong convexity and smoothness parameter, as well as the problem dimension, as they are all universal constants in above problem):

Proposition 11 (Special case of [21], Theorem 5) *Under above setup, we have the following bound for* $k = 1, 2 \cdots$

$$\mathcal{W}_2^2(\pi_\eta^{(k)}, \pi_\eta) \leq 2e^{-c_1 \eta k}\big(\|\psi_0\|_2^2 + 1\big) + c_2 \eta \tag{23a}$$

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

The mean-square error bounds for estimation expectation of a Lipschitz functional is also given by [21].

Proposition 12 (Special case of [21], Eq (27) and Theorem 15) *Under above set-up, given any Lipschitz test function* h*, let* $\overline{h}_{T_0,T} := \frac{1}{T-T_0}\sum_{t=T_0}^{T-1} h(\psi_t)$*, the following bounds hold true:*

$$\left(\mathbb{E}\left[\overline{h}_{T_0,T}\right] - \mathbb{E}_{\pi_\eta}\left[h(X)\right]\right)^2 \leq \frac{\|h\|_{\mathrm{Lip}}^2}{T-T_0}\sum_{t=T_0}^{T} \mathcal{W}_2^2(\pi_\eta^{(t)}, \pi_\eta) \tag{23b}$$

$$\mathrm{var}\left(\overline{h}_{T_0,T}\right) \leq c\frac{\|h\|_{\mathrm{Lip}}^2}{\eta(T-T_0)} \tag{23c}$$

for a universal constant $c > 0$*.*

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

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

Substituting into the MSE bound in Proposition 12, for any choice of burn-in parameter $\beta \in [0,1)$, we have the bound:

$$\mathbb{E}\left(\overline{\psi}_T - \mathbb{E}_{\pi_\eta}[X]\right)^2 \leq c\left(\eta + \frac{1}{\eta(T-T_0)}\right) \leq c' T^{-\min(\alpha, 1-\alpha)} \tag{24}$$

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

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

Lemma 13 *For the* 1*-dimensional probability distribution* π_η *defined above, we have that*

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

for a universal constant $c > 0$*.*

Combining the bound (24) and Lemma 13, we arrive at the lower bound:

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

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

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

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

Taking the limit, we have:

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

which completes the proof of this theorem.
Proof [Proof of Lemma 13]Denote the normalization constant:

$$Z_\eta := \int e^{-F_\eta(x)} dx$$

Since $x^2 \leq F(x) \leq 2x^2$ for any $x \in \mathbb{R}$, we have the bound $\sqrt{\pi/2} \leq Z_\eta \leq \sqrt{\pi}$ for any choice of $\eta > 0$. By definition, we have the expression:

$$\mathbb{E}_{\pi_\eta}[X] = Z_\eta^{-1} \int_0^{+\infty} x \left(e^{-F_\eta(x)} - e^{-F_\eta(-x)} \right) dx$$

Note that $F(x) \leq F(-x)$ for any $x \geq 0$. So we have that $\mathbb{E}_{\pi_\eta}[X] \geq 0$, and the following bound holds:

$$\mathbb{E}_{\pi_\eta}[X] \geq \frac{1}{\sqrt{\pi}} \int_1^2 \left(e^{-F_\eta(x)} - e^{-F_\eta(-x)} \right) dx$$

Given $x \in [1, 2]$ fixed, we lower bound the difference in the density function as follows:

$$e^{-F_\eta(x)} - e^{-F_\eta(-x)} = e^{-x^2} \left(e^{1/2 \int_0^x \delta_\eta(z) dz} - e^{1/4 \int_0^x \delta_\eta(z) dz} \right) \geq \frac{e^{-4}}{4} \int_0^x \delta_\eta(z) dz$$
$$\geq \frac{e^{-4}}{4} \int_{1/2}^1 \frac{z}{\log\left(e + (z\sqrt{\eta})^{-1}\right)} dz \geq \frac{e^{-4}}{8} \cdot \frac{1}{\log\left(e + \frac{2}{\sqrt{\eta}}\right)}$$

Integrating with $x \in [1, 2]$, we arrive at the lower bound:

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

for universal constant $c > 0$. ■

Appendix C. Proof of the non-asymptotic bounds with sharp pre-factor

In this section, we present the proofs for Theorem 4, Theorem 6 and Theorem 7. These three results provide upper bounds on three different metrics (gradient norm, iterate distance, and function value), with the leading-order term exactly matching the optimal normal limit, and sharp high-order terms. We present the proof of Proposition 5 en route (in §C.2), the higher-moment non-asymptotic convergence rates for the process $(v_t)_{t \geq T_0}$ and $(z_t)_{t \geq T_0}$ that is analogous to Proposition 2.

C.1. Proof of Theorem 4

We first establish the results for the single-loop algorithm, and then use it to prove the results with the re-starting loops.

Throughout the proof, we use the following notations for the risk functions

$$r_v(t) := \left(\mathbb{E} \|v_t\|_2^2 \right)^{1/2} \qquad \text{and} \quad r_\theta(t) := \frac{1}{\mu} \left(\mathbb{E} \|\nabla F(\theta_T)\|_2^2 \right)^{1/2}$$

Clearly, by the strong convexity Assumption 1, we have the bound $\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^2 \leq r_\theta(t)^2$.

We start by observing the following decomposition:

$$\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 = \mathbb{E}\left\|z_{T+1}\right\|_2^2 + \mathbb{E}\left\|v_{T+1}\right\|_2^2 - 2\mathbb{E}\langle z_{T+1},\, v_{T+1}\rangle \tag{25}$$

The following lemma provides sharp bounds on the leading-order term $\mathbb{E}\left\|z_{T+1}\right\|_2^2$.

Lemma 14 *Under above set-up, for $T \geq 2T_0 \log T_0$ and any $G \in \mathbb{R}^{d\times d}$, the following bounds hold true for the process $(z_t)_{t\geq T_0}$:*

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \frac{1}{T}\mathrm{Tr}\left(G\Sigma^* G^\top\right) + c\|G\|_{op}^2 \mathcal{H}_T^{(z)} \tag{26a}$$

where the high order term $\mathcal{H}_T^{(z)}$ is defined as

$$\mathcal{H}_T^{(z)} := c\left(\sqrt{\frac{T_0}{T}} + \frac{T_0^\alpha}{T^\alpha}\right)\frac{\sigma_*^2}{T} + c\frac{T_0^2 \log T}{T^2}\left(1 + \frac{T^{2\alpha-3/2}}{T_0^{2\alpha-3/2}}\right)\left\|\nabla F(\theta_0)\right\|_2^2 \tag{26b}$$

See §D.4 for the proof of this lemma.

Invoking Proposition 2, we have the bound for v_T:

$$\mathbb{E}\left\|v_T\right\|_2^2 \leq c\left(\frac{\sigma_*^2}{\mu\eta_T T^2} + \frac{T_0}{\mu^2 T^3 \eta_T^2}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$

For the stepsize choice $\eta_t = \frac{1}{\mu T_0^{1-\alpha} t^\alpha}$, we have the bound

$$\mathbb{E}\left\|v_T\right\|_2^2 \leq c\frac{T_0^{1-\alpha}}{T^{1-\alpha}} \cdot \frac{\sigma_*^2}{T} + c\frac{T_0^{3-2\alpha}}{T^{3-2\alpha}} \cdot \left\|\nabla F(\theta_0)\right\|_2^2 \tag{27}$$

Combining the bounds (26a) and (27) and substituting into the decomposition (25), we arrive at the following bound by applying Young's inequality:

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_T)\right\|_2^2 &\leq \mathbb{E}\left\|z_{T+1}\right\|_2^2 + \mathbb{E}\left\|v_{T+1}\right\|_2^2 + 2\sqrt{\mathbb{E}\left\|z_{T+1}\right\|_2^2} \cdot \sqrt{\mathbb{E}\left\|v_{T+1}\right\|_2^2} \\
&\leq \left(1 + \left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}}\right) \cdot \mathbb{E}\left\|z_{T+1}\right\|_2^2 + \left(1 + \left(\frac{T}{T_0}\right)^{\frac{1-\alpha}{2}}\right)\mathbb{E}\left\|v_{T+1}\right\|_2^2 \\
&\leq \frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T \cdot \left\|\nabla F(\theta_0)\right\|_2^2
\end{aligned}$$

which proves the first claim (11a).

Now we turn to the proof of multi-loop results. By applying the one-loop result to each short epoch, we have the bound for $b = 1, 2, \cdots, \mathscr{E}$:

$$\begin{aligned}
\mathbb{E}\left\|\nabla F(\theta_0^{(b+1)})\right\|_2^2 &\leq \frac{\sigma_*^2}{T^\flat} + c\left(\frac{T_0}{T^\flat}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T^\flat} + c\left(\frac{T_0}{T^\flat}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T \cdot \mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2 \\
&\overset{(i)}{\leq} \frac{2\sigma_*^2}{T^\flat} + \frac{1}{2}\mathbb{E}\left\|\nabla F(\theta_0^{(b)})\right\|_2^2
\end{aligned}$$

In step (i), we use the fact that $T^\flat \geq 2cT_0 \log T_0$ and that $2 \wedge \frac{5-3\alpha}{2} > 1$ for $\alpha \in (0,1)$.

Solving the recursion, we arrive at the bound:

$$\mathbb{E}\left\|\nabla F\big(\theta_0^{(\mathscr{E}+1)}\big)\right\|_2^2 \leq \frac{4\sigma_*^2}{T_0} + 2^{-\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2$$

Substituting this initial condition into the bound (11a), we obtain the final bound:

$$\mathbb{E}\left\|\nabla F\big(\theta_T^{(\mathscr{E}+1)}\big)\right\|_2^2 \leq \frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}\wedge\alpha}\frac{\sigma_*^2}{T} + c\left(\frac{T_0}{T}\right)^{2\wedge\frac{5-3\alpha}{2}}\log T\cdot\left(\frac{4\sigma_*^2}{T_0} + 2^{-\mathscr{E}}\left\|\nabla F(\theta_0)\right\|_2^2\right)$$

Taking $\mathscr{E} \geq \log_2\left(\frac{T_0\|\nabla F(\theta_0)\|_2^2}{4\sigma_*^2} \vee 1\right)$ and substituting with $T = n - \mathscr{E}T^\flat$, we arrive at the conclusion:

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

which proves the bound (11b).

C.2. Proof of Proposition 5

Throughout the proof, we frequently use the following inequalities for the moments of stochastic gradients, which holds true for any $\theta \in \mathbb{R}^d$:

$$\mathbb{E}\left\|\nabla f(\theta,\xi_t)\right\|_2^4 \leq 27\widetilde{\sigma_*}^4 + 27\left(1 + \frac{\widetilde{\ell_\Xi}^4}{\mu^4}\right)\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 \tag{28}$$

To see why this is true, we note that:

$$\begin{aligned}\mathbb{E}\left\|\nabla f(\theta,\xi_t)\right\|_2^4 &\leq 27\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 + 27\mathbb{E}\left\|\varepsilon_t(\theta^*)\right\|_2^4 + 27\mathbb{E}\left\|\varepsilon_t(\theta) - \varepsilon_t(\theta^*)\right\|_2^4\\ &\leq 27\widetilde{\sigma_*}^4 + 27\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4 + 27\widetilde{\ell_\Xi}^4\mathbb{E}\left\|\theta-\theta^*\right\|_2^4\\ &\leq 27\widetilde{\sigma_*}^4 + 27\left(1 + \frac{\widetilde{\ell_\Xi}^4}{\mu^4}\right)\mathbb{E}\left\|\nabla F(\theta)\right\|_2^4\end{aligned}$$

Now we turn to the proof of this proposition. Similar to the proof of Proposition 2, we need the following technical lemmas:

Lemma 15 *Under Assumption 1, 2′, 3, there exists universal constants $c, c' > 0$, when $\eta_t \leq c\big(\frac{1}{L} \wedge \frac{\mu}{\widetilde{\ell_\Xi}^2}\big)$, we have the bound*

$$\sqrt{\mathbb{E}\left\|v_t\right\|_2^4} \leq \left(1-\frac{1}{t}\right)^2\left(1 - \frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{c'}{t^2}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}\right)$$

Lemma 16 *Under Assumption 2′, we have the bound*

$$\sqrt{\mathbb{E}\left\|z_t\right\|_2^4} \leq \frac{cT_0^2\left\|z_0\right\|_2^2}{t^2} + \frac{c\widetilde{\sigma_*}^2}{t} + \frac{c\widetilde{\ell_\Xi}^2}{\mu^2 t^2}\sum_{s=T_0}^{t-1}\sqrt{\mathbb{E}\left\|\nabla F(\theta_s)\right\|_2^4} + \frac{c\widetilde{\ell_\Xi}^2}{t^2}\sum_{s=T_0}^{t-1}s^2\eta_s^2\sqrt{\mathbb{E}\left\|v_s\right\|_2^4}$$

See Section D.5 and D.6 for the proofs of the two lemmas. Taking these two lemmas as given, we now proceed with the proof of the proposition.

The rest of proof goes in parallel with the proof of Proposition 2. We first note that:

$$\sqrt{\mathbb{E}\left\|\nabla F(\theta_t)\right\|_2^4} \leq 4\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4} + 4\sqrt{\mathbb{E}\left\|\nabla F(\theta_t) - \nabla F(\theta_{t-1})\right\|_2^4}$$
$$\leq 4\sqrt{\mathbb{E}\left\|z_t\right\|_2^2} + 4\sqrt{\mathbb{E}\left\|v_t\right\|_2^2} + 4(\eta_t L)^4\sqrt{\mathbb{E}\left\|v_t\right\|_2^4} \leq 4\sqrt{\mathbb{E}\left\|z_t\right\|_2^4} + 6\sqrt{\mathbb{E}\left\|v_t\right\|_2^4}$$

Substituting into the bounds in Lemma 15 and 16, and defining the quantities $H_T := \sup_{T_0 \leq t \leq T} t\sqrt{\mathbb{E}\left\|z_t\right\|_2^4}$, $W_T := T^2\sqrt{\mathbb{E}\left\|v_T\right\|_2^4}$, we arrive at the following recursive inequalities:

$$H_T \leq 2H_{T_0} + 4\widetilde{\sigma_*}^2 + 2C'\widetilde{\ell_\Xi}^2 \sup_{T_0 \leq t \leq T} \frac{1}{t}\sum_{s=T_0}^{t-1} \eta_{s-1}^2 W_s \tag{29a}$$

$$W_T \leq (1 - c\eta_{T-1}\mu)W_{T-1} + \frac{C}{T\mu\eta_{T-1}}H_{T-1} + 2\widetilde{\sigma_*}^2 \tag{29b}$$

Invoking Lemma 10 by replacing (ℓ_Ξ, σ_*) with $(\widetilde{\ell_\Xi}, \widetilde{\sigma_*})$, we obtain the following bounds:

$$H_T \leq c\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \eta_{T_0}}{\mu}W_{T_0} + H_{T_0}\right) \qquad \text{and}$$
$$W_T \leq \frac{c}{\eta_T\mu}\sigma_*^2 + c\left(\frac{T_0}{T\mu^2\eta_{T-1}^2} + e^{-\mu\sum_{t=T_0+1}^{T}\eta_t}T_0^2\right)W_{T_0}$$

For the initial conditions, by applying Khintchine's inequality as well as Young's inequality, we note that:

$$\mathbb{E}\left\|z_{T_0}\right\|_2^4 = \frac{1}{T_0^4}\mathbb{E}\left\|\sum_{t=1}^{T_0}\varepsilon_t(\theta_0)\right\|_2^4 \leq \frac{1}{T_0^4}\mathbb{E}\left(\sum_{t=1}^{T_0}\left\|\varepsilon_t(\theta_0)\right\|_2^2\right)^2 \leq 8\frac{1}{T_0^2}\left(\widetilde{\sigma_*}^4 + \widetilde{\ell_\Xi}^4\left\|\theta_0 - \theta^*\right\|_2^4\right)$$
$$\mathbb{E}\left\|v_{T_0}\right\|_2^4 \leq 8\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4 + \mathbb{E}\left\|z_{T_0}\right\|_2^4 \leq 8\left\|\nabla F(\theta_0)\right\|_2^4 + 8\frac{1}{T_0^2}\left(\widetilde{\sigma_*}^4 + \widetilde{\ell_\Xi}^4\left\|\theta_0 - \theta^*\right\|_2^4\right)$$

Following exactly the same arguments as in the proof of Proposition 2, we arrive at the desired bounds.

C.3. Proof of Theorem 6

We define the quantities $r_\theta(t)$ and $r_v(t)$ the same as in the proof of Theorem 4. Furthermore, we denote the following quantities:

$$\widetilde{r}_v(t) := \left(\mathbb{E}\left\|v_t\right\|_2^4\right)^{1/4} \qquad \text{and} \quad \widetilde{r}_\theta(t) := \frac{1}{\mu}\left(\mathbb{E}\left\|\nabla F(\theta_t)\right\|_2^4\right)^{1/4}$$

Clearly, by the strong convexity Assumption 1, we have the bound $\mathbb{E}\left\|\theta_T - \theta^*\right\|_2^4 \leq r_\theta(t)^4$.

We also note the following decomposition of the gradient:

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

which leads to the following bound under Assumption 4′:

$$\begin{aligned}\big\|(H^*)^{-1}\nabla F(\theta_T) - (\theta_T - \theta^*)\big\|_2 &\leq \int_0^1 \big\|(H^*)^{-1}\left(\nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big) - H^*\right)(\theta_T - \theta^*)\big\|_2 \, d\gamma \\ &\leq \frac{L_2}{\lambda_{\min}(H^*)}\|\theta_T - \theta^*\|_2^2 \leq \frac{L_2}{\lambda_{\min}(H^*)\mu^2}\|\nabla F(\theta_T)\|_2^2 \qquad (30)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\|\theta_T - \theta^*\|_2^2 &\leq \mathbb{E}\left(\big\|(H^*)^{-1}\nabla F(\theta_T)\big\|_2 + \frac{L_2}{\mu^2\lambda_{\min}(H^*)}\|\nabla F(\theta_T)\|_2^2\right)^2 \\ &\leq \mathbb{E}\big\|(H^*)^{-1}\big(v_{T+1} - z_{T+1}\big)\big\|_2^2 + 2\frac{L_2}{\lambda_{\min}(H^*)}\widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)^2}\widetilde{r}_\theta^4(T) \qquad (31)\end{aligned}$$

The leading-order term in the bound (31) admits the following decomposition:

$$\mathbb{E}\big\|(H^*)^{-1}\big(z_{T+1} - v_{T+1}\big)\big\|_2^2 = \mathbb{E}\big\|(H^*)^{-1}z_{T+1}\big\|_2^2 + \mathbb{E}\big\|(H^*)^{-1}v_{T+1}\big\|_2^2 - 2\mathbb{E}\left[\langle (H^*)^{-1}z_T,\, (H^*)^{-1}.v_T\rangle\right]$$

In the following, we bound the three terms in above equation, respectively. Invoking Lemma 14 with $G = (H^*)^{-1}$, we have the bound:

$$\mathbb{E}\big\|(H^*)^{-1}z_{T+1}\big\|_2^2 \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{\alpha\wedge\frac{1}{2}} + \frac{c\|\nabla F(\theta_0)\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \qquad (32a)$$

For the process v_t, Proposition 2 yields the following upper bound:

$$\mathbb{E}\big\|(H^*)^{-1}v_{T+1}\big\|_2^2 \leq \frac{1}{\lambda_{\min}(H^*)^2}\mathbb{E}\|v_{T+1}\|_2^2 \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{3-2\alpha} \qquad (32b)$$

The bound for the cross term is given by the following lemma:

Lemma 17 *Under above set-up, for $T \geq cT_0\log T_0$, for any $d\times d$ deterministic matrix G, the following bound holds true:*

$$\begin{aligned}|\mathbb{E}\left[\langle Gz_t,\, Gv_t\rangle\right]| &\leq c\|G\|_{op}^2\left(\frac{T_0}{t}\right)^{1-\alpha}\left(\frac{\sigma_*^2}{t} + \left(\frac{T_0}{t}\right)^{2-\alpha}\|\nabla F(\theta_0)\|_2^2\right)\log t \\ &\quad + c\frac{\|G\|_{op}^2 L_2}{\mu^2}\left(\frac{T_0}{t}\right)^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right)\end{aligned}$$

See §D.7 for the proof of this lemma.

Substituting with $G = (H^*)^{-1}$, we obtain the bound for the cross term:

$$\left|\mathbb{E}\left[\langle (H^*)^{-1} z_t,\, (H^*)^{-1} v_t\rangle\right]\right| \le \frac{c\sigma_*^2 \log T}{\lambda_{\min}(H^*)^2 T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2 \log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{3-2\alpha} + \frac{cL_2\widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)^2\mu^2 T^{3/2}}\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\left\|\nabla F(\theta_0)\right\|_2^3 \log^2 T}{\lambda_{\min}(H^*)^2\mu^2}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha} \tag{32c}$$

For the rest two terms in the expression (31), we invoke Proposition 5, and obtain the rate:

$$\widetilde{r}_\theta(T) \le \frac{c\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{c\sqrt{\log T}}{\mu}\left\|\nabla F(\theta_0)\right\|_2\left(\frac{T_0}{T}\right)^{1\wedge 3/2-\alpha} \tag{32d}$$

Combining the bounds (32a)-(32d) and substituting into the decomposition (31), we arrive at the bound

$$\begin{aligned}\mathbb{E}\left\|\theta_T-\theta^*\right\|_2^2 \le{}& \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2\log T}{\lambda_{\min}(H^*)^2T}\left(\frac{T_0}{T}\right)^{\alpha\wedge 1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2\log T}{\lambda_{\min}(H^*)^2}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha}\\ &+ \frac{cL_2\widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)\mu^3T^{3/2}} + \frac{cL_2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^3\log^2 T}{\lambda_{\min}(H^*)\mu^3}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha\wedge 3}\\ &+ \frac{cL_2^2\widetilde{\sigma_*}^4}{\lambda_{\min}(H^*)^2\mu^4T^2} + \frac{cL_2^2\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\log^2 T}{\lambda_{\min}(H^*)^2\mu^4}\left(\frac{T_0}{T}\right)^{6-4\alpha\wedge 4}\end{aligned}$$

Noting that $\frac{7}{2}-2\alpha\wedge 3 \ge \frac{3}{2}$ and $6-4\alpha\wedge 4\ge 4$, we complete the proof of the bound (15a).

Now we turn to the proof of the multi-loop result (15b). Invoking Proposition 2 and 5 and noting that $\left\|\nabla F(\theta_t)\right\|_2 \le \left\|z_{t+1}\right\|_2 + \left\|v_{t+1}\right\|_2$, we obtain the bound for $T^\flat \ge cT_0\log T_0$:

$$\mathbb{E}\left\|\nabla F\big(\theta_0^{(b+1)}\big)\right\|_2^2 \le \frac{1}{2}\mathbb{E}\left\|\nabla F\big(\theta_0^{(b)}\big)\right\|_2^2 + \frac{c\sigma_*^2}{T^\flat} \quad \text{and}$$
$$\sqrt{\mathbb{E}\left\|\nabla F\big(\theta_0^{(b+1)}\big)\right\|_2^4} \le \frac{1}{2}\sqrt{\mathbb{E}\left\|\nabla F\big(\theta_0^{(b)}\big)\right\|_2^4} + \frac{c\widetilde{\sigma_*}^2}{T^\flat}$$

Solving the recursion, we have that:

$$\mathbb{E}\left\|\nabla F\big(\theta_0^{(\mathscr{E}+1)}\big)\right\|_2^2 \le 2^{-\mathscr{E}}\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^2 + \frac{2c\sigma_*^2}{T^\flat} \quad \text{and}$$
$$\sqrt{\mathbb{E}\left\|\nabla F\big(\theta_0^{(\mathscr{E}+1)}\big)\right\|_2^4} \le 2^{-\mathscr{E}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4} + \frac{2c\widetilde{\sigma_*}^2}{T^\flat}$$

Taking $\mathscr{E} \ge \log_2\left(\frac{T_0\left\|\nabla F(\theta_0)\right\|_2^2}{4\sigma_*^2}\vee 1\right)$ and substituting into the bound (15a), we have the following guarantee for the multi-loop estimator:

$$\mathbb{E}\left\|\widehat{\theta}_n-\theta^*\right\|_2^2 \le \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*(H^*)^{-1}\big)}{n} + \frac{c\sigma_*^2\log^2 n}{\lambda_{\min}(H^*)^2 n}\left(\frac{T_0}{n}\right)^{\alpha\wedge 1-\alpha} + \frac{cL_2\widetilde{\sigma_*}^3\log^2 n}{\lambda_{\min}(H^*)\mu^3 n^{3/2}} + \frac{cL_2^2\widetilde{\sigma_*}^4\log^2 n}{\lambda_{\min}(H^*)^2\mu^4 n^2}$$

which completes the proof.

C.4. Proof of Theorem 7

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

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

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

$$\begin{aligned} F(\theta) &\leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + \|\theta - \theta^*\|_2 \cdot \int_0^1 \|\nabla^2 F\big(\gamma\theta + (1-\gamma)\theta^*\big) - H^*\|_{\text{op}} d\gamma \cdot \|\theta - \theta^*\|_2 \\ &\leq F(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + L_2 \|\theta - \theta^*\|_2^3 \end{aligned} \tag{33}$$

Similar to Eq (30), we have the bound:

$$\begin{aligned} \left\| (H^*)^{1/2}(\theta - \theta^*) - (H^*)^{-1/2}\nabla F(\theta) \right\|_2 &\leq \int_0^1 \left\| (H^*)^{-1/2} \left(\nabla^2 F\big(\gamma\theta^* + (1-\gamma)\theta_T\big) - H^* \right) (\theta_T - \theta^*) \right\|_2 d\gamma \\ &\leq \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}} \|\theta_T - \theta^*\|_2^2 \leq \frac{L_2}{\sqrt{\lambda_{\min}(H^*)}\mu^2} \|\nabla F(\theta_T)\|_2^2 \end{aligned}$$

Denote the residual $q_t := (H^*)^{1/2}(\theta_t - \theta^*) - (H^*)^{-1/2}\nabla F(\theta_t)$. Substituting into the bound (33), we have that:

$$\begin{aligned} \mathbb{E}\left[F(\theta_T)\right] - F(\theta^*) &\leq \frac{1}{2}\mathbb{E}\left\| (H^*)^{-1/2}\nabla F(\theta) + q_T \right\|_2^2 + L_2 \mathbb{E}\|\theta_T - \theta^*\|_2^3 \\ &\leq \frac{1}{2}\mathbb{E}\left\| (H^*)^{-1/2}(z_{T+1} + v_{T+1}) \right\|_2^2 + 2L_2 \widetilde{r}_\theta^3(T) + \mathbb{E}\|q_T\|_2^2 \\ &\leq \frac{1}{2}\mathbb{E}\left\| (H^*)^{-1/2}(z_{T+1} + v_{T+1}) \right\|_2^2 + 2L_2 \widetilde{r}_\theta^3(T) + \frac{L_2^2}{\lambda_{\min}(H^*)} \widetilde{r}_\theta^4(T) \end{aligned}$$

Invoking Proposition 2, Lemma 14 and 18 with $G = (H^*)^{-1/2}$, we have the bounds

$$\begin{aligned} \mathbb{E}\left\| (H^*)^{-1/2} z_{T+1} \right\|_2^2 &\leq \frac{\mathrm{Tr}\big(\Sigma^*(H^*)^{-1}\big)}{T} + \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{\alpha \wedge \frac{1}{2}} + \frac{c\|\nabla F(\theta_0)\|_2^2 \log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{2\wedge\frac{7}{2}-2\alpha} \\ \mathbb{E}\left\| (H^*)^{-1/2} v_{T+1} \right\|_2^2 &\leq \frac{\mathbb{E}\|v_{T+1}\|_2^2}{\lambda_{\min}(H^*)} \leq \frac{c\sigma_*^2}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{3-2\alpha} \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}\left[\langle (H^*)^{-1/2} z_t,\, (H^*)^{-1/2} v_t \rangle \right] &\leq \frac{c\sigma_*^2 \log T}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{1-\alpha} + \frac{c\|\nabla F(\theta_0)\|_2^2 \log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{3-2\alpha} \\ &\quad + \frac{cL_2 \widetilde{\sigma_*}^3}{\lambda_{\min}(H^*)\mu^2 T^{3/2}}\left(\frac{T_0}{T}\right)^{\frac{1-\alpha}{2}} + \frac{cL_2\|\nabla F(\theta_0)\|_2^3 \log^2 T}{\lambda_{\min}(H^*)\mu^2}\left(\frac{T_0}{T}\right)^{\frac{7}{2}-2\alpha} \end{aligned} \tag{34}$$

Putting them together, we arrive at the bound

$$\mathbb{E}\left[F(\theta_T) - F(\theta^*)\right] \leq \frac{\mathrm{Tr}\big((H^*)^{-1}\Sigma^*\big)}{2T} + \frac{c\sigma_*^2 \log T}{\lambda_{\min}(H^*)T}\left(\frac{T_0}{T}\right)^{\alpha \wedge 1-\alpha} + \frac{c\left\|\nabla F(\theta_0)\right\|_2^2 \log T}{\lambda_{\min}(H^*)}\left(\frac{T_0}{T}\right)^{2\wedge \frac{7}{2}-2\alpha} + cL_2\widetilde{r}_T^3 + c\frac{L_2^2}{\mu}\widetilde{r}_T^4$$

for the quantity $\widetilde{r}_T := \frac{\widetilde{\sigma_*}}{\mu\sqrt{T}} + \frac{\log T}{\mu}\sqrt{\frac{T_0}{T}} \cdot \big(\mathbb{E}\left\|\nabla F(\theta_0)\right\|_2^4\big)^{1/4}$.

For the multi-loop algorithm, applying the same argument on the initial gradient norm as in the proof of Theorem 6, we arrive at the desired bound.

Appendix D. Proof of auxiliary lemmas

For the proofs of auxiliary lemmas, we first describe a simple decomposition result for the process $(z_t)_{t\geq T_0}$ which plays a central role in our analysis.

A key decomposition result The proof for all the results about ROOT-SGD relies on a decomposition of the difference $z_t := v_t - \nabla F(\theta_{t-1})$ that exposes the underlying martingale structure. In particular, beginning with the definition (5) of the updates, for any iterate $t \geq T_0$, we have

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

Unwinding this relation recursively yields

$$z_t = \underbrace{\frac{1}{t}\sum_{s=T_0}^{t}\varepsilon_s(\theta_{s-1})}_{:=M_t} + \frac{T_0}{t} z_{T_0} + \underbrace{\frac{1}{t}\sum_{s=T_0}^{t}(s-1)(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}))}_{:=\Psi_t} \tag{35}$$

It can be seen that both both of the sequences $\{M_t\}_{t\geq T_0}$ and $\{\Psi_t\}_{t\geq T_0}$ are martingales adapted to the filtration $(\mathcal{F}_t)_{t\geq T_0}$. We make use of this martingale decomposition throughout our analysis.

D.1. Proof of Lemma 8

By definition, we note that:

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

Taking the second moments for both sides, we have:

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

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

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

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:

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

Now we study the second term, note that

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

For the cross term, we note that:

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

For the term T_1, we note that:

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

$$\leq (1+\eta_{t-1}L)\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2$$

For the term T_2, we have:

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

So we have:

$$\begin{aligned}
I_3 &\leq 3\mathbb{E}\left(\left\|v_{t-1}\right\|_2 \cdot \left\|\nabla F(\theta_{t-1})\right\|_2\right) \leq 3\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^2 \cdot \mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2}\\
&\leq \frac{t\eta_{t-1}\mu}{8}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{18}{t\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2
\end{aligned}$$

Putting above estimates together, we obtain:

$$\begin{aligned}
\mathbb{E}\left\|v_t\right\|_2^2 &\leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta_{t-1}\mu}{2}\right)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{1}{t^2}\left(2\sigma_*^2 + 4\big(1+\frac{\ell_\Xi^2}{\mu^2}\big)\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2\right)\\
&\quad + \frac{(t-1)\eta_{t-1}\mu}{8t}\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{18}{t^2\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2\\
&\leq \left(1-\frac{1}{t}\right)^2\left(1-\frac{\eta_{t-1}\mu}{4}\right)\mathbb{E}\left\|v_{t-1}\right\|_2^2 + \frac{26}{t^2\mu\eta_{t-1}}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^2 + \frac{2\sigma_*^2}{t^2}
\end{aligned}$$

which finishes the proof.

D.2. Proof of Lemma 9

Taking the squared norm of z_t in the martingale decomposition (35) and applying the triangle inequality yields

$$\mathbb{E}\left\|z_t\right\|_2^2 \leq \frac{2}{t^2}\mathbb{E}\left\|M_t\right\|_2^2 + \frac{T_0^2}{t^2}\left\|z_0\right\|_2^2 + \frac{2}{t^2}\mathbb{E}\left\|\Psi_t\right\|_2^2$$

For the martingale M_t, we have:

$$\mathbb{E}\left\|M_t\right\|_2^2 = \sum_{s=1}^t \mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^2 \leq 2t\sigma_*^2 + 2\ell_\Xi^2\sum_{s=1}^t \mathbb{E}\left\|\theta_{s-1}-\theta^*\right\|_2^2$$

For the martingale Ψ_t, we have:

$$\mathbb{E}\left\|\Psi_t\right\|_2^2 = \sum_{s=1}^t (s-1)^2\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\right\|_2^2 \leq \ell_\Xi^2\sum_{s=1}^t (s-1)^2\eta_{s-1}^2\mathbb{E}\left\|v_{s-1}\right\|_2^2$$

Combining the pieces yields

$$\mathbb{E}\left\|z_t\right\|_2^2 \leq \frac{T_0^2\left\|z_0\right\|_2^2}{t^2} + \frac{4\sigma_*^2}{t} + \frac{4\ell_\Xi^2}{t^2}\sum_{s=1}^t \mathbb{E}\left\|\theta_{s-1}-\theta^*\right\|_2^2 + \frac{\ell_\Xi^2}{t^2}\sum_{s=1}^t (s-1)^2\eta_{s-1}^2\mathbb{E}\left\|v_{s-1}\right\|_2^2$$

Note that the μ-strong convexity condition (cf. Assumption 1) ensures that $\left\|\theta_{s-1}-\theta^*\right\|_2 \leq \frac{1}{\mu}\left\|\nabla F(\theta_{t-1})\right\|_2$. Plugging this bound into the inequality above completes the proof.

D.3. Proof of Lemma 10

Denote $\ell_t := \sum_{s=T_0}^{t} \eta_s$, which is the aggregated step sizes up to time t.

Recursively applying the inequality (19b), and noting that H_t is a non-decreasing sequence and that η_t is non-increasing, we obtain:

$$\begin{aligned} W_T &\leq 2\sigma_*^2 \sum_{t=T_0}^{T-1} e^{-\mu(\ell_T - \ell_t)} + 2CH_{T-1} \sum_{t=T_0}^{T-1} \frac{e^{-\mu(\ell_T-\ell_t)}}{t\mu\eta_{t-1}} + \sum_{t=T_0}^{T-1} e^{-\mu(\ell_T - \ell_{T_0})} W_{T_0} \\ &\leq \frac{2\sigma_*^2}{\eta_T \mu} + \frac{CH_{T-1}}{T(\mu\eta_{T-1})^2} + e^{-\mu(\ell_T-\ell_{T_0})} T_0^2 \mathbb{E}\left\| v_{T_0} \right\|_2^2 \end{aligned}$$

Substituting the bound into Eq (19a), we obtain:

$$\begin{aligned} H_T \leq 4\sigma_*^2 + 2\mathbb{E}\left\| z_{T_0} \right\|_2^2 T_0 + C'\ell_\Xi^2 \frac{2\sigma_*^2}{\mu} \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \eta_s + 2CC'\ell_\Xi^2 H_T \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \frac{1}{s\mu^2} \\ + C'\ell_\Xi^2 T_0^2 \mathbb{E}\left\| v_{T_0} \right\|_2^2 \cdot \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2 \end{aligned}$$

For the quantities involving step size sequences in the inequality above, we have:

$$\begin{aligned} \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \eta_s &\leq \sup_{T_0 \leq t \leq T} \frac{1}{t - T_0 + 1} \sum_{s=T_0}^{t-1} \eta_s \leq \eta_{T_0} \\ \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2 &\leq \frac{1}{T_0} \sum_{s=T_0}^{T-1} e^{-\mu(\ell_s - \ell_{T_0})} \eta_{s-1}^2 \leq \frac{\eta_{T_0}}{T_0 \mu} \end{aligned}$$

For $T_0 > \frac{4CC'\ell_\Xi^2}{\mu^2}$, we have $2CC'\ell_\Xi^2 \sup_{T_0 \leq t \leq T} \frac{1}{t} \sum_{s=T_0}^{t-1} \frac{1}{s\mu^2} \leq \frac{1}{2}$, and consequently:

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

for universal constants $c > 0$.

Substituting back into the bound (19b), for $T \geq T_0 \geq (\mu\eta_T)^{-1}$, we obtain:

$$W_T = T^2 \mathbb{E}\left\| v_T \right\|_2^2 \leq \frac{c'}{\eta_T \mu} \sigma_*^2 + c' \left(\frac{T_0}{T\mu^2 \eta_{T-1}^2} + e^{-\mu(\ell_T - \ell_{T_0})} T_0^2 \right) W_{T_0}$$

D.4. Proof of Lemma 14

By the martingale decomposition (35), for any $t \geq T_0$, we have the identity

$$t^2 \mathbb{E}\left\| G z_t \right\|_2^2 = T_0^2 \mathbb{E}\left\| G z_{T_0} \right\|_2^2 + \mathbb{E}\left([GM]_t \right) + \mathbb{E}\left([G\Psi]_t \right) + 2\mathbb{E}\left([GM, G\Psi]_t \right) \tag{36}$$

For the quadratic variation terms, we note that

$$\mathbb{E}\left([GM]_t \right) = \sum_{s=T_0+1}^{t} \mathbb{E}\left\| G\varepsilon_s(\theta_{s-1}) \right\|_2^2$$

$$
\begin{aligned}
&\leq \sum_{s=T_0+1}^{t} \left(\sqrt{\mathbb{E}\left\| G\varepsilon_s(\theta^*) \right\|_2^2} + \|G\|_{\mathrm{op}} \sqrt{\mathbb{E}\left\| \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*) \right\|_2^2} \right)^2 \\
&\leq \sum_{s=T_0+1}^{t} \left(\sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)} + \ell_\Xi \|G\|_{\mathrm{op}} \sqrt{\mathbb{E}\left\| \theta_{s-1} - \theta^* \right\|_2^2} \right)^2 \\
&\leq (t - T_0)\mathrm{Tr}\left(G\Sigma^* G^\top\right) + 2\sum_{s=T_0+1}^{t} \sqrt{\mathrm{Tr}\left(G\Sigma^* G^\top\right)} \ell_\Xi \|G\|_{\mathrm{op}} r_\theta(s) + \sum_{s=T_0+1}^{t} \ell_\Xi^2 \|G\|_{\mathrm{op}}^2 r_\theta^2(s) \\
&\leq (t - T_0)\mathrm{Tr}\left(G\Sigma^* G^\top\right) + \|G\|_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} \left(2\sigma_* \ell_\Xi r_\theta(s) + \ell_\Xi^2 r_\theta^2(s)\right) \qquad (37)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left([G\Psi]_t\right) &= \sum_{s=T_0+1}^{t} (s-1)^2 \mathbb{E}\left\| G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2}) \right\|_2^2 \\
&\leq \ell_\Xi^2 \|G\|_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)^2 \mathbb{E}\left\| \theta_{s-1} - \theta_{s-2} \right\|_2^2 \\
&\leq \ell_\Xi^2 \|G\|_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)^2 \eta_{s-1}^2 r_v^2(s) \qquad (38)
\end{aligned}
$$

We decompose the cross variation term in two parts, and bound them separately.

$$
\begin{aligned}
\mathbb{E}\left([GM, G\Psi]_t\right) &= \sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}),\, G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle \\
&= \underbrace{\sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta^*),\, G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_1(t)} \\
&\quad + \underbrace{\sum_{s=T_0+1}^{t} (s-1)\mathbb{E}\langle G\varepsilon_s(\theta^*),\, G\varepsilon_s(\theta_{s-1}) - G\varepsilon_s(\theta_{s-2})\rangle}_{:=Q_2(t)}
\end{aligned}
$$

For the term Q_1, Cauchy–Schwartz inequality leads to the bound:

$$
\begin{aligned}
Q_1(t) &\leq \sum_{s=T_0+1}^{t} (s-1)\|G\|_{\mathrm{op}}^2 \sqrt{\mathbb{E}\left\| \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*) \right\|_2^2 \cdot \mathbb{E}\left\| \varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2}) \right\|_2^2} \\
&\leq \ell_\Xi^2 \|G\|_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)\eta_{s-1} \sqrt{\mathbb{E}\left\| \theta_{s-1} - \theta^* \right\|_2^2 \cdot \mathbb{E}\left\| v_{s-1} \right\|_2^2} \\
&\leq \ell_\Xi^2 \|G\|_{\mathrm{op}}^2 \sum_{s=T_0+1}^{t} (s-1)\eta_{s-1} r_v(s) r_\theta(s) \qquad (39)
\end{aligned}
$$

For the term Q_2, we note that

$$\begin{aligned}
Q_2(t) &= \sum_{s=T_0+1}^{t} (s-1) \left(\mathbb{E}\langle G\varepsilon_s(\theta^*),\, G\varepsilon_s(\theta_{s-1})\rangle - \mathbb{E}\langle G\varepsilon_{s-1}(\theta^*),\, G\varepsilon_{s-1}(\theta_{s-2})\rangle \right) \\
&\overset{(i)}{=} (T_0-1)\mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{T_0-1})\rangle + \sum_{s=T_0}^{t-1} \mathbb{E}\langle G\varepsilon_t(\theta^*),\, G\varepsilon_t(\theta_{t-1}) - G\varepsilon_t(\theta_{s-1})\rangle \\
&\overset{(ii)}{\leq} (T_0-1)\sigma_*\ell_\Xi \|G\|_{\mathrm{op}}^2 \sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta_{T_0-1}\right\|_2^2} + \sigma_*\ell_\Xi \|G\|_{\mathrm{op}}^2 \sum_{s=T_0}^{t-1} \sqrt{\mathbb{E}\left\|\theta_{t-1}-\theta_{s-1}\right\|_2^2} \\
&\leq 2\sigma_*\ell_\Xi \|G\|_{\mathrm{op}}^2 \left(T_0 \left\|\theta_0-\theta^*\right\|_2 + t r_\theta(t) + \sum_{s=T_0}^{t-1} r_\theta(s) \right) \qquad (40)
\end{aligned}$$

In step (i), we apply Abel's summation formula, and in step (ii), we use the Cauchy–Schwartz inequality.

Finally, for the initial condition, we have the bound:

$$\mathbb{E}\left\|Gz_{T_0}\right\|_2^2 \leq \|G\|_{\mathrm{op}}^2 \cdot \mathbb{E}\left\|z_{T_0}\right\|_2^2 \leq \|G\|_{\mathrm{op}}^2 \cdot \frac{2(\sigma_*^2 + \ell_\Xi^2 \left\|\theta_0-\theta^*\right\|_2^2)}{T_0} \qquad (41)$$

Collecting the bounds (37)-(41) and substituting into the decomposition (36), we obtain the inequality:

$$\begin{aligned}
\mathbb{E}\left\|Gz_T\right\|_2^2 &\leq \left(1+\frac{T_0}{T}\right) \cdot \frac{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}{T} + c\frac{\|G\|_{\mathrm{op}}^2 \sigma_*\ell_\Xi}{T^2} \sum_{s=T_0}^{T} r_\theta(s) \\
&\quad + c\frac{\|G\|_{\mathrm{op}}^2 \ell_\Xi^2}{T^2} \sum_{s=T_0}^{T} \left(r_\theta(s) + (s-1)\eta_{s-1} r_v(s)\right)^2 + c\frac{T_0\|G\|_{\mathrm{op}}^2 \left(\sigma_* + \ell_\Xi \left\|\theta_0-\theta^*\right\|_2\right)^2}{T^2}
\end{aligned}$$

for a universal constant $c > 0$.

Invoking Proposition 2, we note that:

$$r_\theta(t) \leq c\frac{\sigma_*}{\mu\sqrt{t}} + \frac{\sqrt{T_0}\log t}{\mu^2 t}\left(\ell_\Xi + \frac{1}{\eta_t\sqrt{t}}\right) \left\|\nabla F(\theta_0)\right\|_2 \quad \text{and} \quad r_v(t) \leq c\frac{\sigma_*}{t\sqrt{\mu\eta_t}} + \frac{\sqrt{T_0}}{\mu\eta_t t^{3/2}} \left\|\nabla F(\theta_0)\right\|_2$$

Substituting into above upper bound, we obtain:

$$\begin{aligned}
\mathbb{E}\left\|Gz_T\right\|_2^2 &\leq \frac{\mathrm{Tr}\left(G\Sigma^* G^\top\right)}{T} + c\|G\|_{\mathrm{op}}^2 \left(\frac{\ell_\Xi}{\mu T^{3/2}} + \frac{\ell_\Xi^2 \sum_{s=T_0}^{T} \eta_s}{\mu T^2} + \frac{T_0}{T^2}\right)\sigma_*^2 \\
&\quad + c\|G\|_{\mathrm{op}}^2 \frac{\ell_\Xi^2 T_0 \log T}{\mu^2 T^2} \left(1 + \frac{1}{\mu\ell_\Xi} \sum_{s=T_0}^{T} \frac{1}{\eta_s^2 s^{5/2}}\right) \left\|\nabla F(\theta_0)\right\|_2^2
\end{aligned}$$

For the stepsize choice $\eta_t = \frac{1}{\mu T_0^{1-\alpha} t^\alpha}$, we have the bound

$$\mathbb{E}\left\|Gz_T\right\|_2^2 \leq \frac{\mathrm{Tr}\big(G\Sigma^* G^\top\big)}{T} + c\|G\|_{\mathrm{op}}^2 \left(\frac{T_0}{T}\right)^{1/2\wedge\alpha} \frac{\sigma_*^2}{T} + c\|G\|_{\mathrm{op}}^2 \frac{T_0^2 \log T}{T^2} \left(1 + \frac{T^{2\alpha-3/2}}{T_0^{2\alpha-3/2}}\right) \left\|\nabla F(\theta_0)\right\|_2^2$$

which proves this lemma.

D.5. Proof of Lemma 15

Similar to the proof of Lemma 8, we use the decomposition

$$\mathbb{E}\,\|v_t\|_2^4 \le \left(1-\frac{1}{t}\right)^4 \mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4$$
$$+\frac{4}{t}\left(1-\frac{1}{t}\right)^3 \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \langle v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t),\, \nabla f(\theta_{t-1},\xi_t)\rangle\right)$$
$$+\frac{6}{t^2}\left(1-\frac{1}{t}\right)^2 \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right)$$
$$+\frac{4}{t^3}\left(1-\frac{1}{t}\right) \mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right)$$
$$+\frac{1}{t^4}\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \quad (42)$$

We claim the following bounds on the relevant terms in Eq (42), for stepsize choice $\eta_{t-1} \le \frac{1}{8}\big(\frac{1}{L} \wedge \frac{\mu}{\widetilde{\ell_\Xi}^2}\big)$

$$\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4 \le (1-\mu\eta_{t-1})\mathbb{E}\,\|v_{t-1}\|_2^4 \quad (43a)$$

and

$$\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2 \langle v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t),\, \nabla f(\theta_{t-1},\xi_t)\rangle\right)$$
$$\le \frac{t\mu\eta_{t-1}}{3}\mathbb{E}\,\|v_{t-1}\|_2^4 + \frac{c}{t}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2} \quad (43b)$$

Recall that Eq (28) implies the bound

$$\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4 \le 27\widetilde{\sigma_*}^4 + \frac{27}{(\mu\eta_{t-1})^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4 \quad (43c)$$

Taking these two bounds as given, we now bound the fourth moment $\mathbb{E}\,\|v_t\|_2^4$. First, by Hölder's inequality and Young's inequality, we have the following bounds:

$$\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^2\right)$$
$$\le \left(\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/2}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{1/2}$$
$$\le c\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)$$

and

$$\mathbb{E}\left(\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2^3\right)$$
$$\le \left(\mathbb{E}\,\|v_{t-1}+\nabla f(\theta_{t-1},\xi_t)-\nabla f(\theta_{t-2},\xi_t)\|_2^4\right)^{1/4}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{3/4}$$
$$\le ct\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\cdot\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)+\frac{c}{t}\left(\widetilde{\sigma_*}^4+\frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)$$

Collecting above bounds, we arrive at the conclusion

$$\begin{aligned}\mathbb{E}\left\|v_t\right\|_2^4 &\leq \left(1-\frac{1}{t}\right)^4 (1-\mu\eta_{t-1})\mathbb{E}\left\|v_{t-1}\right\|_2^4 + \frac{c_1}{t^2}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\left\|v_{t-1}\right\|_2^4\right)^{1/2}\\ &\quad + \frac{c_2}{t^4}\left(\widetilde{\sigma_*}^4 + \frac{1}{\mu^2\eta_{t-1}^2}\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4\right)\\ &\leq \left[\left(1-\frac{1}{t}\right)^2\left(1-\frac{\mu\eta_{t-1}}{2}\right)\sqrt{\mathbb{E}\left\|v_{t-1}\right\|_2^4} + \frac{c'}{t^2}\left(\widetilde{\sigma_*}^2 + \frac{1}{\mu\eta_{t-1}}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{t-1})\right\|_2^4}\right)\right]^2\end{aligned}$$

for universal constants $c_1, c_2, c' > 0$. This completes the proof of this lemma.

Proof of Eq (43a): We note the following expansion:

$$\begin{aligned}&\mathbb{E}\left\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^4\\ &\leq \mathbb{E}\left\|v_{t-1}\right\|_2^4 + 4\mathbb{E}\left(\left\|v_{t-1}\right\|_2^2\left\langle v_{t-1},\, \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\rangle\right) + 6\mathbb{E}\left(\left\|v_{t-1}\right\|_2^2\cdot\left\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^2\right)\\ &\quad + 4\mathbb{E}\left(\left\|v_{t-1}\right\|_2\cdot\left\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^3\right) + \mathbb{E}\left\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^4\\ &\leq \left(1 - 4\eta_{t-1}\frac{\mu L}{\mu+L} + 6\eta_{t-1}^2\widetilde{\ell_\Xi}^2\right)\mathbb{E}\left\|v_{t-1}\right\|_2^4 + \left(8 - \frac{4}{(\mu+L)\eta_{t-1}}\right)\mathbb{E}\left(\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}\right\|_2^2\cdot\left\|v_{t-1}\right\|_2^2\right)\\ &\quad + 3\mathbb{E}\left\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^4\end{aligned}$$

For the last term, we note that

$$\begin{aligned}&\mathbb{E}\left\|\nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^4\\ &\leq 8\mathbb{E}\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^4 + 8\mathbb{E}\left\|\varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)\right\|_2^4\\ &\leq 8L^2\eta_{t-1}^2\mathbb{E}\left(\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2})\right\|_2^2\cdot\left\|v_{t-1}\right\|_2^2\right) + 8\widetilde{\ell_\Xi}^4\eta_{t-1}^4\mathbb{E}\left\|v_{t-1}\right\|_2^4\end{aligned}$$

Putting them together, for $\eta_{t-1} \leq \frac{1}{8}\left(\frac{1}{L} \wedge \frac{\mu}{\widetilde{\ell_\Xi}^2}\right)$, we arrive at the contraction bound

$$\begin{aligned}&\mathbb{E}\left\|v_{t-1} + \nabla f(\theta_{t-1},\xi_t) - \nabla f(\theta_{t-2},\xi_t)\right\|_2^4\\ &\leq \left(1 - 4\eta_{t-1}\frac{\mu L}{\mu+L} + 6\eta_{t-1}^2\widetilde{\ell_\Xi}^2 + 24\eta_{t-1}^4\widetilde{\ell_\Xi}^4\right)\mathbb{E}\left\|v_{t-1}\right\|_2^4\\ &\quad + \left(8 - \frac{4}{(L+\mu)\eta_{t-1}} + 24L^2\eta_{t-1}^2\right)\mathbb{E}\left(\left\|\nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}\right\|_2^2\cdot\left\|v_{t-1}\right\|_2^2\right)\\ &\leq (1-\mu\eta_{t-1})\mathbb{E}\left\|v_{t-1}\right\|_2^4\end{aligned}$$

which proves this bound.

Proof of Eq (43b): Denote the following random variables for notational convenience

$$\lambda_{t-1} := v_{t-1} + \nabla F(\theta_{t-1}) - \nabla F(\theta_{t-2}) \qquad \text{and} \qquad \zeta_t := \varepsilon(\theta_{t-1},\xi_t) - \varepsilon(\theta_{t-2},\xi_t)$$

For $\eta_{t-1} \leq \frac{1}{2L}$, it is easy to see the bound $\left\|\lambda_{t-1}\right\|_2 \leq \left\|v_{t-1}\right\|_2$ almost surely. And we note by Assumption 2′ that

$$\mathbb{E}\left(\left\|\zeta_t\right\|_2^4 \mid \mathcal{F}_{t-1}\right) \leq \widetilde{\ell_\Xi}^4\left\|\theta_{t-1} - \theta_{t-2}\right\|_2^4 = \widetilde{\ell_\Xi}^4\eta_{t-1}^4\left\|v_{t-1}\right\|_2^4$$

We note the decomposition

$$\begin{aligned}&\mathbb{E}\left(\|\lambda_{t-1}+\zeta_t\|_2^2\left\langle\lambda_{t-1}+\zeta_t,\,\nabla f(\theta_{t-1},\xi_t)\right\rangle\right)\\&\quad\le\mathbb{E}\left(\|\lambda_{t-1}\|_2^2\left\langle\lambda_{t-1},\,\nabla F(\theta_{t-1})\right\rangle\right)+6\mathbb{E}\left(\|\zeta_t\|_2\cdot\left(\|\lambda_{t-1}\|_2^2+\|\zeta_t\|_2^2\right)\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2\right)\end{aligned}$$

Applying Eq (28) accompanied with Hölder's inequality, we can bound the above terms as follows

$$\mathbb{E}\left(\|\lambda_{t-1}\|_2^2\left\langle\lambda_{t-1},\,\nabla F(\theta_{t-1})\right\rangle\right)\le\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}$$

$$\begin{aligned}\mathbb{E}\left(\|\zeta_t\|_2\,\|\lambda_{t-1}\|_2^2\,\|\nabla f(\theta_{t-1},\xi_t)\|_2\right)&\le 3\widetilde{\ell_\Xi}\eta_{t-1}\mathbb{E}\left(\|v_{t-1}\|_2^3\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\|\nabla F(\theta_{t-1})\|_2\right)\right)\\&\le 3\widetilde{\ell_\Xi}\eta_{t-1}\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(\|\zeta_t\|_2^3\cdot\|\nabla f(\theta_{t-1},\xi_t)\|_2\right)&\le\left(\mathbb{E}\,\|\zeta_t\|_2^4\right)^{3/4}\cdot\left(\mathbb{E}\,\|\nabla f(\theta_{t-1},\xi_t)\|_2^4\right)^{1/4}\\&\le 3\widetilde{\ell_\Xi}^3\eta_{t-1}^3\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\widetilde{\sigma_*}+\frac{\widetilde{\ell_\Xi}}{\mu}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}\right)\end{aligned}$$

Collecting the three terms, and noting that $\eta_{t-1}\le\left(\frac{1}{L}\wedge\frac{\mu}{\widetilde{\ell_\Xi}^2}\right)\le\frac{1}{\widetilde{\ell_\Xi}}\sqrt{\frac{\mu}{L}}\le\frac{1}{\widetilde{\ell_\Xi}}$, we have

$$\begin{aligned}&\mathbb{E}\left(\|\lambda_{t-1}+\zeta_t\|_2^2\left\langle\lambda_{t-1}+\zeta_t,\,\nabla f(\theta_{t-1},\xi_t)\right\rangle\right)\\&\le c\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{3/4}\cdot\left(\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/4}+\widetilde{\ell_\Xi}\eta_{t-1}\widetilde{\sigma_*}\right)\\&\le\frac{t\mu\eta_{t-1}}{3}\mathbb{E}\,\|v_{t-1}\|_2^4+\frac{c}{t}\left(\widetilde{\sigma_*}^2+\frac{1}{\mu\eta_{t-1}}\left(\mathbb{E}\,\|\nabla F(\theta_{t-1})\|_2^4\right)^{1/2}\right)\cdot\left(\mathbb{E}\,\|v_{t-1}\|_2^4\right)^{1/2}\end{aligned}$$

which proves this inequality.

D.6. Proof of Lemma 16

By Eq (35) and Minkowski's inequality, we have the bound

$$\mathbb{E}\,\|z_t\|_2^4\le\frac{T_0^4}{t^4}\mathbb{E}\,\|z_{T_0}\|_2^4+\frac{8}{t^4}\mathbb{E}\,\|M_t\|_2^4+\frac{8}{t^4}\mathbb{E}\,\|\Psi_t\|_2^4$$

Invoking the BDG inequality for Hilbert-space-valued martingales, we have the moment bound

$$\mathbb{E}\,\|M_t\|_2^4\le c\mathbb{E}\left([M]_t^2\right)=c\cdot\mathbb{E}\left(\sum_{s=T_0+1}^{t}\|\varepsilon_s(\theta_{s-1})\|_2^2\right)^2\qquad\text{and}$$

$$\mathbb{E}\left\|\Psi_t\right\|_2^4 \leq c\mathbb{E}\left([\Psi]_t^2\right) \leq c \cdot \mathbb{E}\left(\sum_{s=T_0+1}^{t}(s-1)^2\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\right\|_2^2\right)^2$$

Invoking Cauchy–Schwartz inequality, we note that

$$\begin{aligned}
\mathbb{E}\left\|M_t\right\|_2^4 &\leq c\sum_{s=T_0+1}^{t}\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^4 + 2c\sum_{T_0+1\leq s\leq u\leq t}\mathbb{E}\left(\left\|\varepsilon_s(\theta_{s-1})\right\|_2^2 \cdot \mathbb{E}\left\|\varepsilon_u(\theta_{u-1})\right\|_2^4\right)\\
&\leq c\sum_{s=T_0+1}^{t}\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^4 + 2c\sum_{T_0+1\leq s\leq u\leq t}\sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^4}\cdot\sqrt{\mathbb{E}\left\|\varepsilon_u(\theta_{u-1})\right\|_2^4}\\
&= c\left(\sum_{s=T_0+1}^{t}\sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^4}\right)^2
\end{aligned}$$

Similarly, for the martingale $(\Psi_t)_{t\geq T_0}$, we have the bound

$$\sqrt{\mathbb{E}\left\|\Psi_t\right\|_2^4} \leq c\sum_{s=T_0+1}^{t}(s-1)^2\sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\right\|_2^4}$$

By Eq (28), we have the bound

$$\sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})\right\|_2^4} \leq c\left(\widetilde{\sigma_*}^2 + \frac{\widetilde{\ell_\Xi}^2}{\mu^2}\sqrt{\mathbb{E}\left\|\nabla F(\theta_{s-1})\right\|_2^4}\right)$$

By Assumption 2′, we note that

$$\sqrt{\mathbb{E}\left\|\varepsilon_s(\theta_{s-1})-\varepsilon_s(\theta_{s-2})\right\|_2^4} \leq \widetilde{\ell_\Xi}^2\sqrt{\mathbb{E}\left\|\theta_{s-1}-\theta_{s-2}\right\|_2^4} = \widetilde{\ell_\Xi}^2\eta_{t-1}^2\sqrt{\mathbb{E}\left\|v_{s-1}\right\|_2^4}$$

Collecting the terms above, we arrive at the conclusion.

D.7. Proof of Lemma 17

We first note the following decomposition, which holds true for any $\widetilde{T} \in [0, t-T_0]$

$$|\mathbb{E}\langle tGz_t,\, Gv_t\rangle| \leq \underbrace{(t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}},\, Gv_t\rangle\right|}_{:=Q_3\left(t,\widetilde{T}\right)} + \underbrace{\left|\mathbb{E}\langle G\big(tz_t-(t-\widetilde{T})z_{t-\widetilde{T}}\big),\, Gv_t\rangle\right|}_{:=Q_4\left(t,\widetilde{T}\right)}.$$

We claim the following upper bounds for the terms $Q_3(t,\widetilde{T})$ and $Q_4(t,\widetilde{T})$, for $\widetilde{T} \in \left[cT_0^{1-\alpha}t^\alpha \log t, t/2\right]$:

$$Q_3(t,\widetilde{T}) \leq c\frac{\|G\|_{\text{op}}^2 L_2}{\mu^2}t^{\frac{1+\alpha}{2}}T_0^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2}\log^2 t\left\|\nabla F(\theta_0)\right\|_2^3\right) \tag{44a}$$

$$Q_4(t,\widetilde{T}) \leq c\|G\|_{\text{op}}^2\sqrt{\widetilde{T}T_0^{1-\alpha}t^\alpha}\cdot\left(\frac{\sigma_*^2}{t} + \left(\frac{T_0}{t}\right)^{2-\alpha}\left\|\nabla F(\theta_0)\right\|_2^2\right) \tag{44b}$$

Taking these two bounds as given, we choose the time-lag parameter $\widetilde{T} := cT_0^{1-\alpha}t^{\alpha}\log t$, and arrive at the bound:

$$|\mathbb{E}\langle Gz_t, Gv_t\rangle| \leq c|\!|\!|G|\!|\!|_{\mathrm{op}}^2 \left(\frac{T_0}{t}\right)^{1-\alpha}\left(\frac{\sigma_*^2}{t} + \left(\frac{T_0}{t}\right)^{2-\alpha}\|\nabla F(\theta_0)\|_2^2\right)\log t$$
$$+ c\frac{|\!|\!|G|\!|\!|_{\mathrm{op}}^2 L_2}{\mu^2}\left(\frac{T_0}{t}\right)^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right)$$

which completes the proof of this lemma.

Proof of the bound (44a): To bound the term Q_3, we use the following lemma

Lemma 18 *For $t > T_0$ and $s > 0$, the following bound holds true*

$$\mathbb{E}\,\|\mathbb{E}\,(v_{t+s} \mid \mathcal{F}_t)\|_2^2 \leq c\widetilde{r}_v^2(t)e^{-\mu\sum_{k=1}^{s-1}\eta_k} + c\frac{L_2^2}{\mu^2}\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t)$$

See §D.8 for the proof of this lemma.

Taking Lemma 18 as given, the bound for the term $Q_3(t, \widetilde{T})$ directly follows from Cauchy–Schwartz inequality.

$$Q_3(t,\widetilde{T}) = (t-\widetilde{T})\left|\mathbb{E}\langle Gz_{t-\widetilde{T}},\, G\mathbb{E}\left(v_t \mid \mathcal{F}_{t-\widetilde{T}}\right)\rangle\right| \leq t|\!|\!|G|\!|\!|_{\mathrm{op}}^2\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2}\cdot\sqrt{\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathcal{F}_{t-\widetilde{T}}\right)\right\|_2^2}$$

For the time-lag $\widetilde{T} \leq \frac{t}{2}$, Proposition 2 yields the bound:

$$t\sqrt{\mathbb{E}\left\|z_{t-\widetilde{T}}\right\|_2^2} \leq c\sigma_*\sqrt{t} + T_0\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2 \tag{45a}$$

By Lemma 18, for a non-increasing stepsize sequence, when the time-lag $\widetilde{T}$ satisfies $\mu\widetilde{T}\eta_t \geq c\log t$, we have the bound $e^{-\mu\sum_{k=1}^{s-1}\eta_k} \leq \frac{1}{t^3}$. Therefore, given the stepsize choice $\eta_t = \frac{1}{\mu T_0^{1-\alpha}t^{\alpha}}$, we have the following bound holding true for $\widetilde{T} \geq cT_0^{1-\alpha}t^{\alpha}\log t$:

$$\mathbb{E}\left\|\mathbb{E}\left(v_t \mid \mathcal{F}_{t-\widetilde{T}}\right)\right\|_2^2 \leq cL_2\widetilde{r}_v^2(t)\widetilde{r}_\theta^2(t) \tag{45b}$$

Combining the bounds (45a) and (45b), we have the following bound holds true for the time-lag taking values in the interval $\widetilde{T} \in \left[cT_0^{1-\alpha}t^{\alpha}\log t, t/2\right]$ (the interval is non-empty for any $t \geq cT_0\log T_0$):

$$Q_3(t,\widetilde{T}) \leq c\frac{L_2|\!|\!|G|\!|\!|_{\mathrm{op}}^2}{\mu}\left(\sigma_*\sqrt{t} + T_0\sqrt{\log t}\,\|\nabla F(\theta_0)\|_2\right)\widetilde{r}_v(t)\cdot\widetilde{r}_\theta(t)$$

Noting that $\sigma_* \leq \widetilde{\sigma_*}$ and $\ell_\Xi \leq \widetilde{\ell_\Xi}$, above bounds lead to the inequality:

$$Q_3(t,\widetilde{T}) \leq c\frac{L_2|\!|\!|G|\!|\!|_{\mathrm{op}}^2}{\mu^2}t^{\frac{1+\alpha}{2}}T_0^{\frac{1-\alpha}{2}}\left(\frac{\widetilde{\sigma_*}^3}{t^{3/2}} + \left(\frac{T_0}{t}\right)^{3-3\alpha/2}\log^2 t\,\|\nabla F(\theta_0)\|_2^3\right)$$

which proves the desired result.

Proof of the bound (44b): For the term Q_4, we also apply Cauchy-Schwartz inequality, and obtain the following bound:

$$Q_4(t,\widetilde{T}) \leq \|G\|_{\text{op}}^2 \sqrt{2\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + 2\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2} \cdot \sqrt{\mathbb{E}\|v_t\|_2^2}$$

The mean-squared norms of martingales are just their expected quadratic variation:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([M]_t - [M]_{t-\widetilde{T}}\right) \leq 2\widetilde{T}\sigma_*^2 + 2\sum_{s=t-\widetilde{T}+1}^{t} \ell_\Xi^2 r_\theta^2(s)$$
$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 = \mathbb{E}\left([\Psi]_t - [\Psi]_{t-\widetilde{T}}\right) \leq \ell_\Xi^2 \sum_{s=t-\widetilde{T}+1}^{t} (s-1)^2\eta_{s-1}^2 r_v^2(s)$$

Substituting with the rates in Proposition 2, we have the bounds:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\sigma_*^2 + \frac{\ell_\Xi^2 T_0 \log t}{\mu^4 t^2}(\ell_\Xi^2 + \frac{1}{\eta_t^2 t})\|\nabla F(\theta_0)\|_2^2\right) \quad \text{and} \tag{46a}$$
$$\mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\frac{\ell_\Xi^2\sigma_*^2\eta_t}{\mu} + \frac{T_0^2}{t^3}\|\nabla F(\theta_0)\|_2^2\right) \tag{46b}$$

For the stepsize choice $\eta_t = \frac{1}{\mu T_0^{1-\alpha} t^\alpha}$, we have the bound:

$$\mathbb{E}\left\|M_t - M_{t-\widetilde{T}}\right\|_2^2 + \mathbb{E}\left\|\Psi_t - \Psi_{t-\widetilde{T}}\right\|_2^2 \leq c\widetilde{T}\left(\sigma_*^2 + T_0\left(\frac{T_0}{t}\right)^{\min(2,3-2\alpha)}\|\nabla F(\theta_0)\|_2^2\right)$$

Invoking Proposition 2, we can bound the moment of v_t as:

$$\mathbb{E}\|v_t\|_2^2 \leq c\frac{\sigma_*^2 T_0^{1-\alpha}}{t^{2-\alpha}} + \left(\frac{T_0}{t}\right)^{3-2\alpha}\|\nabla F(\theta_0)\|_2^2$$

Combining above bounds, we conclude that

$$Q_4(t,\widetilde{T}) \leq c\|G\|_{\text{op}}^2\sqrt{\widetilde{T}T_0^{1-\alpha}t^\alpha} \cdot \left(\frac{\sigma_*^2}{t} + (\frac{T_0}{t})^{2-\alpha}\|\nabla F(\theta_0)\|_2^2\right)$$

D.8. Proof of Lemma 18

Given $t > T_0$ fixed, denote $\Delta_s := \mathbb{E}(v_{t+s} \mid \mathcal{F}_t)$ for any $s > 0$.

Taking conditional expectations on both sides of Eq (5a), for $s > 0$, we have that

$$\mathbb{E}[v_{t+s} \mid \mathcal{F}_t] = \frac{t+s-1}{t+s}\mathbb{E}[v_{t+s-1} + \nabla F(\theta_{t+s-1}) - \nabla F(\theta_{t+s-2}) \mid \mathcal{F}_t] + \frac{1}{t+s}\mathbb{E}[\nabla F(\theta_{t+s-1}) \mid \mathcal{F}_t] \tag{47}$$

By the decomposition $\nabla F(\theta_{t+s-1}) = v_{t+s} - z_{t+s}$ and the fact that $(z_t)_{t\geq T_0}$ is a martingale, we note that

$$\mathbb{E}[\nabla F(\theta_{t+s-1}) \mid \mathcal{F}_t] = \mathbb{E}[v_{t+s} \mid \mathcal{F}_t]$$

By the one-point Hessian Lipschitz condition, we note that

$$\begin{aligned}
&\|\nabla F(\theta_{t+s-1}) - \nabla F(\theta_{t+s-2}) + \eta_{t+s-1} H^* v_{t+s-1}\|_2 \\
&= \eta_{t+s-1} \left\| \int_0^1 \left(\nabla^2 F\big(\gamma \theta_{t+s-1} + (1-\gamma)\theta_{t+s-2}\big) - \nabla^2 F(\theta^*) \right) v_{t+s-1} \, d\gamma \right\|_2 \\
&\leq \eta_{t+s-1} L_2 \|v_{t+s-1}\|_2 \cdot \int_0^1 \|\gamma \theta_{t+s-1} + (1-\gamma)\theta_{t+s-2} - \theta^*\|_2 \, d\gamma \\
&\leq \eta_{t+s-1} L_2 \|v_{t+s-1}\|_2 \cdot \left(\|\theta_{t+s-1} - \theta^*\|_2 + \|\theta_{t+s-2} - \theta^*\|_2 \right)
\end{aligned}$$

Substituting into the identity (47), we obtain the following inequality, which holds true almost surely for any $s > 0$:

$$\begin{aligned}
\|\Delta_s\|_2 &\leq \|(I - \eta_{t+s-1} H^*)\Delta_{s-1}\|_2 + \eta_{t+s-1} L_2 \mathbb{E}\left[\|v_{t+s-1}\|_2 \cdot \left(\|\theta_{t+s-1} - \theta^*\|_2 + \|\theta_{t+s-2} - \theta^*\|_2 \right) \mid \mathcal{F}_t \right] \\
&\leq (1 - \eta_{t+s-1}\mu) \|\Delta_{s-1}\|_2 + \eta_{t+s-1} L_2 \mathbb{E}\left[\|v_{t+s-1}\|_2 \cdot \left(\|\theta_{t+s-1} - \theta^*\|_2 + \|\theta_{t+s-2} - \theta^*\|_2 \right) \mid \mathcal{F}_t \right]
\end{aligned}$$

Taking the second moment and applying Cauchy-Schwartz inequality, we arrive at the bound

$$\begin{aligned}
&\sqrt{\mathbb{E}\|\Delta_s\|_2^2} \\
&\leq (1 - \eta_{t+s-1}\mu)\sqrt{\mathbb{E}\|\Delta_{s-1}\|_2^2} + 2\eta_{t+s-1} L_2 \left(\mathbb{E}\|v_{t+s-1}\|_2^4 \cdot \left(\mathbb{E}\|\theta_{t+s-1} - \theta^*\|_2^4 + \mathbb{E}\|\theta_{t+s-2} - \theta^*\|_2^4 \right) \right)^{1/4} \\
&\leq (1 - \eta_{t+s-1}\mu)\sqrt{\mathbb{E}\|\Delta_{s-1}\|_2^2} + 2\eta_{t+s-1} L_2 \widetilde{r}_v(t) \widetilde{r}_\theta(t)
\end{aligned}$$

Solving the recursion, we obtain the bound

$$\mathbb{E}\|\Delta_s\|_2^2 \leq c\widetilde{r}_v^2(t) e^{-\mu \sum_{k=1}^{s-1} \eta_k} + c\frac{L_2^2}{\mu^2} \widetilde{r}_v^2(t) \widetilde{r}_\theta^2(t)$$

which finishes the entire proof.

Appendix E. Discussions

In this paper, we revisited the problem of stochastic optimization for strongly convex and smooth M-estimators, focusing on the ROOT-SGD algorithm with diminishing stepsize. We established both sharp asymptotic and non-asymptotic results, demonstrating that ROOT-SGD converges asymptotically to the optimal normal limit under minimal smoothness conditions that guarantee asymptotic normality.[6] In contrast, we provided a counter-example showing that the Polyak-Ruppert averaging procedure is asymptotically sub-optimal under the same conditions.

On the non-asymptotic side, we derived upper bounds on the gradient norm, estimation error, and excess risk, where the leading term matches the asymptotic risk with near-unity pre-factor, and high-order terms decay exponentially. Additionally, with a one-point Hessian-Lipschitz condition, we established that the additional terms decay at a rate of $O(n^{-3/2})$, achieving optimality without requiring prior knowledge of the sample size.

6. The (Bayesian) Cramér-Rao lower bounds provide the fundamental limit of the *mean-squared error* (MSE) of an estimator in relation to the Fisher information. The standard Cramér-Rao lower bounds are valid only for unbiased estimators, whereas the Bayesian Cramér-Rao lower bound applies to the Bayes risk of *any* estimator [25].

Our findings extend to broader optimization scenarios, suggesting potential applications in non-strongly convex, non-convex, and stochastic approximation problems. Future research could explore these methods in Markovian and distributed data settings, opening new avenues for further development.