Statistical inference with implicit SGD: proximal Robbins-Monro vs. Polyak-Ruppert

Yoonhyung Lee [*1] **Sungdong Lee** [*2] **Joong-Ho Won** [2]

Abstract

The implicit stochastic gradient descent (ISGD), a proximal version of SGD, is gaining interest in the literature due to its stability over (explicit) SGD. In this paper, we conduct an in-depth analysis of the two modes of ISGD for smooth convex functions, namely proximal Robbins-Monro (proxRM) and proximal Poylak-Ruppert (proxPR) procedures, for their use in statistical inference on model parameters. Specifically, we derive non-asymptotic point estimation error bounds of both proxRM and proxPR iterates and their limiting distributions, and propose on-line estimators of their asymptotic covariance matrices that require only a single run of ISGD. The latter estimators are used to construct valid confidence intervals for the model parameters. Our analysis is free of the generalized linear model assumption that has limited the preceding analyses, and employs feasible procedures. Our on-line covariance matrix estimators appear to be the first of this kind in the ISGD literature.

1. Introduction

Consider the optimization problem of the form

$$\min_{\theta} L(\theta) := \mathbf{E}[\ell(Z, \theta)] \tag{1}$$

where $\theta \in \mathbb{R}^p$ is the variable (parameter) of interest, Z is a random variable, and $\mathbf{E}[\cdot]$ denotes the expectation over the distribution of Z. Function ℓ is a real-valued *sample function*, information on which can only be obtained by observing independent copies of Z. For example, if ℓ refers to the negative log-likelihood of the model parameter θ given data Z, then problem (1) reduces to finding the "true parameter" $\theta_\star$ of the model.

[*]Equal contribution [1]Kakao Entertainment Corp. [2]Department of Statistics, Seoul National University. Correspondence to: Joong-Ho Won <wonj@stats.snu.ac.kr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

A popular method for solving problem (1) is the stochastic gradient descent (SGD) method

$$\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_{n-1}), \tag{2}$$

where the gradient is with respect to the second argument of ℓ. The γ_n is the algorithm parameter called either *step size* or *learning rate*. Assuming $\nabla L(\theta) = \mathbf{E}[\nabla \ell(Z, \theta)]$, $\nabla \ell(Z_n, \theta)$ is an unbiased estimator of $\nabla L(\theta)$, and SGD (2) is an instance of stochastic approximation due to Robbins & Monro (1951) for finding the root of ∇L. The past decade witnessed a renewed interest in the Robbins-Monro procedure, mainly due to its adaptivity to large-scale data in machine learning problems (Nemirovski et al., 2009; Bottou, 2010; Bach & Moulines, 2011; Bottou et al., 2018). In particular, SGD and its variants are among the major driving forces of deep learning (LeCun et al., 2015; Abadi et al., 2016). Its convergence property has been extensively studied (Zinkevich, 2003; Nemirovski et al., 2009; Bach & Moulines, 2011). More recently, SGD has been studied as a tool for statistical inference from large datasets (Li et al., 2018; Liang & Su, 2019; Chen et al., 2020).

A notable issue with SGD is its sensitivity to the step size selection. If it is too small, the convergence can be arbitrarily slow; if it is too large, then the iterate $\{\theta_n\}$ can diverge (Bach & Moulines, 2011; Ryu & Boyd, 2014).

As an alternative to SGD, consider the following iteration.

$$\theta_n = \mathrm{prox}_{\gamma_n \ell(Z_n, \cdot)}(\theta_{n-1}), \tag{3}$$

where $\mathrm{prox}_{\gamma f}(\theta) = \mathrm{argmin}_{\theta' \in \mathbb{R}^p} \left\{ f(\theta') + \frac{1}{2\gamma} \|\theta' - \theta\|^2 \right\}$ is the *proximity operator* of f. Here the norm $\|\cdot\|$ is the Euclidean (ℓ_2) norm. Using the optimality condition of the minimand in the definition of the operator, iteration (3) can be written as an implicit equation

$$\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_n). \tag{3$'$}$$

Iteration (3) can be considered as a noisy version of the proximal point algorithm due to Rockafellar (1976), introduced to alleviate the sensitivity to the step size in the (noiseless) gradient descent method. Being relatively new in the literature, iteration (3) has been called in various names:

incremental proximal method (Bertsekas, 2011), stochastic proximal point method (Ryu & Boyd, 2014; Bianchi, 2016), and implicit SGD (ISGD, Toulis et al., 2014; Toulis & Airoldi, 2017a).

Convergence of ISGD is studied for the finite population case (Bertsekas, 2011). For infinite population, its stability (that the iterates do not diverge) and asymptotic rate of convergence are studied in Toulis et al. (2014); Ryu & Boyd (2014). Non-asymptotic estimation error bounds are studied in Toulis & Airoldi (2017a); Patrascu & Necoara (2017); Asi & Duchi (2019). These results can be summarized as that ISGD enjoys the same asymptotic rate of convergence as (explicit) SGD, while the former is much more stable and less sensitive to the choice of the step size.

Asymptotic normality of the ISGD is shown in Toulis & Airoldi (2017a) under several assumptions. However, the result of Toulis & Airoldi (2017a) is limited by the assumption that the sample function ℓ takes the form of $\ell(Z,\theta) = g(X^T\theta, Y)$ where $Z = (X,Y)$, which we shall call the *generalized linear model (GLM) assumption*. While the GLM model family is one of the most widely used in practice and the GLM assumption simplifies the computation of the proximity operator, it may not fit when dealing with the general problem (1). For example, consider estimating the α quantile of a univariate distribution with CDF $F(\theta)$. Then $\ell(Z,\theta) = \max(0, \theta - Z) - \alpha(\theta - Z)$ where Z is drawn from F. This function does not satisfy the GLM assumption. In addition, it is assumed that the Fisher information matrix coincides with the Hessian of the objective function L, which is not in general true unless ℓ is the negative log-likelihood of P. Furthermore, Toulis & Airoldi (2017a) assume that L is both globally Lipschitz and globally strongly convex, a contradiction (Bach & Moulines, 2011; Asi & Duchi, 2019).

These restriction and contradiction are relaxed in some sense by Toulis et al. (2021), who consider the *proximal Robbins-Monro procedure* that idealizes ISGD:

$$\begin{aligned} \theta_n^+ &= \mathrm{prox}_{\gamma_n L}(\theta_{n-1}), \\ \theta_n &= \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_n^+), \end{aligned} \tag{4}$$

and show asymptotic normality of θ_n under the assumption that L is locally strongly convex at $\theta_\star$ and the *gradient* of L is globally Lipschitz (Toulis et al., 2021, Theorem 2.4). Nevertheless, procedure (4) is *not feasible* since θ_n^+ cannot be computed (L is unknown or difficult to compute). ISGD (3) can be understood as a plug-in procedure mimicking (4), since θ_n in (3) is an unbiased estimator of θ_n^+. In the sequel, we shall call the ISGD iteration (3) simply proximal Robbins-Monro (proxRM) and distinguish it from (4) by calling the latter the *idealized* proxRM. Unlike the idealized counterpart, the consistency (convergence) and limiting distribution of proxRM has not been studied well, except for the special case mentioned above.

In the explicit SGD literature, averaging the iterates, known as the Polyak-Ruppert averaging (Polyak & Juditsky, 1992; Ruppert, 1988) has been studied as a means to achieve the optimal asymptotic rate and adapt to large step sizes (Bach & Moulines, 2011). Averaging the ISGD iterates (3), i.e., taking $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$, can thus be called the *proximal Polyak-Ruppert* (proxPR) procedure. Asymptotic normality of proxPR is shown by Asi & Duchi (2019). However, *non-asymptotic* (point) estimation error bounds, which capture the transient behavior of $\{\bar{\theta}_n\}$, has not been studied well, although was conjectured to improve the rate (Toulis et al., 2021).

The goal of this paper is to fill these gaps in the literature, as well as to propose a device for efficient statistical inference with ISGD, in line with the similar works in explicit SGD (Li et al., 2018; Liang & Su, 2019; Chen et al., 2020). The latter goal is important since it enables to construct a valid confidence interval for the true parameter. Building the interval *on-line* is also important as it means a single run of ISGD iteration suffices for legitimate inference.

Contributions. Specifically, our contributions are as follows. (i) We extend the work by Toulis & Airoldi (2017a) on proxRM for GLM models to non-GLM settings. The relevant results include a non-asymptotic error bound on model parameter estimation, stability of the procedure, and its asymptotic normality under strong convexity. (ii) We elucidate that the above properties agree with those of the idealized proxRM (Toulis et al., 2021), thereby asserting the feasibility of the procedure. (iii) We derive a non-asymptotic estimation error bound for proxPR, featuring the transient behavior of averaged ISGD. (iv) We propose consistent *on-line* estimators of the asymptotic covariance matrices of both proxRM and proxPR iterates, enabling statistical inference on the model parameter. In addition to these results toward statistical inference, which requires strong convexity for identifiability of the parameter, (v) we provide a non-asymptotic analysis of the two procedures in the absence of strong convexity. In particular, we show that proxPR achieves the minimax optimal rate up to a logarithmic factor, confirming the conjecture by Toulis et al. (2021). Table 1 summarizes the contributions.

2. Preliminary

Let us begin with formally defining the objective function of problem (1).

$$L(\theta) \triangleq \mathbf{E}[\ell(Z,\theta)] = \int_\Omega \ell(z,\theta) dP(z),$$

where z is an element of the probability space $(\Omega, \mathcal{F}, P)$, for which the probability measure P can be considered as

Table 1. Summary of contributions to implicit SGD. Learning rate schedule (R) is assumed.

	GLM free	Feasibility	Non-asymptotic	Asympt. normality	Inference
Toulis & Airoldi (2017a)	✗	✓	✓ (RM)	✓ (RM)	✗
Patrascu & Necoara (2017)	✓	✓	✓ (RM)	✗	✗
Asi & Duchi (2019)	✓	✓	△ (PR)*	✓ (PR)	✗
Toulis et al. (2021)	✓	✗	✓ (RM)	✓ (RM)	✗
This work	✓	✓	✓ (RM, PR)	✓ (RM)	✓ (RM, PR)

* Restricted to a certain class of "easy" problems. RM = proximal Robbins-Monro; PR = proximal Polyak-Ruppert.

the distribution of a random variable $Z : \Omega \to \Omega : z \mapsto z$. In most practical situations Ω is the Euclidean space $\mathbb{R}^m$, $\mathcal{F}$ is the Borel sets of $\mathbb{R}^m$. The sample function $\ell : \Omega \times \mathbb{R}^p \to \mathbb{R}$ is a real-valued. The ISGD procedure is implemented by sampling $Z_1, \ldots, Z_n, \ldots$ from P independently and applying the update equation (3). Finally, the filtration $\mathcal{F}_n$ is the smallest σ-algebra generated by $Z_1, \ldots, Z_n$.

We make the following basic assumptions on the sample function ℓ.

Assumption A1. *Function $\ell(z, \cdot)$ is real-valued convex function in $\mathbb{R}^p$ for each $z \in \Omega$; Function $\ell(\cdot, \theta)$ is integrable for each $\theta \in \mathbb{R}^p$.*

Assumption A2. *Function $\ell(Z, \cdot)$ is $\beta(Z)$-smooth almost surely (a.s.), with $\mathbf{E}[\beta^2(Z)] = \beta_0^2 < \infty$ around a minimizer $\theta_\star$ of $L(\cdot) = \mathbf{E}[\ell(Z, \cdot)]$. That is, a sample function $\ell(Z, \theta)$ is continuously differentiable in θ and $\|\nabla\ell(Z, \theta) - \nabla\ell(Z, \theta_\star)\| \le \beta(Z) \|\theta - \theta_\star\|$ for all θ, with probability one.*

Definition 2.1 (M-convexity (Ryu & Boyd, 2014))**.** *An extended real-valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is called M-convex at $x \in \mathbb{R}^p$ for a symmetric, positive semidefinite matrix $M \in \mathbb{R}^{p \times p}$ (denoted by $M \succeq 0$) if for $s \in \partial f(x)$*

$$f(y) \ge f(x) + s^T(y - x) + \tfrac{1}{2}\|y - x\|_M^2, \quad \forall y \in \mathbb{R}^p, \quad (5)$$

where $\|z\|_M = (z^T M z)^{1/2}$.

Assumption A3. *Suppose $\theta_\star$ minimizes $L(\theta) = \mathbf{E}[\ell(Z, \theta)]$. Then $\ell(Z, \cdot)$ is $\Lambda(Z)$-convex at $\theta_\star$ a.s. with $\Lambda(Z) \succeq 0$ and $\Lambda_0 = \mathbf{E}[\Lambda(Z)]$ is positive definite so that $\lambda = \lambda_{\min}(\Lambda_0) > 0$, where $\lambda_{\min}(M)$ is the smallest eigenvalue of symmetric matrix M.*

Assumption A4. $\mathbf{E}\|\nabla\ell(Z, \theta_\star)\|^2 \le \sigma^2 < \infty$.

Remark 2.1. *If f is differentiable, then condition (5) is equivalent to*

$$(y - x)^T(\nabla f(y) - \nabla f(x)) \ge \|y - x\|_M^2.$$

Observe that if f is M-convex, then it is μI-convex with $\mu = \lambda_{\min}(M)$, where I is the identity matrix. The latter is equivalent to the standard notion of μ-convexity (if $\mu > 0$, then f is strongly convex) (Bauschke & Combettes, 2011).

Remark 2.2. *From Assumptions A1, A2, and A4 the objective function $L(\theta) = \mathbf{E}[\ell(Z, \theta)]$ is well-defined for all $\theta \in \mathbb{R}^p$. Furthermore, L is continuously differentiable and its gradient has a representation $\nabla L(\theta) = \mathbf{E}[\nabla\ell(Z, \theta)]$ (Bertsekas, 1973). Assumption A3 implies that L is λ-convex at $\theta_\star$. Since $\lambda > 0$, it also implies that the minimizer $\theta_\star$ is unique.*

Throughout, we fix the learning rate schedule as follows.

$$\gamma_n = \gamma_1 n^{-\gamma} \quad \text{for some } \gamma_1 > 0 \text{ and } \gamma > 0. \quad (R)$$

The valid range of the exponent γ depends on the algorithm and the conditions on L; the subsequent discussions will elaborate on this.

Notation. We employ the following asymptotic notation. For a sequence of random vectors/matrices $\{A_n\}$ defined on $(\Omega, \mathcal{F}, P)$ and a positive scalar sequence $\{b_n\}$, $A_n = O(b_n)$ means $\mathbf{E}[\|A_n\|] \le cb_n$ for some $c > 0$ and for all $n = 1, 2, \ldots$. On the other hands, $A_n = o(b_n)$ means $\mathbf{E}[\|A_n\|]/b_n \to 0$ as $n \to \infty$. Here the norm $\|\cdot\|$ refers to the (operator) 2-norm. Notation $b_n \downarrow 0$ means that $\{b_n\}$ is positive and converges monotonically toward zero.

3. Approximation of ISGD by SGD

The results presented in Sects. 4 and 5 rely on the following proposition bounding the difference between ISGD (proxRM) and (explicit) SGD iterates. This result holds without strong convexity.

Proposition 3.1 (Approximation of ISGD by SGD)**.** *In addition to Assumptions A1, A2, and A4, also assume that a minimizer $\theta_\star$ of L exists (not necessarily unique). Then,*

$$\theta_n = \theta_{n-1} - \gamma_n \nabla\ell(Z_n, \theta_{n-1}) + R_n,$$

with

$$\mathbf{E}[\|R_n\| | \mathcal{F}_{n-1}] \le \gamma_n^2\beta_0^2\|\theta_{n-1} - \theta_\star\| + \tfrac{\gamma_n^2}{2}(\beta_0^2 + \sigma^2) \quad (6a)$$

$$\mathbf{E}[\|R_n\|^2 | \mathcal{F}_{n-1}] \le 8\gamma_n^2\beta_0^2\|\theta_{n-1} - \theta_\star\|^2 + 8\gamma_n^2\sigma^2 \quad (6b)$$

$$\mathbf{E}\|R_n\| \le \gamma_n^2[\beta_0^2(r + 1/2) + \sigma^2/2] \quad (6c)$$

$$\mathbf{E}\|R_n\|^2 \le 8\gamma_n^2(\beta_0^2 r^2 + \sigma^2), \quad (6d)$$

where $r = \left(\|\theta_0 - \theta_\star\|^2 + \sigma^2 \sum_{k=1}^\infty \gamma_k^2\right)^{1/2}$. Inequalities (6a) and (6b) hold for $\gamma \in (0, 1]$, and inequalities (6c) and (6d) hold for $\gamma \in (1/2, 1]$.

Proposition 3.1 relies on the following intermediate result:

Lemma 3.1. *Under Assumptions A1, we have*

$$\|\theta_n - \theta_{n-1}\| = \gamma_n \|\nabla \ell(Z_n, \theta_n)\| \leq \gamma_n \|\nabla \ell(Z_n, \theta_{n-1})\|.$$

Proposition 3.1 and Lemma 3.1 jointly play the role of Theorem 3.1 in Toulis & Airoldi (2017a), which states that $\nabla \ell(Z_n, \theta_n)$ has the same direction as $\nabla \ell(Z_n, \theta_{n-1})$ under the GLM assumption, expressing ISGD as a variant of SGD. In the absence of this special relation, knowing that the norm of R_n is $O(\gamma_n^2)$, not just $O(\gamma_n)$ as can be inferred from $\|R_n\|^2 = O(\gamma_n^2)$, is crucial since it allows (in principle) the techniques of bounding the estimation errors of explicit SGD (e.g., Bach & Moulines, 2011) can be employed by controlling the impact of the additional term R_n (e.g., Theorem 4.1). Furthermore, in the proof of asymptotic normality (Theorem 4.3), Proposition 3.1 is used to obtain the $o(\gamma_n^{3/2}) = n^{-\frac{3}{2}\gamma} \cdot o(1)$ error term in the recursive equation for the estimation error $\theta_n - \theta_\star$ (see Eqs. (A.16) and (A.17)). The resulting recursion admits the use of the central limit theorem due to Fabian (1968, Theorem 2.2) for classical stochastic approximation (including the explicit SGD) can be employed almost directly. Proposition 3.1 is also indispensable (through Theorem 4.3) in showing consistency of the proposed on-line estimators of the asymptotic covariance matrices (Theorem 4.4 and Corollary 4.1).

4. Strongly convex objectives

4.1. Proximal Robbins-Monro

4.1.1. Stability

Under Assumptions A1–A4, Bianchi (2016) and Asi & Duchi (2019, Proposition 3.8) show that the proxRM iterate θ_n converges to the unique solution $\theta_\star$ a.s. A finite-sample (non-asymptotic) mean-squared error bound of θ_n in estimating model parameter $\theta_\star$ is obtained by Patrascu & Necoara (2017, Theorem 14) for $\gamma \in (0, 1]$. A simpler bound for $\gamma \in (1/2, 1]$ can be found as follows. We provide the proof in Appendix A since it shows how the analysis of Toulis & Airoldi (2017a) extends to non-GLM settings (and without the contradictory assumptions).

Definition 4.1. $\phi_\gamma(n) \triangleq (n^{1-\gamma} - 1)/(1-\gamma)$ *if* $\gamma \neq 1$, *and* $\phi_\gamma(n) \triangleq \log n$ *if* $\gamma = 1$.

Theorem 4.1 (Non-asymptotic point estimation error bound)**.** *Under Assumptions A1–A4, for any initial step size* $\gamma_1 > 0$ *when* $\gamma \in (1/2, 1)$ *and for* $\gamma_1 > 1/(2\lambda)$ *when* $\gamma = 1$, *there exist a fixed integer* n_0 *and constants* K_1, D_{n_0} *such that*

$$\mathbf{E}\|\theta_n - \theta_\star\|^2 \leq K_1 n^{-\gamma} + \exp\left(-\tfrac{1}{2}\log(1 + 2\lambda\gamma_1)\phi_\gamma(n)\right) \times (\|\theta_0 - \theta_\star\|^2 + D_{n_0}), \quad n \in \mathbb{N}. \tag{7}$$

All of the constants in inequality (7) are explicit, and are presented at the end of the proof given in Appendix A.3.

The result of Theorem 4.1 can be summarized as $\mathbf{E}\|\theta_n - \theta_\star\|^2 = O(\gamma_n)$. This asymptotic rate of $O(n^{-\gamma})$ matches that of SGD (Bach & Moulines, 2011) and the idealized proxRM with the same learning rate schedule (R) with $\gamma \in (0, 1]$. For $\gamma < 1$, the effect of the initial point is forgotten at an exponential rate (second term in the right-hand side of the inequality (7)). For $\gamma = 1$, the rate is also $O(\gamma_n)$ provided that $\gamma_1 \geq (e^2 - 1)/(2\lambda)$, although the "exponential forgetting" behavior is not so much prominent (it is polynomial). Either way, this insensitivity to the initial step size is one that features ISGD in contrast to SGD, where in the latter the impact of the initial point may *exponentially increase* with the initial step size γ_1 in the transient phase (Bach & Moulines, 2011), while in the former it always decreases.

The stability of proxRM can also be formalized in a non-MSE fashion. In order to analyze the error $\theta_n - \theta_\star$ (not $\|\theta_n - \theta_\star\|^2$), we need an additional assumption:

Assumption B1. *The objective function* L *is twice differentiable at* $\theta_\star$.

The Hessian $\nabla^2 L$ of L at $\theta_\star$ is denoted by $\mathcal{H}(\theta_\star)$. Note $0 < \lambda \leq \lambda_{\min}(\mathcal{H}(\theta_\star))$.

Theorem 4.2 (Stability)**.** *Under Assumptions A1–A4 and B1, if* $\gamma \in (1/2, 1)$ *or* $\gamma = 1$ *and* $\gamma_1 \geq (e^2 - 1)/\lambda$, *then the proxRM iterate* $\{\theta_n\}$ *satisfies the following.*

$$\mathbf{E}[\theta_n - \theta_\star] = Q_1^n(\theta_0 - \theta_\star) + o(1) \tag{8}$$

where $Q_1^n = \prod_{i=1}^n [I + \gamma_i \mathcal{H}(\theta_\star)]^{-1}$.

In contrast, if $\{\vartheta_n\}$ denotes the explicit SGD iterate (started from the same initial point), then

$$\mathbf{E}[\vartheta_n - \theta_\star] = P_1^n(\theta_0 - \theta_\star), \quad P_1^n = \prod_{i=1}^n [I - \gamma_i \mathcal{H}(\theta_\star)],$$

ignoring the remainder term in the second-order Taylor expansion of L at $\theta_\star$. To avoid the explosion of the leading eigenvalue of P_1^n, it is desirable to control the initial step size $\gamma_1 < 2/\lambda_{\max}(\mathcal{H}(\theta_\star)) \leq 2/\lambda$. In other words, in SGD the initial step size should not be large. On the other hand, in proxRM the eigenvalues of Q_1^n is always strictly smaller than 1 *regardless of the choice of* γ_1 (when $\gamma < 1$). Thus in proxRM the step sizes can be taken large to promote fast convergence. A similar informal discussion can be found in (Toulis & Airoldi, 2017a, Sect. 2.5); we derive equation (8) formally under weaker assumptions on differentiablility and without the GLM model.

4.1.2. Inference

Asymptotic normality. The following result generalizes that of Toulis & Airoldi (2017a) under weaker differentia-

bility assumptions and without the GLM model. As already mentioned in Sect. 3, the key instrument for deriving this result is Proposition 3.1, which quantifies the degree of approximation to SGD by ISGD. Our proof also fixes a flaw in the proof of Toulis & Airoldi (2017a, Theorem 2.4); see Remark A.1.

In order to establish asymptotic normality, we need additionally the following assumption on the stochastic error:

Assumption B2. *Let* $\sigma_{n,s}^2 = \mathbf{E}(I_{\|\varepsilon_n(\theta_\star)\|^2 \geq s/\gamma_n} \|\varepsilon_n(\theta_\star)\|)$ *where* $\varepsilon_n(\theta_\star) = \nabla\ell(Z_n, \theta_\star) - \nabla L(\theta_\star)$. *Then for all* $s > 0$, $\sum_{i=1}^n \sigma_{i,s}^2 = o(n)$ *if* $\gamma = 1$, *and* $\sigma_{n,s}^2 = o(1)$ *otherwise.*

Theorem 4.3 (Asymptotic normality)**.** *Suppose Assumptions A1– A4 and B1–B2 hold. Then, the proxRM iterate* θ_n *is asymptotically normal, such that*

$$n^{\gamma/2}(\theta_n - \theta_\star) \xrightarrow{d} \mathcal{N}_p(0, \Sigma) \tag{9}$$

where

$$\Sigma = \begin{cases} \gamma_1^2 \mathcal{L}^{-1}_{2\gamma_1 \mathcal{H}(\theta_\star) - I}(\mathcal{I}(\theta_\star)), & \gamma = 1,\ \gamma_1 \geq \frac{e^2-1}{\lambda}, \\ \gamma_1^2 \mathcal{L}^{-1}_{2\gamma_1 \mathcal{H}(\theta_\star)}(\mathcal{I}(\theta_\star)), & \gamma \in (1/2, 1), \end{cases} \tag{10}$$

for $\mathcal{I}(\theta_\star) = \mathbf{E}[\nabla\ell(Z, \theta_\star)\nabla\ell(Z, \theta_\star)^T]$. *Here,* $\mathcal{L}_P^{-1}$ *denotes the inverse operator of the Lyapunov linear map* $\mathcal{L}_P(X) = \frac{1}{2}(PX + XP)$ *for symmetric, positive definite matrix* P.

Remark 4.1. *The inverse of the Lyapunov map has a closed form:*

$$\mathrm{vec}(\mathcal{L}_{2B}^{-1}(Y)) = (I \otimes B + B \otimes I)^{-1} \mathrm{vec}(Y)$$

where $\otimes$ *denotes the Kronecker product and* $\mathrm{vec}(\cdot)$ *refers to the usual vectorization operator for matrices.*

Theorem 4.3 emphasizes that, in order for proxRM to converge (in distribution), $\gamma_1 > 1/(2\lambda) \geq 1/[2\lambda_{\min}(\mathcal{H}(\theta_\star))]$ is necessary, at least for $\gamma = 1$. (Note $\gamma_1 \geq (e^2 - 1)/\lambda$ implies this condition.) Thus in proxRM a large initial step is *promoted*, rather than prohibited for a stability concern as in explicit SGD; see the discussion after Theorem 4.2.

Estimation of the asymptotic covariance matrix. An obvious way of consistently estimating the the covariance matrix Σ of the asymptotic distribution (9) is to run n proxRM iterations B times independently, and take

$$B^{-1} \textstyle\sum_{i=1}^B (\theta_n^{(i)} - \bar{\theta})(\theta_n^{(i)} - \bar{\theta})^T, \tag{11}$$

where $\theta_n^{(i)}$ denotes the nth iterate for the ith run. This estimator is of course not very practical since many runs are required. Here we show a consistent estimator based on only a *single* run can be constructed. Specifically, we propose to use

$$\hat{\Sigma}_n = \begin{cases} \gamma_1^2 \mathcal{L}^{-1}_{2\gamma_1 \hat{H}_n - I}(\hat{I}_n), & \gamma = 1,\ \gamma_1 \geq \frac{e^2-1}{\lambda}, \\ \gamma_1^2 \mathcal{L}^{-1}_{2\gamma_1 \hat{H}_n}(\hat{H}_n), & \gamma \in (1/2, 1), \end{cases}$$

where

$$\begin{aligned} \hat{H}_n &= \frac{1}{n}\sum_{k=1}^n \nabla^2\ell(Z_k, \theta_{k-1}), \\ \hat{I}_n &= \frac{1}{n}\sum_{k=1}^n \nabla\ell(Z_k, \theta_{k-1})\nabla\ell(Z_k, \theta_{k-1})^T \end{aligned} \tag{12}$$

are plug-in estimators of $\mathcal{H}(\theta_\star)$ and $\mathcal{I}(\theta_\star)$, respectively. It is clear that both $\hat{H}_n$ and $\hat{I}_n$ can be computed in an on-line fashion, so can $\hat{\Sigma}_n$. Since $\hat{H}_n$ may not be positive definite in practice, a bit of adjustment on its eigenvalues may be needed to ensure invertibility (see Remark 4.1).

In Theorem 4.4 below, we show that with additional regulatory assumptions along with the positive-definite adjustment, $\hat{\Sigma}_n$ is a consistent estimator of Σ, from which valid inference on $\theta_\star$ can be implemented. The following assumptions are adopted from Chen et al. (2020).

Assumption B3. *For the error sequence* $\varepsilon_n = \nabla L(\theta_{n-1}) - \nabla\ell(Z_n, \theta_{n-1})$, *the fourth conditional moment is bounded as follows.*

$$\mathbf{E}[\|\varepsilon_n\|^4 | \mathcal{F}_{n-1}] \leq \Sigma_3 + \Sigma_4 \|\theta_{n-1} - \theta_\star\|^4$$

for some constants Σ_3 *and* Σ_4.

Assumption B4. *The sample function* $\ell(Z, \cdot)$ *is twice differentiable a.s., and is* M*-Lipschitz continuous at the minimizer* $\theta_\star$ *of* $L(\cdot) = \mathbf{E}[\ell(Z, \cdot)]$. *That is,* $\nabla^2\ell(Z, \theta)$ *exists for all* θ *and* $\|\nabla^2\ell(Z, \theta) - \nabla^2\ell(Z, \theta_\star)\| \leq M\|\theta - \theta_\star\|$ *for all* θ, *with probability one.*

Assumption B5. *The sample function* $\ell(Z, \cdot)$ *is twice differentiable a.s. The second moment of the Hessian is bounded:*

$$\|\mathbf{E}[(\nabla^2\ell(Z, \theta_\star))^2] - [\mathcal{H}(\theta_\star)]^2\| \leq L_4$$

for some constant L_4.

Remark 4.2. *According to Lemma 3.1 of Chen et al. (2020), Assumption B3 is satisfied if* $||\nabla^2\ell(\theta, Z)|| \leq H(Z)$ *for some* H *with a bounded fourth moment, Thus, it can be easily checked that these assumptions hold for common losses such as quadratic or logistic loss. In fact Assumptions B3–B5 correspond to part 3 of Assumption 3.2 and Assumption 4.1 of Chen et al. (2020). Note part 2 of their Assumption 3.2 is not used in the present paper.*

Theorem 4.4 (Consistency of plug-in estimator)**.** *Suppose Assumptions A1–A4, B1, and B3–B5 hold. Let* $\hat{H}_n$ *and* $\hat{I}_n$ *be as given in Equation* (12). *Let* $\beta_+ = 1$ *if* $\gamma = 1$ *and* $\beta_+ = 0$ *otherwise. Choose* $\delta \in (\beta_+/(2\gamma_1), \lambda_{\min}(\mathcal{H}(\theta_\star)))$ *and let* $\tilde{H}_n = P\,\mathrm{diag}(\max(d_1, \delta), \ldots, \max(d_p, \delta))P^T$ *for the spectral decomposition* $P\,\mathrm{diag}(d_1, \ldots, d_p)P^T$ *of* $\hat{H}_n$. *Then, for the asymptotic covariance matrix* (10),

$$\mathbf{E}\left\|\gamma_1^2 \mathcal{L}^{-1}_{2\gamma_1 \tilde{H}_n - I}(\hat{I}_n) - \Sigma\right\| = O(\gamma_n^{1/2})$$

if $\gamma = 1$, $\gamma_1 \geq \frac{e^2-1}{\lambda}$, $2\gamma_1 \mathcal{H}(\theta_\star) \succ I$; *if* $\gamma \in (\frac{1}{2}, 1)$,

$$\mathbf{E}\left\|\gamma_1^2 \mathcal{L}_{2\gamma_1 \tilde{H}_n}(\hat{I}_n) - \Sigma\right\| = O(\gamma_n^{1/2}).$$

Thus the $100(1-\alpha)\%$ confidence interval for the j-th component of $\theta_\star$ can be approximated by $\theta_{n,j} \pm z_{\alpha/2}\hat{\sigma}_{n,j}$ where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, and

$$\hat{\sigma}_{n,j} = \begin{cases} n^{-1/2}\gamma_1 \sqrt{[\mathcal{L}_{2\gamma_1 \tilde{H}_n - I}(\hat{I}_n)]_{jj}}, & \gamma = 1, \\ n^{-\gamma/2}\gamma_1 \sqrt{[\mathcal{L}_{2\gamma_1 \tilde{H}_n}(\hat{I}_n)]_{jj}}, & \gamma \in (1/2, 1). \end{cases} \tag{13}$$

4.2. Proximal Polyak-Ruppert

4.2.1. Stability

With slightly stronger assumptions on the objective function, the stability of proxPR can be analyzed:

Assumption A2′. *Function $\ell(Z,\cdot)$ is β_0-smooth a.s. around the minimizer $\theta_\star$ of $L(\cdot) = \mathbf{E}[\ell(Z,\cdot)]$. That is, a sample function $\ell(Z,\theta)$ is continuously differentiable in θ and $\|\nabla\ell(Z,\theta) - \nabla\ell(Z,\theta_\star)\| \leq \beta_0 \|\theta - \theta_\star\|$ for all θ, with probability one.*

Assumption A4′. $\mathbf{E}\|\nabla\ell(Z,\theta_\star)\|^4 \leq \sigma^4 < \infty$.

Note Assumption A2′ (*resp.* A4′) implies Assumption A2 (*resp.* A4).

Theorem 4.5 (Non-asymptotic point estimation error bound)**.** *Under Assumptions A1, A2′, A3, A4′, B1, and B4, for $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$, the following holds for $\gamma \in (1/3, 1)$.*

$$\begin{aligned}(\mathbf{E}\left\|\bar{\theta}_n - \theta_\star\right\|^2)^{1/2} &\leq \tfrac{1}{\sqrt{n}}\left[\mathrm{tr}\left(\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right)\right]^{1/2} \\ &+ \tfrac{K_1^{1/2}}{\lambda^{1/2}n}\Big((\beta_0 + \gamma_1^{-1})n^{\gamma/2} + 2\beta_0\phi_\gamma^{1/2}(n) \\ &\qquad + \gamma(\beta_0 + \gamma_1^{-1})\phi_{1-\gamma/2}(n) + \tfrac{M}{2}\phi_\gamma(n)\Big) \\ &+ \tfrac{\tilde{\mathcal{A}}}{\lambda^{1/2}n} + \tfrac{M\tilde{\mathcal{B}}}{2\lambda^{1/2}n} \\ &+ \tfrac{\beta_0 + \gamma_1^{-1}}{\lambda^{1/2}}\exp\big(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\big) \\ &\qquad \times (\|\theta_0 - \theta_\star\|^2 + D_{n_0})^{1/2}, \end{aligned} \tag{14}$$

where n_0, K_1, D_{n_0} are the same as in Theorem 4.1.

The constants $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ in inequality (14) are explicit, and are presented at the end of the proof given in Appendix A.7.

Compared with the (explicit) SGD (Bach & Moulines, 2011, Theorem 3), the allowed range of γ is a bit narrower. (This is due to Lemma A.2 and Corollary A.1 in the Appendix A.) However, when $M > 0$, to obtain the optimal $O(1/n)$ rate (in mean square error) independent of γ_n, we need $\gamma \in (1/2, 1]$, just as SGD (for $\gamma = 1$ we get a simpler bound by averaging the bound in Theorem 4.1); when $M = 0$ (L is quadratic) then the rate is $O(1/n)$ for all $\gamma \in (1/3, 1]$. The second slowest term has an order of either $O(n^{-(2-\gamma)})$ or $O(n^{-(1+\gamma)})$ due to the second line in inequality (14), suggesting $\gamma = 2/3$ to get a balance.

The rate of "forgetting the initial condition" consists of two parts, one with the rate of $O(1/n^2)$ involving quantities $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ and the other with an exponential rate of $O(\exp\left(-\frac{1}{2}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\right))$. Furthermore, the constant $\tilde{B}$ can be quite large: it involves the term exponential in $\phi_{\frac{5}{3}\gamma}(k)$, which is increasing if $\gamma < 3/5$ (see Appendix A). Thus, unlike proxRM, the impact of the initial condition many remain for a long time, and this is what we observed in the experiments in Sect. 6. Like in SGD, having a burn-in period before taking averages appears to be beneficial.

The key difference between proxPR the forward Polyak-Ruppert (averaged explicit SGD) is that there is no exponential term multiplied to the constant $\tilde{\mathcal{B}}$ in (14), which corresponds to the constant "A" in Theorem 3 of Bach & Moulines (2011). *Thus there is no "catastrophic term" in proxPR.* This suggests that we can take γ_1 large.

Remark 4.3. *In Toulis et al. (2016), a non-asymptotic error bound with rate $O(1/n)$ for any $\gamma \in (1/2, 1]$ is obtained for the averaged iterate, however under the contradictory assumption that the objective function L is both globally Lipschitz and globally strongly convex, as well as the GLM assumption. In particular, their Lemma 5 crucially depends on the Lipschitz continuity of L.*

4.2.2. Inference

Asymptotic normality. The asymptotic normality of proxPR, i.e., that of $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$, is established in Asi & Duchi (2019, Theorem 3.11) for $\gamma \in (1/2, 1)$. The asymptotic covariance matrix is $n^{-1}\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}$, which achieves the Cramér-Rao lower bound. Its rate of vanishing matches that of the asymptotic covariance matrix of the proxRM iterate θ_n, but $\bar{\theta}_n$ is statistically more efficient.

Estimation of the asymptotic covariance matrix. The proof of Theorem 4.4 involves consistency of $\hat{H}_n$ and $\hat{I}_n$ in estimating $\mathcal{H}(\theta_\star)$ and $\mathcal{I}(\theta_\star)$ respectively (see Lemma A.4) and that $\tilde{H}_n$ is asymptotically equivalent to $\hat{H}_n$. Hence $\tilde{H}_n^{-1}\hat{I}_n\tilde{H}_n^{-1}$ is a consistent estimator of $\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}$, the asymptotic covariance matrix of the scaled Polyak-Ruppert average $\sqrt{n}\bar{\theta}_n = \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}\theta_k$. It follows:

Corollary 4.1 (Plug-in estimator)**.** *Suppose Assumptions A1–A4, B1, and B3–B5 hold. Then, for $\tilde{H}_n$ and $\hat{I}_n$ as appeared in Theorem 4.4 and for $\gamma \in (1/2, 1)$,*

$$\mathbf{E}\left\|\tilde{H}_n^{-1}\hat{I}_n\tilde{H}_n^{-1} - \mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right\| = O(\gamma_n^{1/2}).$$

An asymptotic $100(1-\alpha)\%$ confidence interval for the j-th component of $\theta_\star$ is given by $\bar{\theta}_{n,j} \pm z_{\alpha/2}\tilde{\sigma}_{n,j}$ where

$$\tilde{\sigma}_{n,j} = n^{-1/2}\sqrt{[\tilde{H}_n^{-1}\hat{I}_n\tilde{H}_n^{-1}]_{jj}}, \quad \gamma \in (1/2, 1). \quad (15)$$

5. Non-strongly convex objectives

5.1. Proximal Robbins-Monro

In this and the next subsection, we do not assume that the objective function L is strongly convex. However we do assume that the minimum is attained. In other words, we replace Assumption A3 with a weaker one:

Assumption A3′. *There exists a minimizer $\theta_\star$ of the objective L. The minimizer may not be unique.*

Since the minimizer is not unique, we derive a finite-sample bound on the objective value.

Theorem 5.1 (Non-asymptotic optimization error bound)**.** *Under Assumptions A1, A2′, A3′, and A4, the following holds for the proxRM iterate (3):*

$$\mathbf{E}\left[L(\theta_n) - L(\theta_\star)\right] \le \begin{cases} \Gamma_1 \cdot \frac{\delta_0 + \gamma_1^2(1+\phi_{2\gamma}(n))}{\phi_{1-\gamma/2}(n) - \phi_{1-\gamma/2}(n_1-1)}, & \gamma \in (0, \frac{1}{2}], \\ \Gamma_2 \cdot \frac{\delta_0 + \gamma_1^2\zeta(2\gamma)}{\phi_{1-\gamma/2}(n) - \phi_{1-\gamma/2}(n_1-1)}, & \gamma \in (\frac{1}{2}, \frac{2}{3}), \\ \Gamma_3 \cdot \frac{\delta_0 + \gamma_1^2\zeta(2\gamma)}{\phi_{\gamma}(n) - \phi_{\gamma}(n_1-1)}, & \gamma \in [\frac{2}{3}, 1], \end{cases}$$

for $n \ge n_1 \triangleq \max\{\inf\{n \in \mathbb{N} : \gamma_n \le 1/\beta_0\}, 3\}$, where $\delta_0 = \|\theta_0 - \theta_\star\|^2$; $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$ is the Riemann zeta function.

The constants Γ_1, Γ_2, and Γ_3 are explicit, and are presented at the end of the proof given in Appendix A.8.

This bound is decreasing only if $\gamma > 2/5$. Thus the rate is $O(n^{-(1-5\gamma/2)})$ if $2/5 < \gamma < 1/2$, $O(n^{-1/4}\log n)$ if $\gamma = 1/2$, $O(n^{-\gamma/2})$ if $1/2 < \gamma < 2/3$, $O(n^{-(1-\gamma)})$ if $2/3 < \gamma < 1$, and $O(1/\log n)$ if $\gamma = 1$. The best rate is attained for $\gamma = 2/3$ which is $O(n^{-1/3})$. This is the same rate as the explicit SGD counterpart (Bach & Moulines, 2011, Theorem 4), while the latter is valid for $\gamma \in [1/2, 1]$. Furthermore, no catastrophic term appears in the bounds in Theorem 5.1.

The convergence rates and stability implied in Theorem 5.1 fully agree with Toulis et al. (2021, Theorem 2) obtained for the idealized, but infeasible, proxRM procedure ($\gamma \in (1/2, 1]$). Therefore, realizing the ideal procedure (4) by equation (3) loses nothing.

5.2. Proximal Poylak-Ruppert

Theorem 5.2 (Non-asymptotic optimization error bound)**.** *Under Assumptions A1, A2′, A3′, and A4, the following holds.*

$$\mathbf{E}\left[L(\bar{\theta}_n) - L(\theta_\star)\right] \le \begin{cases} \frac{\tilde{\Gamma}_1}{n} + \frac{\sigma^2\gamma_1^2}{(1-2\gamma)n^\gamma} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)n^\gamma}, & \gamma \in (0, \frac{1}{2}), \\ \frac{\tilde{\Gamma}_2}{n} + \frac{\sigma^2\gamma_1^2(1+\log n)}{\sqrt{n}} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)\sqrt{n}}, & \gamma = \frac{1}{2}, \\ \frac{\tilde{\Gamma}_3}{n} + \frac{\sigma^2\gamma_1^2\zeta(2\gamma)}{n^{1-\gamma}} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)n^\gamma}, & \gamma \in (\frac{1}{2}, 1), \\ \frac{\tilde{\Gamma}_3}{n} + \sigma^2\gamma_1^2\zeta(2) + \frac{2\tilde{\sigma}^2\gamma_1\log n}{(1-\gamma)n}, & \gamma = 1 \end{cases}$$

for $n \ge n_* \triangleq \inf\{k \in \mathbb{N} : (1-\gamma_k\beta_0) \ge 1/2\}$. Also,

$$\mathbf{E}\left[L(\bar{\theta}_n) - L(\theta_\star)\right] \le \begin{cases} \frac{\beta_0}{2}\left(\delta_0 + \frac{\sigma^2\gamma_1^2}{(1-2\gamma)(2-2\gamma)}n^{1-2\gamma}\right), & \gamma \in (0, \frac{1}{2}), \\ \frac{\beta_0}{2}\left(\delta_0 + \sigma^2\gamma_1^2\frac{\log(n+1)}{n}\right), & \gamma = \frac{1}{2}, \\ \frac{\beta_0}{2}\left(\delta_0 + \frac{\sigma^2\gamma_1^2\zeta(2\gamma)}{n}\right), & \gamma \in (\frac{1}{2}, 1], \end{cases}$$

for $n < n_*$, where $\delta_0 = \|\theta_0 - \theta_\star\|^2$.

The constants $\tilde{\Gamma}_1$, $\tilde{\Gamma}_2$, $\tilde{\Gamma}_3$, and $\tilde{\sigma}^2$ are explicit, and are presented at the end of the proof given in Appendix A.9.

The convergence rate can be summarized as $O(n^{-\gamma} \vee n^{-(1-\gamma)})$. Thus with $\gamma = 1/2$, we attain the minimax optimal asymptotic rate of $O(n^{-1/2})$ (Ruppert, 1988) up to a logarithmic factor. We have thereby confirmed the conjecture by Toulis et al. (2021, Remark 3.1) that the minimax rate may be achieved by averaging the (ideal) proxRM iterates, i.e., proxPR. Moreover, our procedure is feasible. Compared to explicit SGD (Bach & Moulines, 2011, Theorem 6), again no catastrophic term is involved and the initial step size γ_1 can be taken large to obtain the best rate.

6. Experiments

6.1. Point estimation and optimization

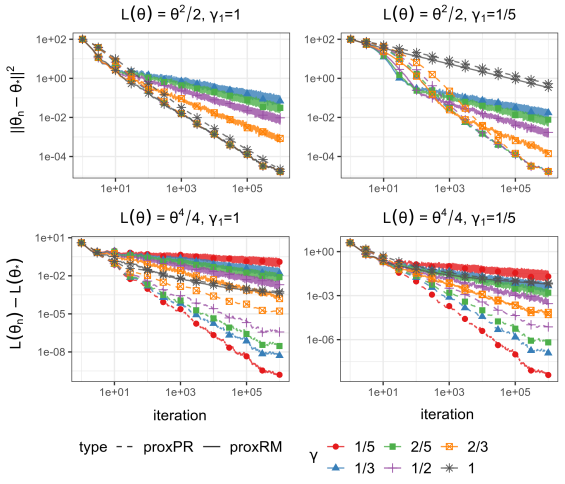

Figure 1. Error reduction for various learning rates.

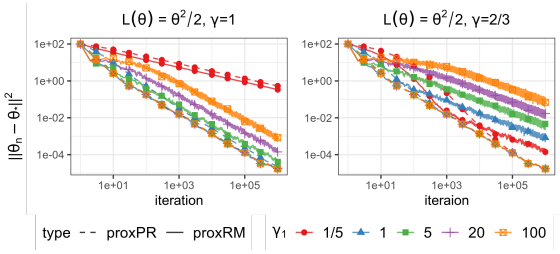

Figure 2. Error reduction for various values of initial step size.

Following Bach & Moulines (2011), we examined the convergence behavior of proxRM and proxPR using two univariate functions: $L(\theta) = \frac{1}{2}\theta^2$ (strongly convex) and $L(\theta) = \frac{1}{4}\theta^4$ (non-strongly convex); the sample functions were chosen $\ell(Z,\theta) = \frac{1}{2}\theta^2 + Z\theta$ and $\ell(Z,\theta) = \frac{1}{4}\theta^4 + Z\theta$, where $Z \sim N(0,4^2)$. We fixed the initial point $\theta_0 = 10$ for the quadratic and $\theta_0 = 2$ for the quartic function, and observed 100 independent runs of $n = 10^6$ ISGD iterations for initial step size $\gamma_1 \in \{1/5, 1, 5, 20, 100\}$ and exponent $\gamma \in \{1/5, 1/3, 2/5, 1/2, 2/3, 1\}$.

Figs. 1 and 2 plot the squared estimation ($\|\theta_n - \theta_\star\|^2$) or optimization ($L(\theta_n) - L(\theta_\star)$) error averaged over all replications for each iteration. The trend generally aligns with the results in Sects. 4 and 5. First, several instances of proxPR showed slower reduction than proxRM in the early stage of the iteration, as expected by the theory. For the strongly convex case, the error reduction was proportional to γ in proxRM and at the same fastest convergence rate in proxPR for all values except for $\gamma = 1$, for which too small an initial step size showed a noticeable slowdown. For a given γ, the rate was proportional to γ_1 in proxRM, ending up with parallel lines in Fig. 2; in proxPR, the rate was independent of γ and γ_1. Note that an initial step size as large as 100 was allowed, with no explosion, confirming the stability of ISGD over SGD. For the non-strongly convex case, $\gamma = 1$ was the fastest among proxRM if γ_1 was not too small but became the slowest among proxPR, for which slower decays led to faster convergence.

6.2. Interval estimation and inference

To evaluate the performance of the plug-in estimators of the asymptotic covariance matrices (Theorem 4.4 and Corollary 4.1), we conducted statistical inference on model parameters in linear regression and quantile estimation models using ISGD; see Chen et al. (2020); Toulis et al. (2021). Based on the observations on proxPR in the previous subsection, we discarded the first tenth of the iterations when calculating $\tilde{H}_n, \hat{I}_n$, and $\bar{\theta}_n$. Then, we gathered the average rate of the nominal 95% confidence interval (13), (15) covering θ_* for each coordinate ("cover"), mean squared error from θ_* ("MSE"), and the length of the 95% confidence interval ("lenCI") from $B = 500$ independent replications.

We also compared the plug-in estimator of each run with the multi-run estimator (11) using the Frobenius norm (normalized by dimension p); its average is reported ("covdiff").

Linear regression. We generated $Z_n = (y_n, \mathbf{x}_n)$ where $y_n = \mathbf{x}_n^T\theta_* + \epsilon_n$, $\mathbb{R}^p \ni \mathbf{x}_n \sim N(0,\Sigma)$, and $\epsilon_n \sim N(0,1)$, for $\ell(Z_n,\theta) = \frac{1}{2}(y_n - \mathbf{x}_n^T\theta)^2$. Three different covariance structures were considered: identity ($\Sigma = I_p$), Toeplitz ($\Sigma_{ij} = (0.5)^{|i-j|}$), and equicorrelation ($\Sigma_{ij} = 0.2$ for $i \neq j$, $\Sigma_{ii} = 1$). We fixed $\theta_\star = (1,\ldots,1)^T$ and ran $n = 10^5$ iterations of ISGD for $\gamma \in \{0.6, 1.0\}$, $p \in \{5, 20, 100, 200\}$ with $\theta_0 = 0$ for each type of Σ. Table 2 collects the results. The coverage rates generally observed the nominal level, with proxRM showing a slightly better coverage, possibly because proxPR exhibited shorter confidence intervals.

Quantile estimation. Since the sample function for quantile estimation introduced in Sect. 1 is nondifferentiable, we instead used a smoothed version in which the $\max(0,\cdot)$ part is replaced by a quadratic function $(4\mu)^{-1}(x+\mu)^2$ on $[-\mu,\mu]$. The smoothing parameter μ makes $\nabla\ell(Z,\theta)$ $(2\mu)^{-1}$-Lipschitz and introduces a bias. We nevertheless set $\theta_\star$ to be the 99%-ile of the standard normal and let $Z \sim N(0,1)$. The $n = 10^6$ iterations were started with $\theta_0 = 0$ for each replication, where $\gamma \in \{0.6, 1\}$ and $\mu \in \{10^{-1}, 10^{-2}, 10^{-3}\}$; we used $\gamma_1 = 250$ when $\gamma = 1$ and $\gamma_1 = 30$ when $\gamma = 0.6$. Table 3 summarizes the results. While proxPR exhibits smaller MSE and covdiff, the coverage rate of proxRM was close to 95% and there are few cases that proxPR deteriorates. This result may be misleading since the actual minimizer of L is not $\theta_\star$ here, and the length of the confidence interval is smaller for proxPR. On the other hand, if the genuine goal is to find the quantile, the better coverage of (biased) proxRM might be a blessing.

7. Conclusion

In both point estimation of the model parameter for the strongly convex case and optimization of non-strongly convex functionals, the behavior of implicit SGD as revealed by our non-asymptotic analysis, either proxRM or proxPR, resembles the explicit SGD when the gradient is uniformly bounded a.s. on a bounded domain (Bach & Moulines, 2011, Theorems 2, 5, 7), which replaces Assumption(s) A4 (and, with strong convexity, A2). Thus ISGD effectively imposes the latter condition without requiring it, operating under weaker assumptions on the gradient map.

Like its explicit counterpart, our analysis (and experiments) states that averaging brings to ISGD robustness to the initial step size and lack of strong convexity. On the other hand, averaged ISGD (proxPR) may suffer from slow progress in the early phase of iterations. Burn-in is a viable option.

It is interesting to note that the plug-in estimators of the asymptotic covariance matrices of proxRM and proxPR,

Table 2. Statistical inference for linear regression model parameters.

			ProxRM				ProxPR			
Σ	γ	p	cover	MSE	lenCI	covdiff	cover	MSE	lenCI	covdiff
Identity	0.6	5	95.92	4.991e-3	0.285	5.817e-4	95.00	1.177e-5	0.013	1.299e-6
		20	97.54	4.424e-3	0.298	1.640e-3	95.01	1.270e-5	0.014	2.657e-6
		100	99.59	3.348e-3	0.334	4.180e-3	95.11	1.611e-5	0.016	7.309e-6
		200	99.97	2.511e-3	0.352	5.783e-3	95.05	1.782e-5	0.017	1.129e-5
	1	5	94.56	5.278e-5	0.028	3.436e-6	94.52	1.170e-5	0.013	1.344e-6
		20	95.14	5.245e-5	0.028	1.050e-5	94.70	1.151e-5	0.013	2.361e-6
		100	95.33	5.260e-5	0.029	2.353e-5	94.39	1.184e-5	0.013	5.456e-6
		200	95.48	5.197e-5	0.029	3.299e-5	94.44	1.192e-5	0.013	7.547e-6
Toeplitz	0.6	5	95.88	4.890e-3	0.285	6.979e-4	94.96	1.766e-5	0.017	2.037e-6
		20	97.40	4.439e-3	0.298	1.617e-3	94.96	2.082e-5	0.018	4.247e-6
		100	99.61	3.347e-3	0.334	4.185e-3	94.99	2.679e-5	0.020	1.194e-5
		200	99.96	2.525e-3	0.352	5.779e-3	95.19	2.959e-5	0.021	1.894e-5
	1	5	94.92	5.486e-5	0.029	4.976e-6	94.32	1.809e-5	0.016	1.874e-6
		20	95.05	5.518e-5	0.029	1.066e-5	93.95	1.985e-5	0.017	4.767e-6
		100	94.96	5.520e-5	0.029	2.487e-5	93.61	2.067e-5	0.017	9.691e-6
		200	95.47	5.396e-5	0.029	3.424e-5	93.88	2.083e-5	0.017	1.337e-5
EquiCorr	0.6	5	95.64	4.968e-3	0.285	6.030e-4	94.68	1.306e-5	0.014	1.490e-6
		20	97.42	4.454e-3	0.298	1.604e-3	94.90	1.532e-5	0.015	3.125e-6
		100	99.60	3.394e-3	0.334	4.159e-3	95.20	1.985e-5	0.018	8.990e-6
		200	99.95	2.583e-3	0.353	5.773e-3	95.13	2.228e-5	0.019	1.412e-5
	1	5	94.72	5.307e-5	0.029	3.199e-6	94.00	1.307e-5	0.014	1.518e-6
		20	95.15	5.326e-5	0.029	1.058e-5	94.49	1.403e-5	0.014	2.853e-6
		100	95.19	5.337e-5	0.029	2.391e-5	94.25	1.480e-5	0.015	6.827e-6
		200	95.45	5.267e-5	0.029	3.334e-5	94.36	1.497e-5	0.015	9.483e-6

Table 3. Statistical inference for quantile estimation.

Method	γ	μ	cover	MSE	lenCI	covdiff
ProxRM	0.6	1e-3	95.20	1.339e-3	0.148	1.185e-4
		1e-2	95.60	1.315e-3	0.146	9.965e-5
		1e-1	95.40	1.191e-3	0.140	8.716e-5
	1	1e-3	96.60	4.933e-5	0.028	1.689e-6
		1e-2	96.00	4.887e-5	0.028	1.021e-6
		1e-1	91.40	5.769e-5	0.027	6.967e-7
ProxPR	0.6	1e-3	73.80	2.626e-5	0.016	4.363e-6
		1e-2	73.40	2.587e-5	0.015	3.262e-6
		1e-1	96.40	1.266e-5	0.015	2.755e-6
	1	1e-3	94.60	1.603e-5	0.016	1.246e-6
		1e-2	95.60	1.589e-5	0.015	5.692e-7
		1e-1	87.40	2.872e-5	0.015	5.506e-7

indispensable for valid statistical inference on the model parameter, does not benefit from averaging. Since this paper appears to be the first to propose estimators based only on a single run, investigation of more statistically efficient estimators as well as online ones as in SGD (Chen et al., 2020) will make a promising avenue of future research.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019.

Bach, F. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24: 451–459, 2011. Full-length version: https://hal.archives-ouvertes.fr/hal-00608041.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

Bertsekas, D. P. Stochastic optimization problems with nondifferentiable cost functionals. *J. Optim. Theory Appl.*, 12(2):218–231, 1973.

Bertsekas, D. P. Incremental proximal methods for large scale convex optimization. *Math. Program., Ser. B.*, 129 (163), 2011.

Bianchi, P. Ergodic convergence of a stochastic proximal point algorithm. *SIAM J. Optim.*, 26(4):2235–2260, 2016.

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

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60 (2):223–311, 2018.

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 2020.

Fabian, V. On asymptotic normality in stochastic approximation. *Ann. Math. Stat.*, 39(4):1327–1332, 1968.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, T., Liu, L., Kyrillidis, A., and Caramanis, C. Statistical inference using sgd. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

Liang, T. and Su, W. J. Statistical inference for the population landscape via moment-adjusted stochastic gradients. *J. R. Stat. Soc.*, 2019.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

Patrascu, A. and Necoara, I. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *J. Mach. Learn. Res. (JMLR)*, 18(1):7204–7245, 2017.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30 (4):838–855, 1992.

Robbins, H. and Monro, S. A stochastic approximation method. *Ann. Math. Stat.*, 22(3):400–407, 1951.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

Ruppert, D. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Ryu, E. K. and Boyd, S. Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Manuscript*, 2014.

Toulis, P. and Airoldi, E. M. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist.*, 45(4):1694–1727, 2017a.

Toulis, P. and Airoldi, E. M. Supplement to 'asymptotic and finite-sample properties of estimators based on stochastic gradients', 2017b.

Toulis, P., Airoldi, E., and Rennie, J. Statistical analysis of stochastic gradient methods for generalized linear models. In *International Conference on Machine Learning*, pp. 667–675. PMLR, 2014.

Toulis, P., Tran, D., and Airoldi, E. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pp. 1290–1298. PMLR, 2016.

Toulis, P., Horel, T., and Airoldi, E. M. The proximal Robbins-Monro method. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(1):188–212, 2021.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Int. Conf. Mach. Learn. (ICML 2003)*, pp. 928–936, 2003.

A. Proofs of main results

A.1. Preliminary

In this section, we collect known facts useful for the subsequent proofs.

Proposition A.1 (Bach & Moulines (2011), p. 17)**.**

$$\|a+b\|^k \leq 2^{k-1}(\|a\|^k + \|b\|^k), \quad k = 1, 2, 3, 4.$$

Proposition A.2.

$$\gamma_n \leq \gamma_{n-1} \leq 2\gamma_n, \quad n \geq 2;$$
$$\gamma_n^{1/2} - \gamma_{n+1}^{1/2} \geq \frac{\gamma_1^{1/2}}{4n^{\gamma/2}}, \quad n \geq 3.$$

Proposition A.3. *Let $\phi_\gamma(n) = (n^{1-\gamma} - 1)/(1-\gamma)$ if $\gamma > 0$ and $\gamma \neq 1$, and $\phi_\gamma(n) = \log n$ if $\gamma = 1$. Then,*

$$\sum_{k=1}^{n} \frac{1}{k^\gamma} \leq 1 + \phi_\gamma(n) \leq \begin{cases} \frac{n^{1-\gamma}}{1-\gamma}, & \gamma \in (0,1), \\ 1 + \log n, & \gamma = 1, \\ \zeta(\gamma) := \sum_{k=1}^{\infty} \frac{1}{k^\gamma} < \infty, & \gamma > 1 \end{cases}$$

and

$$\frac{1}{2}[\phi_\gamma(n) - \phi_\gamma(m)] \leq \sum_{k=m+1}^{n} \frac{1}{k^\gamma} \leq \phi_\gamma(n) - \phi_\gamma(m), \quad n > m \geq 1.$$

The last inequality is from Bach & Moulines (2011, p. 13).

A.2. Proof of Proposition 3.1

Proposition 3.1 can be proved by combining the following intermediate result and Lemma 3.1.

Lemma A.1 (Theorem 3.3 and Corollary 3.3 of Asi & Duchi (2019))**.** *In addition to Assumptions A1, A2, and A4, also assume that a minimizer $\theta_\star$ of L exists (not necessarily unique). Then, the iterates from the ISGD update (3) with learning rate sequence (R) satisfies*

$$\mathbf{E}[\|\theta_n - \theta_\star\|^2 | \mathcal{F}_{n-1}] \leq \|\theta_{n-1} - \theta_\star\|^2 + \sigma^2\gamma_n^2, \quad \gamma \in (0, 1],$$
$$\mathbf{E}\|\theta_n - \theta_\star\|^2 \leq \|\theta_0 - \theta_\star\|^2 + \sigma^2 \sum_{k=1}^{\infty} \gamma_k^2 = r^2 < \infty, \quad \gamma \in (1/2, 1].$$

for all $n = 1, 2, \ldots$.

Proof of Proposition 3.1. Note that

$$\begin{aligned} \theta_n &= \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_n) \\ &= [\theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_{n-1})] + [\gamma_n \nabla \ell(Z_n, \theta_{n-1}) - \gamma_n \nabla \ell(Z_n, \theta_n)]. \end{aligned}$$

Thus $R_n = \gamma_n[\nabla \ell(Z_n, \theta_{n-1}) - \nabla \ell(Z_n, \theta_n)]$. Then,

$$\begin{aligned} \|R_n\| &= \gamma_n \|\nabla \ell(Z_n, \theta_{n-1}) - \nabla \ell(Z_n, \theta_n)]\| \\ &\leq \gamma_n \beta(Z_n) \|\theta_{n-1} - \theta_n\| \leq \gamma_n^2 \beta(Z_n) \|\nabla \ell(Z_n, \theta_{n-1})\|, \end{aligned} \tag{A.1}$$

where the first inequality is due to Assumption A2, and the second due to Lemma 3.1.

We show the bounds on $\|R_n\|$. First, observe from the triangular inequality and Assumption A2 that

$$\|\nabla \ell(Z_n, \theta_{n-1})\| \leq \beta(Z_n) \|\theta_{n-1} - \theta_\star\| + \|\nabla \ell(Z_n, \theta_\star)\|. \tag{A.2}$$

Then,

$$\begin{aligned}\mathbf{E}\left[\|R_n\|\,|\mathcal{F}_{n-1}\right] &\leq \gamma_n^2\,\mathbf{E}\left[\beta^2(Z_n)\right]\|\theta_{n-1}-\theta_\star\| \\ &\quad + (1/2)\gamma_n^2\,\mathbf{E}\left[\beta^2(Z_n)\right] + (1/2)\gamma_n^2\,\mathbf{E}\,\|\nabla\ell(Z_n,\theta_\star)\|^2 \\ &\leq \gamma_n^2\beta_0^2\|\theta_{n-1}-\theta_\star\| + (1/2)\gamma_n^2\beta_0^2 + (1/2)\gamma_n^2\sigma^2\end{aligned} \tag{A.3}$$

using $2ab \leq a^2+b^2$ and Assumption A2. Therefore for $\gamma \in (1/2, 1]$ with which $r < \infty$,

$$\mathbf{E}\,\|R_n\| \leq \gamma_n^2\beta_0^2\,\mathbf{E}\,\|\theta_{n-1}-\theta_\star\| + (1/2)\gamma_n^2(\beta_0^2+\sigma^2) \leq \gamma_n^2[\beta_0^2(r+1/2)+\sigma^2/2].$$

The last inequality is due to Lemma A.1 and Jensen's inequality.

To establish bounds on $\|R_n\|^2$, from inequality (A.2),

$$\|\nabla\ell(Z_n,\theta_{n-1})\|^2 \leq 2\beta^2(Z_n)\,\|\theta_{n-1}-\theta_\star\|^2 + 2\,\|\nabla\ell(Z_n,\theta_\star)\|^2$$

using $(a+b)^2 \leq 2a^2+2b^2$. So

$$\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_{n-1})\|^2\,|\mathcal{F}_{n-1}\right] \leq 2\beta_0^2\,\|\theta_{n-1}-\theta_\star\|^2 + 2\sigma^2 \tag{A.4}$$

from Assumption A4. Then, by the definition of R_n,

$$\begin{aligned}\mathbf{E}\left[\|R_n\|^2\,|\mathcal{F}_{n-1}\right] &= \gamma_n^2\,\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_n)-\nabla\ell(Z_n,\theta_{n-1})\|^2\,|\mathcal{F}_{n-1}\right] \\ &= \gamma_n^2\,\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_n)\|^2\,|\mathcal{F}_{n-1}\right] - 2\gamma_n^2\,\mathbf{E}\left[\nabla\ell(Z_n,\theta_n)^T\nabla\ell(Z_n,\theta_{n-1})|\mathcal{F}_{n-1}\right] \\ &\quad + \gamma_n^2\,\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_{n-1})\|^2\,|\mathcal{F}_{n-1}\right] \\ &\leq 4\gamma_n^2\,\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_{n-1})\|^2\,|\mathcal{F}_{n-1}\right] \leq 8\gamma_n^2\beta_0^2\,\|\theta_{n-1}-\theta_\star\|^2 + 8\gamma_n^2\sigma^2\end{aligned}$$

using the Cauchy-Schwarz inequality and Lemma 3.1. Then, for $\gamma \in (1/2, 1]$,

$$\mathbf{E}\,\|R_n\|^2 \leq 8\gamma_n^2\beta_0^2\,\mathbf{E}\,\|\theta_{n-1}-\theta_\star\|^2 + 8\gamma_n^4\sigma^2 \leq 8\gamma_n^2(\beta_0^2r^2+\sigma^2) \tag{A.5}$$

from Lemma A.1. □

A.3. Proof of Theorem 4.1

We enhance the result by Toulis & Airoldi (2017a, Lemma 2.2 in the Supplement):

Lemma A.2. *Consider positive sequences a_n such that $\sum_{k=1}^{\infty} a_k = A < \infty$, $b_n \downarrow 0$, and $c_n \downarrow 0$, and there is n' such that $c_n/b_n < 1$ for all $n > n'$. Let*

$$\delta_n = \frac{1}{a_n}\left(\frac{a_{n-1}}{b_{n-1}} - \frac{a_n}{b_n}\right), \quad \zeta_n = \frac{c_n}{b_{n-1}}\frac{a_{n-1}}{a_n}$$

and suppose $\delta_n \downarrow \delta \geq 0$ and $\zeta_n \downarrow 0$. Then, for a nonnegative sequence y_n such that

$$y_n \leq \frac{1+c_n}{1+(1+\delta)b_n}y_{n-1} + a_n, \tag{A.6}$$

there holds

$$y_n \leq K_0\frac{a_n}{b_n} + Q_1^n y_0 + Q_{n_0+1}^n[(1+c_1)^{n_0}A + B] \tag{A.7}$$

for every $n = 1, 2, \ldots$, where n_0 is an integer such that $\delta_n + \zeta_n < 1+\delta$ and $c_n < (1+\delta)b_n$ for all $n \geq n_0$, $K_0 = [1+(1+\delta)b_1]/(1+\delta-\delta_{n_0}-\zeta_{n_0})$, $B = K_0 a_{n_0}/b_{n_0}$, and $Q_i^n = \prod_{j=i}^n(1+c_j)/(1+(1+\delta)b_j)$ if $n \geq i$ and $Q_i^n = 1$ otherwise.

Proof. See Sect. B. □

Corollary A.1. *Let $\alpha > \beta > \gamma$ where $\alpha > 1$ and $\gamma \in (0, 1]$. Consider a nonnegative sequence $\{y_n\}$ such that*

$$y_n \leq (1 - \eta n^{-\gamma} + \nu n^{-\beta}) y_{n-1} + a_1 n^{-\alpha} \tag{A.8}$$

with $\eta > 0$, $\nu \geq 0$, and $a_1 \geq 0$. If $\gamma = 1$, assume that $\eta > \alpha - \gamma$. Then, there holds

$$y_n \leq K_1 n^{-(\alpha-\gamma)} + K_2(n) \exp\left(-\tfrac{1}{2}\log(1+\eta)\phi_\gamma(n)\right)(y_0 + D_{n_0}), \tag{A.9}$$

where

$$\phi_\gamma(n) = \begin{cases} (n^{1-\gamma} - 1)/(1-\gamma), & \gamma \in (0,1), \\ \log n, & \gamma = 1, \end{cases} \quad \textit{and} \quad \delta = \begin{cases} 0, & \gamma \in (0,1), \\ 1/(\frac{\eta}{\alpha-\gamma} - 1), & \gamma = 1. \end{cases}$$

and

$$K_1 = K_0 \frac{a_1(1+\delta)}{\eta}, \tag{A.10a}$$

$$K_2(n) = \begin{cases} \exp\left(\nu(1+\eta)\sum_{k=1}^{\infty} k^{-\beta}\right) < \infty, & \beta > 1, \\ \exp\left(\nu(1+\eta)\phi_\beta(n)\right), & \beta \leq 1, \end{cases} \tag{A.10b}$$

$$D_{n_0} = (1+\eta)^{n_0}\left([1+\nu(1+\eta)]^{n_0} A + B\right), \tag{A.10c}$$

$$A = a_1 \sum_{k=1}^{\infty} k^{-\alpha} < \infty, \quad B = K_0(1+\delta)\eta^{-1} a_1 n_0^{-(\alpha-\gamma)}. \tag{A.10d}$$

The n_0 is an integer such that

$$n^\gamma\left([n/(n-1)]^{\alpha-\gamma} - 1\right) + \nu(1+\eta)[n/(n-1)]^{\alpha-\gamma} n^{-\gamma} < \eta \tag{A.11a}$$

$$\nu(1+\eta)n^{-\gamma} < \eta \tag{A.11b}$$

for all $n \geq n_0$. The constant K_0 is given by

$$K_0 = \frac{(1+\eta)/(1+\delta)}{1 - \left(\frac{n_0^{-\gamma}}{\eta}([\frac{n_0}{n_0-1}]^{\alpha-\gamma} - 1) + \frac{\nu(1+\eta)}{\eta} n_0^{-(\beta-\gamma)}[\frac{n_0}{n_0-1}]^{\alpha-\gamma}\right)}.$$

In particular, if $\nu = 0$, then $K_2(n) \equiv 1$.

Proof. See Sect. B. □

Proof of Theorem 4.1. From Proposition 3.1, one can obtain

$$\begin{aligned}
\mathbf{E}\left\|\theta_n - \theta_\star\right\|^2 = {} & \mathbf{E}\left\|\theta_{n-1} - \theta_\star\right\|^2 && \text{(A.12a)} \\
& - 2\gamma_n \mathbf{E}[(\theta_{n-1} - \theta_\star)^T \nabla \ell(Z_n, \theta_{n-1})] && \text{(A.12b)} \\
& + \gamma_n^2 \mathbf{E}\left\|\nabla \ell(Z_n, \theta_{n-1})\right\|^2 && \text{(A.12c)} \\
& + 2\,\mathbf{E}[(\theta_{n-1} - \theta_\star)^T R_n] && \text{(A.12d)} \\
& - 2\,\mathbf{E}[R_n^T \nabla \ell(Z_n, \theta_{n-1})] && \text{(A.12e)} \\
& + \mathbf{E}\left\|R_n\right\|^2. && \text{(A.12f)}
\end{aligned}$$

The second term (A.12b) is upper bounded as follows.

$$\begin{aligned}
& -2\gamma_n \mathbf{E}[(\theta_{n-1} - \theta_\star)^T \nabla \ell(Z_n, \theta_{n-1})] \\
& \quad = -2\gamma_n \mathbf{E}\,\mathbf{E}[(\theta_{n-1} - \theta_\star)^T \nabla \ell(Z_n, \theta_{n-1}) | \mathcal{F}_{n-1}] \\
& \quad = -2\gamma_n \mathbf{E}[(\theta_{n-1} - \theta_\star)^T \mathbf{E}[\nabla \ell(Z_n, \theta_{n-1}) | \mathcal{F}_{n-1}]] \\
& \quad = -2\gamma_n \mathbf{E}[(\theta_{n-1} - \theta_\star)^T \nabla L(\theta_{n-1})] \\
& \quad = -2\gamma_n \mathbf{E}[(\theta_{n-1} - \theta_\star)^T (\nabla L(\theta_{n-1}) - \nabla L(\theta_\star))] \\
& \quad \leq -2\gamma_n \lambda \mathbf{E}\left\|\theta_{n-1} - \theta_\star\right\|^2.
\end{aligned} \tag{A.13}$$

The last inequality is due to Assumption A3 and Remark 2.2; the penultimate inequality is due to $\nabla L(\theta_\star) = 0$.

In order to bound the third term (A.12c), from inequality (A.4) in the proof of Proposition 3.1,

$$\begin{aligned}\gamma_n^2 \,\mathbf{E}\,\|\nabla \ell(Z_n, \theta_{n-1})\|^2 &\leq 2\gamma_n^2\beta_0^2 \,\mathbf{E}\,\|\theta_{n-1} - \theta_\star\|^2 + 2\gamma_n^2\sigma^2 \\ &\leq 2\gamma_n^2(\beta_0^2 r^2 + \sigma^2).\end{aligned}$$

The last line follows from Lemma A.1.

Also for the term (A.12d), from the Cauchy-Schwarz and inequality (6a) in Proposition 3.1 we have

$$\begin{aligned}\mathbf{E}\,[(\theta_{n-1} - \theta_\star)^T R_n | \mathcal{F}_{n-1}] &\leq \mathbf{E}\,[\|\theta_{n-1} - \theta_\star\| \, \|R_n\| \, | \mathcal{F}_{n-1}] \\ &= \|\theta_{n-1} - \theta_\star\| \,\mathbf{E}\,[\|R_n\| \, | \mathcal{F}_{n-1}] \\ &\leq \gamma_n^2\beta_0^2 \,\|\theta_{n-1} - \theta_\star\|^2 + \gamma_n^2(\beta_0^2 + \sigma^2) \,\|\theta_{n-1} - \theta_\star\| \,.\end{aligned}$$

Thus

$$2\,\mathbf{E}\,[(\theta_{n-1} - \theta_\star)^T R_n] \leq 2\gamma_n^2[\beta_0^2 r^2 + (\beta_0^2 + \sigma^2)r]$$

using Lemma A.1 and Jensen's inequality.

For the term (A.12e), we obtain

$$\begin{aligned}-\,\mathbf{E}\,[R_n^T \nabla \ell(Z_n, \theta_{n-1}) | \mathcal{F}_{n-1}] &= -\,\mathbf{E}\,[\gamma_n[\nabla \ell(Z_n, \theta_{n-1}) - \nabla \ell(Z_n, \theta_n)]^T \nabla \ell(Z_n, \theta_{n-1}) | \mathcal{F}_{n-1}] \\ &= -\gamma_n \,\mathbf{E}\,[\|\nabla \ell(Z_n, \theta_{n-1})\|^2 \, | \mathcal{F}_{n-1}] + \gamma_n \,\mathbf{E}\,\nabla \ell(Z_n, \theta_n)^T \nabla \ell(Z_n, \theta_{n-1}) | \mathcal{F}_{n-1}] \\ &\leq -\gamma_n \,\mathbf{E}\,[\|\nabla \ell(Z_n, \theta_{n-1})\|^2 \, | \mathcal{F}_{n-1}] + \gamma_n \,\mathbf{E}\,\|\nabla \ell(Z_n, \theta_n)\| \, \|\nabla \ell(Z_n, \theta_{n-1})\| \, | \mathcal{F}_{n-1}] \\ &\leq -\gamma_n \,\mathbf{E}\,[\|\nabla \ell(Z_n, \theta_{n-1})\|^2 \, | \mathcal{F}_{n-1}] + \gamma_n \,\mathbf{E}\,[\|\nabla \ell(Z_n, \theta_{n-1})\|^2 \, | \mathcal{F}_{n-1}] \\ &= 0,\end{aligned}$$

where the last inequality is due to Lemma 3.1. Thus we have $-2\,\mathbf{E}\,[R_n^T \nabla \ell(Z_n, \theta_{n-1})] \leq 0$.

The final term (A.12) is bounded by inequality (6d) in Proposition 3.1.

Combining these results yields

$$\mathbf{E}\,\|\theta_n - \theta_\star\|^2 \leq (1 - 2\lambda\gamma_n) \,\mathbf{E}\,\|\theta_{n-1} - \theta_\star\|^2 + 2[6\beta_0^2 r^2 + 5\sigma^2 + (\beta_0^2 + \sigma^2)r]\gamma_n^2.$$

Now we can apply Corollary A.1 by setting $y_n = \mathbf{E}\,\|\theta_n - \theta_\star\|^2$, $\alpha = 2\gamma$, $a_1 = 2[6\beta_0^2 r^2 + 5\sigma^2 + (\beta_0^2 + \sigma^2)r]$, γ_1^2, $\eta = 2\lambda\gamma_1 > 0$, and $\nu = 0$. This proves the claim.

Below we put the explicit values of the constants:

$$\begin{aligned}\delta &= \begin{cases} 0, & \gamma \in (1/2, 1), \\ 1/(2\lambda\gamma_1 - 1), & \gamma = 1, \end{cases} \\ K_0 &= \frac{1/(1+\delta) + \gamma_1}{1 - \frac{n_0^\gamma}{2\lambda\gamma_1}([\frac{n_0}{n_0 - 1}]^\gamma - 1)}, \\ K_1 &= K_0 \frac{2[6\beta_0^2 r^2 + 5\sigma^2 + (\beta_0^2 + \sigma^2)r]\gamma_1(1+\delta)}{2\lambda}, \\ K_2(n) &\equiv 1, \\ D_{n_0} &= (1 + 2\lambda\gamma_1)^{n_0}(A + B), \\ A &= 2[6\beta_0^2 r^2 + 5\sigma^2 + (\beta_0^2 + \sigma^2)r]\gamma_1^2 \sum_{k=1}^{\infty} k^{-2\gamma} < \infty, \\ B &= 2K_0[6\beta_0^2 r^2 + 5\sigma^2 + (\beta_0^2 + \sigma^2)r]\gamma_1 n_0^{-\gamma}.\end{aligned}$$

□

A.4. Proof of Theorem 4.2

Proof of Theorem 4.2. Suppose the following holds.

$$\mathbf{E}\left[\nabla \ell(Z_n, \theta_n)\right] = \mathcal{H}(\theta_\star)\,\mathbf{E}\left[\theta_n - \theta_\star\right] + O(\gamma_n^{1/2}). \tag{A.14}$$

Taking expectations on the update equation $\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_n)$, equation (A.14) entails

$$\mathbf{E}\left[\theta_n - \theta_\star\right] = \mathbf{E}\left[\theta_{n-1} - \theta_\star\right] - \gamma_n \mathcal{H}(\theta_\star)\,\mathbf{E}\left[\theta_n - \theta_\star\right] + O(\gamma_n^{1/2}),$$

or

$$\mathbf{E}\left[\theta_n - \theta_\star\right] = [I + \gamma_n \mathcal{H}(\theta_\star)]^{-1}\,\mathbf{E}\left[\theta_{n-1} - \theta_\star\right] + [I + \gamma_n \mathcal{H}(\theta_\star)]^{-1} O(\gamma_n^{1/2}).$$

By recursively applying the above equation, we see

$$\mathbf{E}\left[\theta_n - \theta_\star\right] = Q_1^n(\theta_0 - \theta_\star) + Q_1^n O(\sum_{k=1}^{n} \gamma_k^{1/2}) = Q_1^n(\theta_0 - \theta_\star) + Q_1^n O(n^{1-\gamma/2}).$$

Note $\|Q_1^n\| = \left\|\prod_{i=1}^n [I + \gamma_i \mathcal{H}(\theta_\star)]^{-1}\right\| \le (\prod_{i=1}^n (1 + \lambda\gamma_i))^{-1}$ where $\lambda \le \lambda_{\min}(\mathcal{H}(\theta_\star))$. From inequality (B.5), $Q_1^n = O(\exp(-\kappa n^{1-\gamma}))$ if $\gamma \in (1/2, 1)$ and $Q_1^n = O(n^{-\kappa})$ if $\gamma = 1$, where $\kappa = \log(1 + \lambda\gamma_1) > 0$. In either case, we have $Q_1^n O(n^{1-\gamma/2}) = o(1)$ as desired (when $\gamma = 1$, we are given $\gamma_1 \ge (e^2 - 1)/\lambda$ with which $\kappa \ge 1$).

To see equation (A.14) holds, recall that the objective function L is twice differentiable at $\theta_\star$. It follows that

$$\begin{aligned} \nabla \ell(Z_n, \theta_n) &= W_n + \mathcal{H}(\theta_\star)(\theta_n - \theta_\star) + o(\|\theta_n - \theta_\star\|), \\ &= W_n + \mathcal{H}(\theta_\star)(\theta_n - \theta_\star) + o(\gamma_n^{1/2}), \\ W_n &= \nabla \ell(Z_n, \theta_n) - \nabla L(\theta_n), \end{aligned} \tag{A.15}$$

From Proposition 3.1, we see $\nabla \ell(Z_n, \theta_n) = \nabla \ell(Z_n, \theta_{n-1}) - \gamma_n^{-1} R_n$ and $\mathbf{E}[R_n] = O(\gamma_n^2)$. In other words, $W_n = \nabla \ell(Z_n, \theta_{n-1}) - L(\theta_n) - \gamma_n^{-1} R_n$ and

$$\mathbf{E}[W_n] = \nabla L(\theta_{n-1}) - \nabla L(\theta_n) + O(\gamma_n).$$

The difference of the first two terms are bounded as follows.

$$\begin{aligned} \|\nabla L(\theta_n) - \nabla L(\theta_{n-1})\| &= \|\mathbf{E}[\nabla \ell(Z, \theta_n) - \nabla \ell(Z, \theta_{n-1})]\| \\ &\le \mathbf{E}\,\|\nabla \ell(Z, \theta_n) - \nabla \ell(Z, \theta_{n-1})\| \quad \text{(Jensen's inequality)} \\ &\le \mathbf{E}\,\|\nabla \ell(Z, \theta_n) - \nabla \ell(Z, \theta_\star)\| + \mathbf{E}\,\|\nabla \ell(Z, \theta_{n-1}) - \nabla \ell(Z, \theta_\star)\| \\ &\le \mathbf{E}\left[\beta(Z)\,\|\theta_n - \theta_\star\|\right] + \mathbf{E}\left[\beta(Z)\,\|\theta_{n-1} - \theta_\star\|\right] \quad \text{(Assumption A2)} \\ &\le (\mathbf{E}\left[\beta^2(Z)\right])^{1/2} (\mathbf{E}\,\|\theta_n - \theta_\star\|^2)^{1/2} \\ &\quad + (\mathbf{E}\left[\beta^2(Z)\right])^{1/2} (\mathbf{E}\,\|\theta_{n-1} - \theta_\star\|^2)^{1/2} \quad \text{(Cauchy-Schwarz)} \\ &= O(\gamma_n^{1/2}), \end{aligned}$$

where the last inequality is due to Assumption A2 and Theorem 4.1. Hence $\mathbf{E}[W_n] = O(\gamma_n^{1/2})$. Finally, by taking expectations on both sides of equation (A.15), equation (A.14) follows. □

A.5. Proof of Theorem 4.3

The proof utilizes Fabian's central limit theorem for stochastic approximation Fabian (1968, Theorem 2.2) to show the asymptotic normality (9).

Lemma A.3 (Fabian)**.** *Let U_n, V_n, $T_n \in \mathbb{R}^p$ and Φ_n, $\Gamma_n \in \mathbb{R}^{p \times p}$ satisfy the following equation*

$$U_n = (I - n^{-\alpha}\Gamma_n) U_{n-1} + n^{-(\alpha+\beta)/2} \Phi_n V_n + n^{-\alpha-\beta/2} T_n$$

for $\alpha \in (0,1]$ and $\beta \geq 0$, where $\Gamma_n \to \Gamma \succ 0$, $\Phi_n \to \Phi$, $T_n \to T$ or $\mathbf{E}\|T_n - T\| \to 0$, $\mathbf{E}[V_n|\mathcal{F}_{n-1}] = 0$, $C > \left\|\mathbf{E}[V_n V_n^T - \Sigma|\mathcal{F}_{n-1}\right\|] \to 0$ for some C and $\mathcal{F}_n$ a non-decreasing sequence of σ-fields such that Γ_n, Φ_n, and V_n are $\mathcal{F}_n$-measurable. Suppose $\sigma_{j,r}^2 = \mathbf{E}[I_{\|V_j\|^2 \geq rj^\alpha} \|V_j\|^2]$, and either $\lim_{j\to\infty} \sigma_{j,r}^2 = 0$ for every $r > 0$ or $\alpha = 1$ and $\lim_{n\to\infty} n^{-1} \sum_{j=1}^n \sigma_{j,r}^2 = 0$ for every $r > 0$. Let $\beta_+ = \beta$ if $\alpha = 1$, $\beta_+ = 0$ if $\alpha \neq 1$ and assume $\Gamma \succ (\beta_+/2)I$. Then the asymptotic distribution of $n^{\beta/2}U_n$ is normal with mean $(\Gamma - (\beta_+/2)I)^{-1}T$ and covariance matrix $\mathcal{L}^{-1}_{2\Gamma-\beta_+ I}(\Phi\Sigma\Phi^T)$, where $\mathcal{L}_P : X \mapsto (1/2)(PX + XP)$ is the Lyapunov linear map.

If $P \succ 0$, the inverse linear map $\mathcal{L}_P^{-1}$ is well-defined; if furthermore $C \succ 0$, then $\mathcal{L}_P^{-1}(C) \succ 0$.

Proof of Theorem 4.3. Using Proposition 3.1, the implicit update equation becomes

$$\begin{aligned}
\theta_n &= \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_n) \\
&= \theta_{n-1} - \gamma_n \nabla \ell(Z_n, \theta_{n-1}) + R_n \\
&= \theta_{n-1} - \gamma_n [\nabla L(\theta_{n-1}) + \nabla \ell(Z_n, \theta_{n-1}) - \nabla L(\theta_{n-1})] + O(\gamma_n^2) \\
&= \theta_{n-1} - \gamma_n \nabla L(\theta_{n-1}) + \gamma_n \varepsilon_n + O(\gamma_n^2),
\end{aligned} \tag{A.16}$$

where $\varepsilon_n = \nabla L(\theta_{n-1}) - \nabla \ell(Z_n, \theta_{n-1})$. Since L is twice differentiable at $\theta_\star$, we further have

$$\begin{aligned}
\theta_n - \theta_\star &= \theta_{n-1} - \theta_\star - \gamma_n \mathcal{H}(\theta_\star)(\theta_{n-1} - \theta_\star) + \gamma_n o(\|\theta_{n-1} - \theta_\star\|) + \gamma_n \varepsilon_n + O(\gamma_n^2) \\
&= \theta_{n-1} - \theta_\star - \gamma_n \mathcal{H}(\theta_\star)(\theta_{n-1} - \theta_\star) + \gamma_n o(\gamma_{n-1}^{1/2}) + \gamma_n \varepsilon_n + O(\gamma_n^2) \\
&= \theta_{n-1} - \theta_\star - \gamma_n \mathcal{H}(\theta_\star)(\theta_{n-1} - \theta_\star) + o(\gamma_n^{3/2}) + \gamma_n \varepsilon_n + O(\gamma_n^2) \\
&= [I - \gamma_n \mathcal{H}(\theta_\star)](\theta_{n-1} - \theta_\star) + \gamma_n \varepsilon_n + o(\gamma_n^{3/2}),
\end{aligned} \tag{A.17}$$

where the second line is due to Theorem 4.1 and $\theta_{n-1} \to \theta_\star$ a.s. For the third line, recall that $\gamma_n = \gamma_1 n^{-\gamma}$ (R).

In order to apply Lemma A.3, observe that $\mathbf{E}[\varepsilon_n|\mathcal{F}_{n-1}] = 0$ and

$$\begin{aligned}
\mathbf{E}[\varepsilon_n \varepsilon_n^T|\mathcal{F}_{n-1}] &= \mathbf{E}[(\nabla L(\theta_{n-1}) - \nabla \ell(Z_n, \theta_{n-1}))(\nabla L(\theta_{n-1}) - \nabla \ell(Z_n, \theta_{n-1}))^T|\mathcal{F}_{n-1}] \\
&= \mathbf{E}[(\nabla L(\theta_{n-1}) - \nabla \ell(Z, \theta_{n-1}))(\nabla L(\theta_{n-1}) - \nabla \ell(Z, \theta_{n-1}))^T] \\
&= \mathcal{I}(\theta_{n-1})
\end{aligned}$$

since $\theta_{n-1} \in \mathcal{F}_{n-1}$. To meet the conditions for Lemma A.3, it suffices to show that $\mathcal{I}$ is continuous at $\theta_\star$. Consider a non-random convergent sequence $\{\vartheta_n\}$ such that $\vartheta_n \to \theta_\star$. Fix $\epsilon > 0$. Then there exists n_0 such that for all $n \geq n_0$, $\|\vartheta_n - \theta_\star\| \leq \epsilon$. Then,

$$\begin{aligned}
\left\|\nabla \ell(Z, \vartheta_n) \nabla \ell(Z, \vartheta_n)^T\right\| &\leq \|\nabla \ell(Z, \vartheta_n)\|^2 \\
&\leq 2\beta^2(Z) \|\vartheta_n - \vartheta_\star\|^2 + 2 \|\nabla \ell(Z, \theta_\star)\|^2 \\
&\leq 2\beta^2(Z)\epsilon^2 + 2 \|\nabla \ell(Z, \theta_\star)\|^2
\end{aligned}$$

for all $n \geq n_0$. The last line is integrable due to Assumptions A2 and A4. Therefore, by the dominated convergence theorem, $\mathcal{I}(\vartheta_n) = \mathbf{E}\left[\nabla \ell(Z, \vartheta_n) \nabla \ell(Z, \vartheta_n)^T\right] - \nabla L(\vartheta_n) L(\vartheta_n)^T < \infty$ and $\mathcal{I}(\vartheta_n) \to \mathcal{I}(\theta_\star)$ and $\mathcal{I}$ is continuous at $\theta_\star$.

Letting $U_n = \theta_n - \theta_\star$, $V_n = \varepsilon_n$, $T_n = o(1)$, $\Phi_n = \Phi = \gamma_1 I$, $\Gamma_n = \Gamma = \gamma_1 \mathcal{H}(\theta_\star)$, $T = 0$, $\Sigma = \mathcal{I}(\theta_\star)$, and $\alpha = \beta = \gamma$ in Lemma A.3 results in the desired asymptotic normality. □

Remark A.1. *The approximation of ISGD to SGD in Proposition 3.1 is crucial since otherwise ε_n would equal to $\nabla L(\theta_n) - \nabla \ell(Z_n, \theta_n)$. Since $\theta_n \notin \mathcal{F}_{n-1}$, we do not have $\mathbf{E}[\varepsilon_n|\mathcal{F}_{n-1}] = 0$. That is, ε_n is not a martingale difference sequence and it is difficult to see if $\mathbf{E}[\varepsilon_n \varepsilon_n^T|\mathcal{F}_{n-1}]$ would converge to a known quantity. Toulis et al. (Toulis & Airoldi, 2017a;b) instead employ $\nabla \ell(Z_n, \theta_\star)$ in place of ε_n, but assume $\nabla^2 \ell(Z_n, \theta_\star)$ converges to $\mathcal{H}(\theta_\star) = \nabla^2 L(\theta_\star)$ almost surely, which rarely holds in general.*

A.6. Proof of Theorem 4.4

The following result can be deduced from the proof of Chen et al. (2020, Lemma 4.1).

Lemma A.4. *Suppose Assumptions A1–A4, B1, and B3–B5 hold. Then,*

$$\mathbf{E}\left\|\hat{H}_n - \mathcal{H}(\theta_\star)\right\| = O(\gamma_n^{1/2}), \quad \textit{and} \quad \mathbf{E}\left\|\hat{I}_n - \mathcal{I}(\theta_\star)\right\| = O(\gamma_n^{1/2}).$$

A key in the proof of Chen et al. (2020, Lemma 4.1) is Chen et al. (2020, Lemma 3.2) showing that $\mathbf{E}\,\|\theta_n - \theta_\star\| = O(n^{-\gamma/2}) = O(\gamma_n^{1/2})$ in (explicit) SGD, which can be replaced by Theorem 4.1 in implicit SGD.

Proof of Theorem 4.4. Let $B = \gamma_1\mathcal{H}(\theta_\star) - (\beta_+/2)I$, $\tilde{B}_n = \gamma_1\tilde{H}_n - (\beta_+/2)I$, and $\hat{B}_n = \gamma_1\hat{H}_n - (\beta_+/2)I$, where $\beta_+ = 1$ if $\gamma = 1$ and $\beta_+ = 0$ if $\gamma \in (0.5, 1)$. By construction, $\lambda_{\min}(\tilde{B}_n) \geq \gamma_1\delta - \beta_+/2 > 0$ and $\lambda_{\min}(B) \geq \gamma_1\lambda_{\min}(\mathcal{H}(\theta_\star)) - \beta_+/2 > 0$.

Recall that $\operatorname{vec}\left(\mathcal{L}_{2\hat{B}_n}^{-1}(\hat{I}_n)\right) = (I\otimes\hat{B}_n + \hat{B}_n\otimes I)^{-1}\operatorname{vec}(\hat{I}_n)$ and $\operatorname{vec}\left(\mathcal{L}_{2B}^{-1}(\mathcal{I}(\theta_\star)\right) = (I\otimes B + B\otimes I)^{-1}\operatorname{vec}(\mathcal{I}(\theta_\star)$. It suffices to show that

$$\mathbf{E}\left\|\operatorname{vec}\left(\mathcal{L}_{2\hat{B}_n}^{-1}(\hat{I}_n)\right) - \operatorname{vec}\left(\mathcal{L}_{2B}^{-1}(\mathcal{I}(\theta_\star)\right)\right\| = O(\gamma_n^{1/2}). \tag{A.18}$$

To see this, let

$$\begin{aligned} E_B &= (I\otimes\tilde{B}_n + \tilde{B}_n\otimes I) - (I\otimes B + B\otimes I), \quad E_I = \hat{I}_n - \mathcal{I}(\theta_\star), \\ F_B &= (I\otimes\tilde{B}_n + \tilde{B}_n\otimes I)^{-1} - (I\otimes B + B\otimes I)^{-1}. \end{aligned}$$

Then,

$$\begin{aligned} &\operatorname{vec}\left(\mathcal{L}_{2\hat{B}_n}^{-1}(\hat{I}_n)\right) - \operatorname{vec}\left(\mathcal{L}_{2B}^{-1}(\mathcal{I}(\theta_\star)\right) = \\ &\quad [(I\otimes B + B\otimes I)^{-1} + F_B](\operatorname{vec}(\mathcal{I}(\theta_\star) + E_I) - (I\otimes B + B\otimes I)^{-1}\operatorname{vec}(\mathcal{I}(\theta_\star) \\ &\quad = (I\otimes B + B\otimes I)^{-1}E_I + F_BE_I + F_B\operatorname{vec}(\mathcal{I}(\theta_\star)). \end{aligned}$$

Since the eigenvalues of $I\otimes A + A\otimes I$ consist of $\lambda_i(A) + \lambda_j(A)$ for $i, j = 1, \ldots, p$ if A is a $p\times p$ symmetric matrix,

$$\|F_B\| \leq \left\|(I\otimes\tilde{B}_n + \tilde{B}_n\otimes I)^{-1}\right\| + \left\|(I\otimes B + B\otimes I)^{-1}\right\| \leq \frac{1}{2\gamma_1\delta - \beta_+} + \frac{1}{2\lambda_{\min}(B)}.$$

Therefore, by Lemma A.4,

$$\mathbf{E}\left\|(I\otimes B + B\otimes I)^{-1}E_I + F_BE_I\right\| \leq \left(\frac{1}{2\gamma_1\delta - \beta_+} + \frac{1}{\lambda_{\min}(B)}\right)\mathbf{E}\,\|E_I\| = O(\gamma_n^{1/2}). \tag{A.19}$$

In order to bound $F_B\operatorname{vec}(\mathcal{I}(\theta_\star))$, recall from Chen et al. (2020, Lemma C.1) that

$$\|F_B\| \leq \|E_B\|\left\|(I\otimes B + B\otimes I)^{-1}\right\|^2 \leq \frac{1}{4\lambda_{\min}^2(B)}\|E_B\|$$

under the event $\left\|(I\otimes B + B\otimes I)^{-1}E_B\right\| < 1/2$. Thus

$$\begin{aligned} \mathbf{E}\,\|F_B\| &\leq \frac{1}{4\lambda_{\min}^2(B)}\mathbf{E}\,\|E_B\|\,P(\left\|(I\otimes B + B\otimes I)^{-1}E_B\right\| < 1/2) \\ &\quad + \Big(\frac{1}{2\gamma_1\delta - \beta_+} + \frac{1}{2\lambda_{\min}(B)}\Big)P(\left\|(I\otimes B + B\otimes I)^{-1}E_B\right\| \geq 1/2) \\ &\leq \frac{1}{4\lambda_{\min}^2(B)}\mathbf{E}\,\|E_B\| + \Big(\frac{1}{2\gamma_1\delta - \beta_+} + \frac{1}{2\lambda_{\min}(B)}\Big)\frac{1}{\lambda_{\min}(B)}\mathbf{E}\,\|E_B\|, \end{aligned}$$

where the last line is due to Markov's inequality

$$P(\left\|(I\otimes B + B\otimes I)^{-1}E_B\right\| \geq 1/2) \leq 2\left\|(I\otimes B + B\otimes I)^{-1}\right\|\mathbf{E}\,\|E_B\|.$$

It remains to bound $\mathbf{E}\,\|E_B\|$. If $\hat{H}_n \succeq \delta I$, then $\tilde{H}_n = \hat{H}_n$. Otherwise, $\lambda_{\min}(\hat{H}_n) < \delta$ and

$$\begin{aligned}\|E_B\| &= \left\|(I\otimes\tilde{B}_n + \tilde{B}_n\otimes I) - (I\otimes B + B\otimes I)\right\| \\ &\le \left\|(I\otimes\tilde{B}_n + \tilde{B}_n\otimes I) - (I\otimes\hat{B}_n + \hat{B}_n\otimes I)\right\| \\ &\quad + \left\|(I\otimes\hat{B}_n + \hat{B}_n\otimes I) - (I\otimes B + B\otimes I)\right\| \\ &\le 2\delta + \left\|(I\otimes\hat{B}_n + \hat{B}_n\otimes I) - (I\otimes B + B\otimes I)\right\|.\end{aligned}$$

Thus, by using Lemma A.4,

$$\begin{aligned}\mathbf{E}\,\|E_B\| &\le \mathbf{E}\left\|(I\otimes\hat{B}_n + \hat{B}_n\otimes I) - (I\otimes B + B\otimes I)\right\| + 2\delta P(\lambda_{\min}(\hat{H}_n) < \delta) \\ &\le \mathbf{E}\left\|I\otimes(\hat{B}_n - B) + (\hat{B}_n - B)\otimes I\right\| \\ &\quad + 2\delta P\left(\left\|\hat{H}_n - \mathcal{H}(\theta_\star)\right\| \ge \lambda_{\min}(\mathcal{H}(\theta_\star)) - \delta\right) \\ &\le 2\,\mathbf{E}\left\|\hat{H}_n - \mathcal{H}(\theta_\star)\right\| + \frac{2\delta}{\lambda_{\min}(\mathcal{H}(\theta_\star)) - \delta}\,\mathbf{E}\left\|\hat{H}_n - \mathcal{H}(\theta_\star)\right\|,\end{aligned}$$

where the second line is due to Weyl's inequalty

$$\lambda_{\min}(\mathcal{H}(\theta_\star)) \le \lambda_{\max}(\mathcal{H}(\theta_\star) - \hat{H}_n) + \lambda_{\min}(\hat{H}_n) \le \left\|\mathcal{H}(\theta_\star) - \hat{H}_n\right\| + \delta$$

and the third line is Markov's inequality. Hence,

$$\begin{aligned}\mathbf{E}\,\|F_B\,\mathrm{vec}\,\mathcal{I}(\theta_\star)\| \le \|\mathcal{I}(\theta_\star)\|_F &\left(\frac{1}{4\lambda_{\min}^2(B)} + \left(\frac{1}{2\gamma_1\delta - \beta_+} + \frac{1}{2\lambda_{\min}(B)}\right)\frac{1}{\lambda_{\min}(B)}\right)\times \\ &\left(2 + \frac{2\delta}{\lambda_{\min}(\mathcal{H}(\theta_\star)) - \delta}\right)\mathbf{E}\left\|\hat{H}_n - \mathcal{H}(\theta_\star)\right\| = O(\gamma_n^{1/2})\end{aligned} \tag{A.20}$$

by Lemma A.4.

Combining inequalities (A.19) and (A.20), we obtain inequality (A.18). □

A.7. Proof of Theorem 4.5

Lemma A.5. *Under Assumptions A1, A2′, A3, A4′, B4, and B1, if $\gamma \in (1/3, 1]$, it follows*

$$\begin{aligned}\mathbf{E}\,\|\theta_n - \theta_\star\|^4 \le \tilde{K}_1 n^{-2\gamma} + \exp&\left(\nu(1+\lambda\gamma_1/2)\phi_{\frac{5}{3}\gamma}(n) - \tfrac{1}{2}\log(1+\lambda\gamma_1/2)\phi_\gamma(n)\right) \\ &\times\left(\|\theta_0 - \theta_\star\|^4 + \tfrac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\|\theta_0 - \theta_\star\|^3 + \tilde{D}_{\tilde{n}_0}\right)\end{aligned}$$

where

$$\nu = \tfrac{6(\beta_0^2+\sigma^2)}{\lambda}\beta_0^{2/3}\gamma_1^{5/3} + 14\beta_0^2\gamma_1^2 + 16\beta_0^3\gamma_1^3 + 8\beta_0^4\gamma_1^4 + 2K_2[(10\sigma^2 + \tfrac{32\sigma(\beta_0^2+\sigma^2)}{\lambda})\gamma_1^2 + \tfrac{16(\beta_0^2+\sigma^2)^2}{\lambda}\gamma_1^3 + 8\gamma_1^4],$$

and

$$K_2 = (K_1 + \|\theta_0 - \theta_\star\|^2 + D_{n_0})/\gamma_1$$

with K_1 and D_{n_0} as defined in Theorem 4.1. Here, $\tilde{n}_0$ is an integer such that

$$n^\gamma\left([n/(n-1)]^{2\gamma} - 1\right) + \nu(1+\eta)[n/(n-1)]^{2\gamma}n^{-\gamma} < \eta \text{ and } \nu(1+\lambda\gamma_1/2)n^{-\gamma} < \eta \tag{A.21}$$

for all $n \ge \tilde{n}_0$, and

$$\tilde{K}_1 = \frac{1+\lambda\gamma_1/2}{1 - \left(\frac{\tilde{n}_0^{-\gamma}}{\eta}([\frac{\tilde{n}_0}{\tilde{n}_0-1}]^{2\gamma} - 1) + \frac{\nu(1+\lambda\gamma_1/2)}{\lambda\gamma_1/2}\tilde{n}_0^{-2\gamma/3}[\frac{\tilde{n}_0}{\tilde{n}_0-1}]^{2\gamma}\right)}\frac{C_1\gamma_1^3}{\lambda\gamma_1/2}, \tag{A.22a}$$

$$C_1 = \tfrac{2(\beta_0^2+\sigma^2)}{\lambda}\beta_0^2 + (16\sigma^4 + \tfrac{56(\beta_0^2+\sigma^2)}{\lambda}\sigma^3)\gamma_1 \tag{A.22b}$$

$$\tilde{D}_{\tilde{n}_0} = (1+\lambda\gamma_1/2)^{\tilde{n}_0}\big([1+\nu(1+\lambda\gamma_1/2)]^{\tilde{n}_0}\tilde{A} + \tilde{B}\big), \tag{A.22c}$$

$$\tilde{A} = C_1\gamma_1^3\sum_{k=1}^{\infty} k^{-3\gamma} < \infty, \tag{A.22d}$$

$$\tilde{B} = \frac{1+\lambda\gamma_1/2}{1-\left(\frac{\tilde{n}_0^{-\gamma}}{\eta}([\frac{\tilde{n}_0}{\tilde{n}_0-1}]^{2\gamma}-1) + \frac{\nu(1+\lambda\gamma_1/2)}{\lambda\gamma_1/2}\tilde{n}_0^{-2\gamma/3}[\frac{n_0}{\tilde{n}_0-1}]^{2\gamma}\right)}\frac{C_1\gamma_1^3}{\lambda\gamma_1/2}\tilde{n}_0^{-2\gamma}. \tag{A.22e}$$

Proof of Theorem 4.5. We follow the decomposition by Bach & Moulines (2011, section C) for explicit SGD. The only difference is that

$$\nabla\ell(Z_n,\theta_n) = \frac{1}{\gamma_n}(\theta_{n-1}-\theta_n)$$

instead of $\nabla\ell(Z_n,\theta_{n-1}) = \frac{1}{\gamma_n}(\theta_{n-1}-\theta_n)$, which leads to

$$\begin{aligned}\|\nabla\ell(Z_n,\theta_{n-1})\| &\le \|\nabla\ell(Z_n,\theta_{n-1}) - \nabla\ell(Z_n,\theta_n)\| + \|\nabla\ell(Z_n,\theta_n)\| \\ &\le \beta_0\|\theta_{n-1}-\theta_n\| + \frac{1}{\gamma_n}\|\theta_{n-1}-\theta_n\| = \frac{\beta_0\gamma_n+1}{\gamma_n}\|\theta_{n-1}-\theta_n\| \\ &\le \frac{\beta_0\gamma_1+1}{\gamma_n}\|\theta_{n-1}-\theta_n\|\end{aligned}$$

and

$$\begin{aligned}\left\|\frac{1}{n}\sum_{k=1}^{n}\nabla\ell(Z_k,\theta_{k-1})\right\| \le{}& \frac{\beta_0\gamma_1+1}{n}\sum_{k=1}^{n-1}|\gamma_{k+1}^{-1}-\gamma_k^{-1}|\,\|\theta_k-\theta_\star\| \\ &+ \frac{\beta_0\gamma_1+1}{n\gamma_n}\|\theta_n-\theta_\star\| + \frac{\beta_0\gamma_1+1}{n\gamma_1}\|\theta_0-\theta_\star\|.\end{aligned}$$

This in turn yields

$$\begin{aligned}(\mathbf{E}\left\|\bar{\theta}_n-\theta_\star\right\|^2)^{1/2} \le{}& \frac{1}{\sqrt{n}}\left[\mathrm{tr}\left(\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right)\right]^{1/2} \\ &+ \frac{\beta_0\gamma_1+1}{\lambda^{1/2}n}\sum_{k=1}^{n-1}|\gamma_{k+1}^{-1}-\gamma_k^{-1}|(\mathbf{E}\|\theta_k-\theta_\star\|^2)^{1/2} \\ &+ \frac{\beta_0\gamma_1+1}{\lambda^{1/2}n\gamma_n}(\mathbf{E}\|\theta_n-\theta_\star\|^2)^{1/2} + \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1 n}\|\theta_0-\theta_\star\| \\ &+ \frac{M}{2\lambda^{1/2}n}\sum_{k=1}^{n}(\mathbf{E}\|\theta_k-\theta_\star\|^4)^{1/2} + \frac{2\beta_0}{\lambda^{1/2}n}(\sum_{k=0}^{n-1}\mathbf{E}\|\theta_k-\theta_\star\|^2)^{1/2},\end{aligned}$$

which corresponds to inequality (26) of (Bach & Moulines, 2011). The above bound can be simplified to

$$\begin{aligned}(\mathbf{E}\left\|\bar{\theta}_n-\theta_\star\right\|^2)^{1/2} \le{}& \frac{1}{\sqrt{n}}\left[\mathrm{tr}\left(\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right)\right]^{1/2} + \frac{1/\gamma_1+3\beta_0}{\lambda^{1/2}n}\|\theta_0-\theta_\star\| \\ &+ \frac{\beta_0\gamma_1+1}{\lambda^{1/2}n\gamma_n}(\mathbf{E}\|\theta_n-\theta_\star\|^2)^{1/2} && \text{(A.23a)} \\ &+ \frac{2\beta_0}{\lambda^{1/2}n}(\sum_{k=1}^{n}\mathbf{E}\|\theta_k-\theta_\star\|^2)^{1/2} && \text{(A.23b)} \\ &+ \frac{\beta_0\gamma_1+1}{\lambda^{1/2}n}\sum_{k=1}^{n-1}|\gamma_{k+1}^{-1}-\gamma_k^{-1}|(\mathbf{E}\|\theta_k-\theta_\star\|^2)^{1/2} && \text{(A.23c)} \\ &+ \frac{M}{2\lambda^{1/2}n}\sum_{k=1}^{n}(\mathbf{E}\|\theta_k-\theta_\star\|^4)^{1/2} && \text{(A.23d)}\end{aligned}$$

The terms (A.23a)–(A.23c) can be further bounded using Theorem 4.1 and Minkowski's inequality:

$$\begin{aligned}
\text{(A.23a)} &\le \frac{(\beta_0\gamma_1+1)K_1^{1/2}}{\lambda^{1/2}\gamma_1 n^{1-\gamma/2}} \\
&\quad + \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1 n^{1-\gamma}} \exp\left(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\right)\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)^{1/2} \\
&\le \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1}\frac{K_1^{1/2}}{n^{1-\gamma/2}} + \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1}\exp\left(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\right)\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)^{1/2}, \\
\text{(A.23b)} &\le \frac{2\beta_0}{\lambda^{1/2}n}\left[K_1\sum_{k=1}^n k^{-\gamma} + \sum_{k=1}^n \exp\left(-\tfrac{1}{2}\log(1+2\lambda\gamma_1)\phi_\gamma(k)\right)\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)\right]^{1/2} \\
&\le \frac{2\beta_0 K_1^{1/2}}{\lambda^{1/2}}\frac{\phi_\gamma^{1/2}(n)}{n} + \frac{2\beta_0}{\lambda^{1/2}}\frac{\mathcal{A}^{1/2}}{n}\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)^{1/2}, \\
\text{(A.23c)} &\le \frac{\gamma(\beta_0+1/\gamma_1)}{\lambda^{1/2}n}\left[K_1^{1/2}\sum_{k=1}^n k^{\gamma/2-1}\right. \\
&\quad \left. + \sum_{k=1}^n k^{\gamma-1}\exp\left(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(k)\right)\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)^{1/2}\right] \\
&\le \frac{\gamma K_1^{1/2}(\beta_0+1/\gamma_1)}{\lambda^{1/2}}\frac{\phi_{1-\gamma/2}(n)}{n} + \frac{\gamma(\beta_0+1/\gamma_1)}{\lambda^{1/2}}\frac{\mathcal{A}}{n}\left(\|\theta_0-\theta_\star\|^2+D_{n_0}\right)^{1/2},
\end{aligned}$$

where

$$\mathcal{A} = \textstyle\sum_{k=1}^\infty \exp\left(-\frac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(k)\right) < \infty.$$

The first inequality for (A.23c) is due to $|\gamma_{k+1}^{-1}-\gamma_k^{-1}| = \frac{1}{\gamma_1}[(k+1)^\gamma - k^\gamma] = \frac{k^\gamma}{\gamma_1}[(1+1/k)^\gamma - 1] \le \frac{\gamma}{\gamma_1}k^{\gamma-1}$.

In order to bound the term (A.23d), from Lemma A.5,

$$\begin{aligned}
\left(\mathbf{E}\left\|\theta_n-\theta_\star\right\|^4\right)^{1/2} &\le \tilde{K}_1^{1/2}n^{-\gamma} + \exp\left(\tfrac{1}{2}\nu(1+\lambda\gamma_1/2)\phi_{\frac{5}{3}\gamma}(n) - \tfrac{1}{4}\log(1+\lambda\gamma_1/2)\phi_\gamma(n)\right) \\
&\quad \times \left(\left\|\theta_0-\theta_\star\right\|^4 + \frac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\left\|\theta_0-\theta_\star\right\|^3 + \tilde{D}_{n_0}\right)^{1/2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\text{(A.23d)} &\le \frac{M}{2\lambda^{1/2}n}\sum_{k=1}^n\frac{\tilde{K}_1^{1/2}}{k^\gamma} + \frac{M}{2\lambda^{1/2}n}\left(\left\|\theta_0-\theta_\star\right\|^4 + \frac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\left\|\theta_0-\theta_\star\right\|^3 + D_{n_0}\right)^{1/2} \\
&\quad \times \sum_{k=1}^n \exp\left(\tfrac{1}{2}\nu(1+\lambda\gamma_1/2)\phi_{\frac{5}{3}\gamma}(k) - \tfrac{1}{4}\log(1+\lambda\gamma_1/2)\phi_\gamma(k)\right) \\
&\le \frac{M\tilde{K}_1^{1/2}}{2\lambda^{1/2}}\frac{\phi_\gamma(n)}{n} + \frac{M}{2\lambda^{1/2}}\frac{\mathcal{B}}{n}\left(\left\|\theta_0-\theta_\star\right\|^4 + \frac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\left\|\theta_0-\theta_\star\right\|^3 + D_{n_0}\right)^{1/2},
\end{aligned}$$

where

$$\mathcal{B} = \textstyle\sum_{k=1}^\infty \exp\left(\frac{1}{2}\nu(1+\lambda\gamma_1/2)\phi_{\frac{5}{3}\gamma}(k) - \frac{1}{4}\log(1+\lambda\gamma_1/2)\phi_\gamma(k)\right) < \infty.$$

Combining these bounds, we get

$$\begin{aligned}
\left(\mathbf{E}\left\|\bar{\theta}_n-\theta_\star\right\|^2\right)^{1/2} &\le \frac{1}{\sqrt{n}}\left[\operatorname{tr}\left(\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right)\right]^{1/2} + \frac{1/\gamma_1+3\beta_0}{\lambda^{1/2}}\frac{\left\|\theta_0-\theta_\star\right\|}{n} \\
&\quad + \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1}\frac{\tilde{K}_1^{1/2}}{n^{1-\gamma/2}} + \frac{\beta_0\gamma_1+1}{\lambda^{1/2}\gamma_1}\exp\left(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\right)\left(\left\|\theta_0-\theta_\star\right\|^2+\tilde{D}_{n_0}\right)^{1/2} \\
&\quad + \frac{2\beta_0\tilde{K}_1^{1/2}}{\lambda^{1/2}}\frac{\phi_\gamma^{1/2}(n)}{n} + \frac{2\beta_0}{\lambda^{1/2}}\frac{\mathcal{A}^{1/2}}{n}\left(\left\|\theta_0-\theta_\star\right\|^2+\tilde{D}_{n_0}\right)^{1/2}
\end{aligned}$$

$$+\frac{\gamma\tilde{K}_1^{1/2}(\beta_0+1/\gamma_1)}{\lambda^{1/2}}\frac{\phi_{1-\gamma/2}(n)}{n}+\frac{\gamma(\beta_0+1/\gamma_1)}{\lambda^{1/2}}\frac{\mathcal{A}}{n}(\|\theta_0-\theta_\star\|^2+\tilde{D}_{n_0})^{1/2}$$
$$+\frac{M\tilde{K}_1^{1/2}}{2\lambda^{1/2}}\frac{\phi_\gamma(n)}{n}+\frac{M}{2\lambda^{1/2}}\frac{\mathcal{B}}{n}(\|\theta_0-\theta_\star\|^4+\frac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\|\theta_0-\theta_\star\|^3+\tilde{D}_{n_0})^{1/2},$$

which further simplifies to

$$\begin{aligned}(\mathbf{E}\left\|\bar{\theta}_n-\theta_\star\right\|^2)^{1/2}\le{}&\frac{1}{\sqrt{n}}\left[\mathrm{tr}\left(\mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}\right)\right]^{1/2}\\
&+\frac{\tilde{K}_1^{1/2}}{\lambda^{1/2}n}\left((\beta_0+\gamma_1^{-1})n^{\gamma/2}+2\beta_0\phi_\gamma^{1/2}(n)+\gamma(\beta_0+\gamma_1^{-1})\phi_{1-\gamma/2}(n)+\frac{M}{2}\phi_\gamma(n)\right)\\
&+\frac{1}{\lambda^{1/2}n}\Big((\gamma_1^{-1}+3\beta_0)\|\theta_0-\theta_\star\|+2\beta_0\mathcal{A}^{1/2}(\|\theta_0-\theta_\star\|^2+\tilde{D}_{n_0})^{1/2}\\
&\qquad+\gamma(\beta_0+\gamma_1^{-1})\mathcal{A}(\|\theta_0-\theta_\star\|^2+\tilde{D}_{n_0})^{1/2}\Big)\\
&+\frac{M\mathcal{B}}{2\lambda^{1/2}n}(\|\theta_0-\theta_\star\|^4+\tfrac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\|\theta_0-\theta_\star\|^3+\tilde{D}_{n_0})^{1/2}\\
&+\frac{\beta_0+\gamma_1^{-1}}{\lambda^{1/2}}\exp\left(-\tfrac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(n)\right)(\|\theta_0-\theta_\star\|^2+\tilde{D}_{n_0})^{1/2}.\end{aligned}$$

Below we put the explicit values of the constants:

$$\begin{aligned}\tilde{\mathcal{A}}&=(\gamma_1^{-1}+3\beta_0)\|\theta_0-\theta_\star\|+2\beta_0\mathcal{A}^{1/2}(\|\theta_0-\theta_\star\|^2+D_{n_0})^{1/2}+\gamma(\beta_0+\gamma_1^{-1})\mathcal{A}(\|\theta_0-\theta_\star\|^2+D_{n_0})^{1/2},\\
\tilde{\mathcal{B}}&=\mathcal{B}\cdot(\|\theta_0-\theta_\star\|^4+\tfrac{8(\beta_0^2+\sigma^2)}{\lambda}\gamma_1\|\theta_0-\theta_\star\|^3+D_{n_0})^{1/2},\\
\mathcal{A}&=\textstyle\sum_{k=1}^\infty\exp\big(-\frac{1}{4}\log(1+2\lambda\gamma_1)\phi_\gamma(k)\big),\\
\mathcal{B}&=\textstyle\sum_{k=1}^\infty\exp\left(\frac{1}{2}\nu(1+\lambda\gamma_1/2)\phi_{\frac{5}{3}\gamma}(k)-\frac{1}{4}\log(1+\lambda\gamma_1/2)\phi_\gamma(k)\right),\\
\nu&=\tfrac{6(\beta_0^2+\sigma^2)}{\lambda}\beta_0^{2/3}\gamma_1^{5/3}+14\beta_0^2\gamma_1^2+16\beta_0^3\gamma_1^3+8\beta_0^4\gamma_1^4+2K_2[(10\sigma^2+\tfrac{32\sigma(\beta_0^2+\sigma^2)}{\lambda})\gamma_1^2+\tfrac{16(\beta_0^2+\sigma^2)^2}{\lambda}\gamma_1^3+8\gamma_1^4],\\
K_2&=(K_1+\|\theta_0-\theta_\star\|^2+D_{n_0})/\gamma_1.\end{aligned}$$

□

A.8. Proof of Theorem 5.1

Proof of Theorem 5.1. Pick a minimizer $\theta_\star$ of L. From Assumption A2′, ∇L is also β_0-Lipschitz continuous. Therefore,

$$\begin{aligned}L(\theta_n)-L(\theta_\star)&\le L(\theta_{n-1})-L(\theta_\star)+\nabla L(\theta_{n-1})^T(\theta_n-\theta_{n-1})+\frac{\beta_0}{2}\|\theta_n-\theta_{n-1}\|^2\\
&\le L(\theta_{n-1})-L(\theta_\star)+\nabla L(\theta_{n-1})^T(\theta_n-\theta_{n-1})+\frac{\beta_0\gamma_n^2}{2}\|\nabla\ell(Z_n,\theta_n)\|^2,\end{aligned}\tag{A.24}$$

where the last inequality is due to equation (3′).

In order to find a recursive relation for the sequence $\Delta_n\triangleq\mathbf{E}\left[L(\theta_n)-L(\theta_\star)\right]$, observe that

$$\begin{aligned}\|\nabla\ell(Z_n,\theta_n)\|^2&=\|\nabla\ell(Z_n,\theta_n)-\nabla\ell(Z_n,\theta_\star)+\nabla\ell(Z_n,\theta_\star)\|^2\\
&\le 2\|\nabla\ell(Z_n,\theta_n)-\nabla\ell(Z_n,\theta_\star)\|^2+2\|\nabla\ell(Z_n,\theta_\star)\|^2\\
&\le 2\beta_0^2\|\theta_n-\theta_\star\|^2+2\|\nabla\ell(Z_n,\theta_\star)\|^2.\end{aligned}$$

Let $r_n^2=\delta_0+\sigma^2\sum_{k=1}^n\gamma_k^2=\|\theta_0-\theta_\star\|^2+\sigma^2\sum_{k=1}^n\gamma_k^2$. Since $\mathbf{E}\|\theta_n-\theta_\star\|^2\le r_n^2$ due to Lemma A.1, we see

$$\mathbf{E}\|\nabla\ell(Z_n,\theta_n)\|^2\le 2\beta_0^2\,\mathbf{E}\|\theta_n-\theta_\star\|^2+2\,\mathbf{E}\|\nabla\ell(Z_n,\theta_\star)\|^2\le 2\beta_0^2r_n^2+2\sigma^2.\tag{A.25}$$

In addition, from equation (3′) and Lemma 3.1,

$$\nabla L(\theta_{n-1})^T(\theta_n-\theta_{n-1})=-\gamma_n\nabla L(\theta_{n-1})^T\nabla\ell(Z_n,\theta_n)$$

$$
\begin{aligned}
&= -\gamma_n \nabla L(\theta_{n-1})^T \nabla \ell(Z_n, \theta_{n-1}) \\
&\quad + \nabla L(\theta_{n-1})^T (\gamma_n [\nabla \ell(Z_n, \theta_{n-1}) + \nabla \ell(Z_n, \theta_n)]) \\
&\leq -\gamma_n \nabla L(\theta_{n-1})^T \nabla \ell(Z_n, \theta_{n-1}) + \|\nabla L(\theta_{n-1})\| \, \|R_n\| ,
\end{aligned}
$$

which leads to

$$
\begin{aligned}
\mathbf{E}\left[\nabla L(\theta_{n-1})^T (\theta_n - \theta_{n-1}) | \mathcal{F}_{n-1}\right] &\leq -\gamma_n \nabla L(\theta_{n-1})^T \nabla L(\theta_{n-1}) + \|\nabla L(\theta_{n-1})\| \, \mathbf{E}\left[\|R_n\| \, | \mathcal{F}_{n-1}\right] \\
&\leq -\gamma_n \left\|\nabla L(\theta_{n-1})\right\|^2 \\
&\quad + \|\nabla L(\theta_{n-1}) - \nabla L(\theta_\star)\| \left(\gamma_n^2 \beta_0^2 \left\|\theta_{n-1} - \theta_\star\right\| + \frac{\beta_0^2 + \sigma^2}{2} \gamma_n^2\right) \\
&\leq -\gamma_n \left\|\nabla L(\theta_{n-1})\right\|^2 \\
&\quad + \beta_0 \left\|\theta_{n-1} - \theta_\star\right\| \left(\gamma_n^2 \beta_0^2 \left\|\theta_{n-1} - \theta_\star\right\| + \frac{\beta_0^2 + \sigma^2}{2} \gamma_n^2\right) \\
&= -\gamma_n \left\|\nabla L(\theta_{n-1})\right\|^2 \\
&\quad + \gamma_n^2 \beta_0^3 \left\|\theta_{n-1} - \theta_\star\right\|^2 + \frac{\beta_0(\beta_0^2 + \sigma^2)}{2} \gamma_n^2 \left\|\theta_{n-1} - \theta_\star\right\| .
\end{aligned}
$$

Then by convexity of L,

$$
L(\theta_{n-1}) - L(\theta_\star) \leq \nabla L(\theta_{n-1})^T (\theta_{n-1} - \theta_\star) \leq \|\nabla L(\theta_{n-1})\| \, \|\theta_{n-1} - \theta_\star\| .
$$

Thus

$$
\begin{aligned}
\Delta_{n-1} = \mathbf{E}\left[L(\theta_{n-1}) - L(\theta_\star)\right] &\leq \mathbf{E}\left[\|\nabla L(\theta_{n-1})\| \, \|\theta_{n-1} - \theta_\star\|\right] \\
&\leq (\mathbf{E}\left\|\nabla L(\theta_{n-1})\right\|^2)^{1/2} (\mathbf{E}\left\|\theta_{n-1} - \theta_\star\right\|^2)^{1/2} \\
&\leq \mathbf{E}\left\|\nabla L(\theta_{n-1})\right\|^{2^{1/2}} r_n
\end{aligned}
$$

to obtain

$$
\mathbf{E}\left\|\nabla L(\theta_{n-1})\right\|^2 \geq \frac{1}{r_n^2} \Delta_{n-1}^2 .
$$

Therefore

$$
\begin{aligned}
\mathbf{E}\left[\nabla L(\theta_{n-1})^T (\theta_n - \theta_{n-1})\right] &\leq -\gamma_n \mathbf{E}\left\|\nabla L(\theta_{n-1})\right\|^2 + \gamma_n^2 \beta_0^3 \mathbf{E}\left\|\theta_{n-1} - \theta_\star\right\|^2 + \frac{\beta_0(\beta_0^2 + \sigma^2)}{2} \gamma_n^2 \mathbf{E}\left\|\theta_{n-1} - \theta_\star\right\| \\
&\leq -\frac{\gamma_n}{r_n^2} \Delta_{n-1}^2 + \gamma_n^2 \beta_0^3 r_n^2 + \gamma_n^2 \frac{\beta_0(\beta_0^2 + \sigma^2)}{2} r_n . \qquad \text{(A.26)}
\end{aligned}
$$

Combining inequalities (A.24), (A.25), and (A.26), we have

$$
\Delta_n \leq \Delta_{n-1} - \frac{\gamma_n}{r_n^2} \Delta_{n-1}^2 + \frac{1}{2} \gamma_n^2 \bar{\beta}_n \sigma^2 , \qquad \text{(A.27)}
$$

where $\bar{\beta}_n = \sigma^{-2}(4\beta_0^3 r_n^2 + \beta_0^3 r_n + \beta_0 \sigma^2 r_n + 2\beta_0 \sigma^2)$.

If we let $\psi_n(t) = t - \frac{\gamma_n}{r_n^2} t^2$, then $\Delta_n \leq \psi_n(\Delta_{n-1}) + \frac{1}{2} \gamma_n^2 \bar{\beta}_n \sigma^2$. Since ψ is increasing for $t \in [0, \frac{r_n^2}{2\gamma_n}]$ and $\Delta_n = \mathbf{E}\left[L(\theta_n) - L(\theta_\star)\right] \leq \mathbf{E}\left[\nabla L(\theta_\star)^T (\theta_n - \theta_\star) + \frac{\beta_0}{2} \left\|\theta_n - \theta_\star\right\|^2\right] = \frac{\beta_0}{2} \mathbf{E}\left\|\theta_n - \theta_\star\right\|^2 \leq \frac{\beta_0}{2} r_n^2$, if we let $n_1 = \max\{\inf\{n \in \mathbb{N} : \gamma_n \leq 1/\beta_0\}, 3\}$, then $\psi_n(\Delta_n)$ is increasing for $n \geq n_1$.

Now let us consider a surrogate sequence defined as

$$
\tilde{\Delta}_n = \tilde{\Delta}_{n-1} - \frac{\gamma_n}{r_n^2} \tilde{\Delta}_{n-1}^2 + \frac{1}{2} \gamma_n^2 \bar{\beta}_n \sigma^2 , \quad \tilde{\Delta}_{n_1} = \Delta_{n_1} .
$$

Then $\Delta_n \le \tilde{\Delta}_n$ for $n \ge n_1$, since if we suppose $\Delta_{n-1} \le \tilde{\Delta}_{n-1}$, then

$$\Delta_n \le \psi_n(\Delta_{n-1}) + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \le \psi_n(\tilde{\Delta}_{n-1}) + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 = \tilde{\Delta}_n$$

from the monotonicity of ψ_n.

It suffices to bound $\tilde{\Delta}_n$. Let $\varepsilon_n = (4\bar{\beta}_n^{1/2}\sigma\gamma_1^{3/2})^{-1}\min\{r_1, r_n n^{3\gamma/2-1}\}$, which is decreasing. Since $\gamma_n^{1/2} - \gamma_{n-1}^{1/2} \ge \frac{\gamma_1^{1/2}}{4n^{\gamma/2}}$ for $n \ge 3$ (Proposition A.2), we have for all $n \ge n_1$,

$$\begin{aligned}\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2} &\ge \gamma_n^{1/2}(1+\varepsilon_{n+1})^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2} \\ &\ge \frac{\gamma_1^{1/2}}{4n^{\gamma/2}}(1+\varepsilon_{n+1})^{1/2} \ge \frac{\gamma_1^{1/2}}{4n^{\gamma/2}}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\varepsilon_n\bar{\beta}_n^{1/2}\sigma\gamma_n^2 r_n^{-1} &= (1/4)\bar{\beta}_n^{-1/2}\sigma^{-1}\gamma_1^{-3/2}\bar{\beta}_n^{1/2}\sigma\gamma_1^2 n^{-2\gamma} r_n^{-1}\min\{r_1, r_n n^{3\gamma/2-1}\} \\ &= \frac{\gamma_1^{1/2}}{4n^{2\gamma}}\min\Big\{\frac{r_1}{r_n}, n^{3\gamma/2-1}\Big\} \le \frac{\gamma_1^{1/2}}{4n^{2\gamma}} \le \frac{\gamma_1^{1/2}}{4n^{\gamma/2}}.\end{aligned}$$

Thus

$$\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2} \ge \varepsilon_n\bar{\beta}_n^{1/2}\sigma\gamma_n^2 r_n^{-1}, \quad n \ge n_1. \tag{A.28}$$

Let $n_2 = \inf\{n \ge n_1 : \tilde{\Delta}_{n-1}^2 \ge (1+\varepsilon_n)\bar{\beta}_n\gamma_n^2\sigma^2 r_n^2/2\}$. Assume for now n_2 is finite. We want to show that

$$\tilde{\Delta}_{n-1}^2 \ge (1+\varepsilon_n)\bar{\beta}_n\gamma_n\sigma^2 r_n^2/2, \quad n \ge n_2. \tag{A.29}$$

Indeed, inequality (A.29) is true for $n = n_2$ by construction. If $\tilde{\Delta}_{n-1} \ge (1+\varepsilon_n)^{1/2}\bar{\beta}_n^{1/2}\gamma_n^{1/2}\sigma r_n/\sqrt{2}$, then since ψ_n is increasing,

$$\begin{aligned}\tilde{\Delta}_n &= \psi_n(\tilde{\Delta}_{n-1}) + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \\ &\ge \psi_n((1+\varepsilon_n)^{1/2}\bar{\beta}_n^{1/2}\gamma_n^{1/2}\sigma r_n/\sqrt{2}) + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \\ &= (1+\varepsilon_n)^{1/2}\bar{\beta}_n^{1/2}\gamma_n^{1/2}\sigma r_n/\sqrt{2} - \frac{\gamma_n}{2r_n^2}(1+\varepsilon_n)\bar{\beta}_n\gamma_n\sigma^2 r_n^2 + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \\ &= \big(\tfrac{1+\varepsilon_n}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_n^{1/2}\sigma r_n - \tfrac{1}{2}(1+\varepsilon_n)\gamma_n^2\bar{\beta}_n\sigma^2 + \tfrac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \\ &= \big(\tfrac{1+\varepsilon_{n+1}}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_{n+1}^{1/2}\sigma r_n - \tfrac{1}{2}\varepsilon_n\gamma_n^2\bar{\beta}_n\sigma^2 - \big(\tfrac{1+\varepsilon_{n+1}}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_{n+1}^{1/2}\sigma r_n + \big(\tfrac{1+\varepsilon_n}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_n^{1/2}\sigma r_n \\ &= \big(\tfrac{1+\varepsilon_{n+1}}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_{n+1}^{1/2}\sigma r_n - \tfrac{1}{2}\varepsilon_n\gamma_n^2\bar{\beta}_n\sigma^2 + \tfrac{1}{\sqrt{2}}\bar{\beta}_n^{1/2}\sigma r_n[\gamma_n^{1/2}(1+\varepsilon_n)^{1/2} - \gamma_{n+1}^{1/2}(1+\varepsilon_{n+1})^{1/2}] \\ &\overset{(A.28)}{\ge} \big(\tfrac{1+\varepsilon_{n+1}}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_{n+1}^{1/2}\sigma r_n - \tfrac{1}{2}\varepsilon_n\gamma_n^2\bar{\beta}_n\sigma^2 + \tfrac{1}{\sqrt{2}}\varepsilon_n\gamma_n^2\bar{\beta}_n\sigma^2 \\ &\ge \big(\tfrac{1+\varepsilon_{n+1}}{2}\big)^{1/2}\bar{\beta}_n^{1/2}\gamma_{n+1}^{1/2}\sigma r_n.\end{aligned}$$

Hence inequality (A.29) holds for all $n \ge n_2$.

Therefore, for $n \ge n_2$,

$$\begin{aligned}\tilde{\Delta}_n &= \tilde{\Delta}_{n-1} - \frac{\gamma_n}{r_n^2}\tilde{\Delta}_{n-1}^2 + \frac{1}{2}\gamma_n^2\bar{\beta}_n\sigma^2 \\ &\overset{(A.29)}{\le} \tilde{\Delta}_{n-1} - \frac{\gamma_n}{r_n^2}\tilde{\Delta}_{n-1}^2 + \frac{\gamma_n}{2}\frac{\tilde{\Delta}_{n-1}^2}{1+\varepsilon_n} = \tilde{\Delta}_{n-1} - \frac{\gamma_n}{r_n^2}\frac{\varepsilon_n}{1+\varepsilon_n}\tilde{\Delta}_{n-1}^2\end{aligned}$$

Divide the preceding inequality by $\tilde{\Delta}_{n-1}\tilde{\Delta}_n$ to obtain

$$\tilde{\Delta}_{n-1}^{-1} \le \tilde{\Delta}_n^{-1} - \frac{\gamma_n}{r_n^2}\frac{\varepsilon_n}{1+\varepsilon_n}\frac{\tilde{\Delta}_{n-1}}{\tilde{\Delta}_n} \le \tilde{\Delta}_n^{-1} - \frac{\gamma_n}{r_n^2}\frac{\varepsilon_n}{1+\varepsilon_n},$$

where the last inequality is from $\tilde{\Delta}_n \le \tilde{\Delta}_{n-1}$. It follows $\tilde{\Delta}_{n_2-1}^{-1} \le \tilde{\Delta}_n^{-1} - \sum_{k=n_2}^n \frac{1}{r_k^2}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k$, or

$$\tilde{\Delta}_n \le \frac{1}{\tilde{\Delta}_{n_2}^{-1} + \sum_{k=n_2}^n \frac{1}{r_k^2}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k}, \quad n \ge n_2.$$

By the definition of n_2, $\tilde{\Delta}_{n_2-1} \le (1+\varepsilon_{n_2})^{1/2}\bar{\beta}_{n_2}^{1/2}\gamma_{n_2}\sigma r_{n_2}/\sqrt{2}$. Hence $\tilde{\Delta}_{n_2}^{-1} \ge \tilde{\Delta}_{n_2-1}^{-1} \ge (1+\varepsilon_{n_2})^{-1/2}\bar{\beta}_{n_2}^{-1/2}\gamma_{n_2}^{-1}\sigma^{-1}r_{n_2}^{-1}\sqrt{2}$. Now from inequality (A.28), since both ε_n and γ_n are decreasing, and $\bar{\beta}_n$ is increasing,

$$\begin{aligned}\gamma_{n+1}^{-1/2}(1+\varepsilon_{n+1})^{-1/2} - \gamma_n^{-1/2}(1+\varepsilon_n)^{-1/2} &\ge \bar{\beta}_n^{1/2}\sigma r_n^{-1}\frac{\varepsilon_n}{(1+\varepsilon_n)^{1/2}(1+\varepsilon_{n+1})^{1/2}}\frac{\gamma_n^2}{\gamma_n^{1/2}\gamma_{n+1}^{1/2}} \\ &\ge \sigma\bar{\beta}_n^{1/2}r_n^{-1}\frac{\varepsilon_n}{1+\varepsilon_n}\gamma_n, \quad n \ge n_1.\end{aligned}$$

Also since $\gamma_{n_1}^{1/2}(1+\varepsilon_{n_1})^{1/2} \ge \varepsilon_{n_1}\bar{\beta}_{n_1}^{1/2}\sigma\gamma_{n_1}r_{n_1}^{-1}$,

$$\gamma_{n_1}^{-1/2}(1+\varepsilon_{n_1})^{-1/2} \ge \sigma\frac{\varepsilon_{n_1}}{1+\varepsilon_{n_1}}\bar{\beta}_{n_1}^{1/2}r_{n_1}^{-1}\gamma_{n_1}, \quad n \ge n_1.$$

Therefore

$$\begin{aligned}\gamma_n^{-1/2}(1+\varepsilon_n)^{-1/2} &= \sum_{k=n_1}^{n-1}[\gamma_{k+1}^{-1/2}(1+\varepsilon_{k+1})^{-1/2} - \gamma_k^{-1/2}(1+\varepsilon_k)^{-1/2}] + \gamma_{n_1}^{-1/2}(1+\varepsilon_{n_1})^{-1/2} \\ &\ge \sigma\sum_{k=n_1}^{n-1}\frac{\bar{\beta}_k^{1/2}}{r_k}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k + \sigma\frac{\bar{\beta}_{n_1}^{1/2}}{r_{n_1}}\frac{\varepsilon_{n_1}}{1+\varepsilon_{n_1}}\gamma_{n_1} \\ &\ge \sigma\sum_{k=n_1}^{n-1}\frac{\bar{\beta}_k^{1/2}}{r_k}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k.\end{aligned}$$

Since $\bar{\beta}_n^{1/2}/r_n$ is decreasing,

$$\tilde{\Delta}_{n_2}^{-1} \ge \frac{\sqrt{2}}{\gamma_{n_2}}\frac{1}{r_{n_2}^2}\sum_{k=n_1}^{n_2-1}\frac{\bar{\beta}_k^{1/2}}{r_k}\frac{r_{n_2}}{\bar{\beta}_{n_2}^{1/2}}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k \ge \frac{\sqrt{2}}{\gamma_{n_2}}\frac{1}{r_{n_2}^2}\sum_{k=n_1}^{n_2-1}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k. \tag{A.30}$$

This entails, as r_n^2 is increasing, for $n \ge n_2$,

$$\tilde{\Delta}_n \le \frac{1}{\frac{\sqrt{2}}{\gamma_{n_2}}\frac{1}{r_{n_2}^2}\sum_{k=n_1}^{n_2-1}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k + \sum_{k=n_2}^n\frac{1}{r_k^2}\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k} \le \frac{\max\{\frac{\gamma_{n_2}}{\sqrt{2}},1\}r_n^2}{\sum_{k=n_1}^n\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k}. \tag{A.31}$$

Given the derivation of inequality (A.30), inequality (A.31) holds for $n_1 \le n < n_2$, thus it is true even if n_2 is infinite. Therefore, inequality (A.31) holds for $n \ge n_1$.

In order to obtain the rate, note

$$r_n^2 = \delta_0 + \gamma_1^2\sum_{k=1}^n k^{-2\gamma} \le \begin{cases} r^2 \triangleq \delta_0 + \gamma_1^2\zeta(2\gamma) < \infty, & \gamma \in (1/2,1], \\ \delta_0 + \gamma_1^2(1+\phi_{2\gamma}(n)), & \gamma \in (0,1/2]. \end{cases}$$

Also, for $\gamma \in [2/3,1]$, $\varepsilon_n = (4\bar{\beta}_n^{1/2}\sigma\gamma_1^{1/2})^{-1}r_1$. Since $\bar{\beta}_n \le \bar{\beta}_\infty \triangleq \sigma^2[4\beta_0^3r^2 + (\beta_0^3+\beta_0\sigma^2)r + 2\beta_0\sigma^2] < \infty$,

$$\begin{aligned}\sum_{k=n_1}^n\frac{\varepsilon_k}{1+\varepsilon_k}\gamma_k &= \sum_{k=n_1}^n\frac{1}{\varepsilon_k^{-1}+1}\gamma_k = \sum_{k=n_1}^n\frac{1}{\frac{4\sigma\gamma_1^{1/2}}{r_1}\bar{\beta}_k^{1/2}+1}\gamma_k \\ &\ge \frac{\gamma_1}{\frac{4\sigma\gamma_1^{1/2}}{r_1}\bar{\beta}_\infty^{1/2}+1}\sum_{k=n_1}^n\frac{1}{k^\gamma} \ge \frac{\gamma_1/2}{\frac{4\sigma\gamma_1^{1/2}}{r_1}\bar{\beta}_\infty^{1/2}+1}[\phi_\gamma(n) - \phi_\gamma(n_1-1)].\end{aligned}$$

If $\gamma \in (0, 2/3)$, then $\varepsilon_n = (4\bar{\beta}_n^{1/2}\sigma\gamma_1^{1/2})^{-1} r_n n^{3\gamma/2-1}$. Since $\bar{\beta}_n^{1/2}/r_n$ is decreasing,

$$\sum_{k=n_1}^{n} \frac{1}{\varepsilon_k^{-1}+1}\gamma_k = \sum_{k=n_1}^{n} \frac{1}{4\sigma\gamma_1^{1/2}\frac{\bar{\beta}_k^{1/2}}{r_k}k^{1-3\gamma/2}+1}\gamma_k = \sum_{k=n_1}^{n} \frac{\gamma_1}{4\sigma\gamma_1^{1/2}\frac{\bar{\beta}_k^{1/2}}{r_k}k^{1-\gamma/2}+k^{\gamma}}$$
$$\geq \frac{\gamma_1}{4\sigma\gamma_1^{1/2}\frac{\bar{\beta}_1^{1/2}}{r_1}+1}\sum_{k=n_1}^{n}\frac{1}{k^{1-\gamma/2}} \geq \frac{\gamma_1/2}{4\sigma\gamma_1^{1/2}\frac{\bar{\beta}_1^{1/2}}{r_1}+1}[\phi_{1-\gamma/2}(n)-\phi_{1-\gamma/2}(n_1-1)].$$

Therefore,

$$\Delta_n \leq \begin{cases} \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_1^{1/2}r_1^{-1}+\gamma_1}\frac{\delta_0+\gamma_1^2(1+\phi_{2\gamma}(n))}{\phi_{1-\gamma/2}(n)-\phi_{1-\gamma/2}(n_1-1)}, & \gamma\in(0,1/2], \\ \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_1^{1/2}r_1^{-1}+\gamma_1}\frac{\delta_0+\gamma_1^2\zeta(2\gamma)}{\phi_{1-\gamma/2}(n)-\phi_{1-\gamma/2}(n_1-1)}, & \gamma\in(1/2,2/3), \\ \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_\infty^{1/2}r_1^{-1}+\gamma_1}\frac{\delta_0+\gamma_1^2\zeta(2\gamma)}{\phi_{\gamma}(n)-\phi_{\gamma}(n_1-1)}, & \gamma\in[2/3,1], \end{cases}$$

for $n \geq n_1$ (note $\gamma_{n_2} \leq \gamma_{n_1}$).

Below we put the explicit values of the constants:

$$\Gamma_1 = \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_1^{1/2}r_1^{-1}+\gamma_1}, \quad \Gamma_2 = \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_1^{1/2}r_1^{-1}+\gamma_1} \quad \Gamma_3 = \frac{2\max\{\frac{\gamma_{n_1}}{\sqrt{2}},1\}}{4\sigma\gamma_1^{3/2}\bar{\beta}_\infty^{1/2}r_1^{-1}+\gamma_1}$$
$$r_n^2 = \delta_0 + \sigma^2\sum_{k=1}^{n}\gamma_k^2,$$
$$\bar{\beta}_n = \sigma^{-2}(4\beta_0^3 r_n^2 + \beta_0^3 r_n + \beta_0\sigma^2 r_n + 2\beta_0\sigma^2),$$
$$\bar{\beta}_\infty = \sigma^2[4\beta_0^3 r^2 + (\beta_0^3+\beta_0\sigma^2)r + 2\beta_0\sigma^2] < \infty,$$
$$r = \delta_0 + \sigma^2\sum_{k=1}^{\infty}\gamma_k^2.$$

□

A.9. Proof of Theorem 5.2

Proof of Theorem 5.2. Pick a minimizer $\theta_\star$ of L. From the main iteration (3′), we have $\theta_k - \theta_\star = \theta_{k-1} - \theta_\star - \gamma_k\nabla\ell(Z_k,\theta_k)$ and

$$\begin{aligned} \|\theta_k-\theta_\star\|^2 &= \|\theta_{k-1}-\theta_\star\|^2 - 2\gamma_k\nabla\ell(Z_k,\theta_k)^T(\theta_{k-1}-\theta_\star) + \gamma_k^2\|\nabla\ell(Z_k,\theta_k)\|^2 \\ &= \|\theta_{k-1}-\theta_\star\|^2 - 2\gamma_k\nabla\ell(Z_k,\theta_{k-1})^T(\theta_{k-1}-\theta_\star) \\ &\quad + 2\gamma_k(\nabla\ell(Z_k,\theta_{k-1})-\nabla\ell(Z_k,\theta_k))^T(\theta_{k-1}-\theta_\star) + \gamma_k^2\|\nabla\ell(Z_k,\theta_k)\|^2 \\ &\leq \|\theta_{k-1}-\theta_\star\|^2 - 2\gamma_k\nabla\ell(Z_k,\theta_{k-1})^T(\theta_{k-1}-\theta_\star) \\ &\quad + 2\|R_k\|\|\theta_{k-1}-\theta_\star\| + \gamma_k^2\|\nabla\ell(Z_{k-1},\theta_k)\|^2, \end{aligned}$$

where the final inequality is due to Lemma 3.1. Therefore

$$\begin{aligned} \mathbf{E}[\|\theta_k-\theta_\star\|^2|\mathcal{F}_{k-1}] &\leq \|\theta_{k-1}-\theta_\star\|^2 - 2\gamma_k\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star) \\ &\quad + 2\mathbf{E}[\|R_k\||\mathcal{F}_{k-1}]\|\theta_{k-1}-\theta_\star)\| + \gamma_k^2\mathbf{E}[\|\nabla\ell(Z_{k-1},\theta_k)\|^2|\mathcal{F}_{k-1}]. \end{aligned} \tag{A.32}$$

From the cocoercivity of Lipschitz continuous gradient operator (Bauschke & Combettes, 2011), we see

$$\beta_0^{-1}\|\nabla\ell(Z_k,\theta_{k-1})-\nabla\ell(Z_k,\theta_\star)\|^2 \leq (\nabla\ell(Z_k,\theta_{k-1})-\nabla\ell(Z_k,\theta_\star))^T(\theta_{k-1}-\theta_\star)$$

and thus

$$\|\nabla\ell(Z_k,\theta_{k-1})\|^2 = \|\ell(Z_k,\theta_{k-1})-\ell(Z_k,\theta_\star)+\ell(Z_k,\theta_\star)\|^2$$

$$\leq 2\left\|\ell(Z_k,\theta_{k-1})-\ell(Z_k,\theta_\star)\right\|^2+2\left\|\ell(Z_k,\theta_\star)\right\|^2$$
$$\leq 2\beta_0(\nabla\ell(Z_k,\theta_{k-1})-\nabla\ell(Z_k,\theta_\star))^T(\theta_{k-1}-\theta_\star)+2\left\|\ell(Z_k,\theta_\star)\right\|^2.$$

Therefore from inequality (A.32)

$$\begin{aligned}\mathbf{E}\left\|\theta_k-\theta_\star\right\|^2 &\leq \mathbf{E}\left\|\theta_{k-1}-\theta_\star\right\|^2-2\gamma_k\,\mathbf{E}\left[\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)\right]\\ &\quad+2\gamma_k^2\beta_0^2\,\mathbf{E}\left\|\theta_{k-1}-\theta_\star\right\|^2+(\beta_0^2+\sigma^2)\gamma_k^2\,\mathbf{E}\left\|\theta_{k-1}-\theta_\star\right\|^2\\ &\quad+2\gamma_k^2\beta_0\,\mathbf{E}\left[\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)\right]+2\gamma_k^2\sigma^2\\ &\leq \mathbf{E}\left\|\theta_{k-1}-\theta_\star\right\|^2-2\gamma_k(1-\beta_0\gamma_k)\,\mathbf{E}\left[\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)\right]+2\gamma_k^2\tilde\sigma^2,\end{aligned}$$

where $\tilde\sigma^2=\beta_0^2r^2+r(\beta_0^2+\sigma^2)/2+\sigma^2$; the last inequality is due to Lemma A.1. Let δ_k denote $\mathbf{E}\left\|\theta_k-\theta_\star\right\|^2$. The previous inequality implies

$$2\gamma_k(1-\beta_0\gamma_k)\,\mathbf{E}\left[\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)\right]\leq\delta_{k-1}-\delta_k+2\gamma_k^2\tilde\sigma^2.$$

If we let $n_*=\inf\{k\in\mathbb{N}:(1-\gamma_k\beta_0)\geq 1/2\}$, then $\mathbf{E}\left[\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)\right]\leq\gamma_k^{-1}(\delta_{k-1}-\delta_k+2\gamma_k^2\tilde\sigma^2)$ for all $k\geq n_*$.

Lipschitz continuity of ∇L implies $L(\theta_{k-1})-L(\theta_\star)\leq\frac{\beta_0}{2}\left\|\theta_{k-1}-\theta_\star\right\|^2$. Also, by convexity, $L(\theta_{k-1})-L(\theta_\star)\leq\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star)$. Therefore, for $n>n_*$,

$$\begin{aligned}L(\bar\theta_n)-L(\theta_\star)&=L(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k)-L(\theta_\star)\\ &\leq\frac{1}{n}\sum_{k=1}^{n}(L(\theta_{k-1})-L(\theta_\star))\\ &=\frac{1}{n}\sum_{k=1}^{n_*}(L(\theta_{k-1})-L(\theta_\star))+\frac{1}{n}\sum_{k=n_*+1}^{n}(L(\theta_{k-1})-L(\theta_\star))\\ &\leq\frac{\beta_0}{2n}\sum_{k=1}^{n_*}\left\|\theta_{k-1}-\theta_\star\right\|^2+\frac{1}{n}\sum_{k=n_*+1}^{n}\nabla L(\theta_{k-1})^T(\theta_{k-1}-\theta_\star).\end{aligned}$$

Then it follows

$$\mathbf{E}\left[L(\bar\theta_n)-L(\theta_\star)\right]\leq\frac{1}{n}\Big(\frac{\beta_0}{2}\sum_{k=1}^{n_*}\delta_{k-1}+\sum_{k=n_*+1}^{n}\gamma_k^{-1}(\delta_{k-1}-\delta_k+2\gamma_k^2\tilde\sigma^2)\Big).$$

Observe that $\delta_{k-1}\leq r_k^2\triangleq\delta_0+\sigma^2\sum_{j=1}^k\gamma_j^2\leq\delta_0+\sigma^2\gamma_1^2[1+\phi_{2\gamma}(k)]$. Since $1+\phi_{2\gamma}(k)\leq k^{1-2\gamma}/(1-2\gamma)$ if $\gamma<1/2$ and $\phi_{2\gamma}(k)=\log k$ if $\gamma=1/2$, and $\sum_{j=1}^k j^{-2\gamma}\leq\sum_{j=1}^\infty j^{-2\gamma}=\zeta(2\gamma)<\infty$ if $\gamma>1/2$ where $\zeta(\cdot)$ is the Riemann zeta function, it follows that

$$\sum_{k=1}^{n_*}r_k^2\leq\begin{cases}n_*\delta_0+\frac{\sigma^2\gamma_1^2}{1-2\gamma}\phi_{2\gamma-1}(n_*), & \gamma<1/2,\\ n_*\delta_0+\sigma^2\gamma_1^2\log(n_*+1), & \gamma=1/2,\\ n_*[\delta_0+\sigma^2\gamma_1^2\zeta(2\gamma)], & \gamma>1/2.\end{cases}$$

On the other hand,

$$\begin{aligned}\sum_{n_*+1}^{n}\gamma_k^{-1}(\delta_{k-1}-\delta_k)&=\gamma_{n_*+1}^{-1}\delta_{n_*}+\sum_{k=n_*+1}^{n-1}\delta_k(\gamma_{k+1}^{-1}-\gamma_k^{-1})-\gamma_n^{-1}\delta_n\\ &\leq r_n^2\gamma_{n_*+1}^{-1}+r_n^2\sum_{k=n_*+1}^{n-1}(\gamma_{k+1}^{-1}-\gamma_k^{-1})-\gamma_n^{-1}\delta_n\\ &=r_n^2\gamma_n^{-1}-\gamma_n^{-1}\delta_n\leq r_n^2\gamma_n^{-1}\end{aligned}$$

(Bach & Moulines, 2011, p. 27). Thus

$$\begin{aligned}
\mathbf{E}\left[L(\bar{\theta}_n) - L(\theta_\star)\right] &\leq \frac{1}{n}\Big(\frac{\beta_0}{2}\sum_{k=1}^{n_*} r_k^2 + r_n^2\gamma_n^{-1} + 2\tilde{\sigma}^2\sum_{k=1}^{n}\gamma_k\Big) \\
&\leq \frac{1}{n}\Big(\frac{\beta_0}{2}\sum_{k=1}^{n_*} r_k^2 + \delta_0\gamma_n^{-1} + \sigma^2\gamma_1^2[1+\phi_{2\gamma}(n)]\gamma_n^{-1} + 2\tilde{\sigma}^2\gamma_1\phi_\gamma(n)\Big) \\
&= \frac{\beta_0}{2n}\sum_{k=1}^{n_*} r_k^2 + \frac{\delta_0}{\gamma_1 n^{1-\gamma}} + \sigma^2\gamma_1^2\frac{1+\phi_{2\gamma}(n)}{n^{1-\gamma}} + 2\tilde{\sigma}^2\gamma_1\frac{\phi_\gamma(n)}{n} \\
&\leq \begin{cases} \frac{\beta_0[n_*\delta_0 + \frac{\sigma^2\gamma_1^2}{1-2\gamma}\phi_{2\gamma-1}(n_*)]}{2n} + \frac{\sigma^2\gamma_1^2}{(1-2\gamma)n^\gamma} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)n^\gamma}, & \gamma < 1/2, \\ \frac{\beta_0[n_*\delta_0 + \sigma^2\gamma_1^2\log(n_*+1)]}{2n} + \sigma^2\gamma_1^2\frac{1+\log n}{\sqrt{n}} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)\sqrt{n}}, & \gamma = 1/2. \end{cases}
\end{aligned}$$

When $\gamma > 1/2$, we can replace the $1 + \phi_{2\gamma}(n)$ with $\zeta(2\gamma)$, hence

$$\mathbf{E}\left[L(\bar{\theta}_n) - L(\theta_\star)\right] \leq \begin{cases} \frac{\beta_0 n_*[\delta_0 + \sigma^2\gamma_1^2\zeta(2\gamma)]}{2n} + \frac{\sigma^2\gamma_1^2\zeta(2\gamma)}{n^{1-\gamma}} + \frac{2\tilde{\sigma}^2\gamma_1}{(1-\gamma)n^\gamma}, & \gamma \in (1/2, 1), \\ \frac{\beta_0 n_*[\delta_0 + \sigma^2\gamma_1^2\zeta(2\gamma)]}{2n} + \sigma^2\gamma_1^2\zeta(2) + \frac{2\tilde{\sigma}^2\gamma_1\log n}{(1-\gamma)n}, & \gamma = 1. \end{cases}$$

The preceding argument immediately yields that for $n \leq n_*$,

$$\mathbf{E}\left[L(\bar{\theta}_n) - L(\theta_\star)\right] \leq \begin{cases} \frac{\beta_0}{2}\Big(\delta_0 + \frac{\sigma^2\gamma_1^2}{(1-2\gamma)(2-2\gamma)}n^{1-2\gamma}\Big), & \gamma < 1/2, \\ \frac{\beta_0}{2}\Big(\delta_0 + \sigma^2\gamma_1^2\frac{\log(n+1)}{n}\Big), & \gamma = 1/2, \\ \frac{\beta_0}{2}\Big(\delta_0 + \frac{\sigma^2\gamma_1^2\zeta(2\gamma)}{n}\Big), & \gamma > 1/2, \end{cases}$$

Since $\phi_{2\gamma-1}(n) \leq n^{2-2\gamma}/(2-2\gamma)$ when $\gamma < 1/2$.

Below we put the explicit values of the constants:

$$\begin{aligned}
\tilde{\Gamma}_1 &= \frac{\beta_0[n_*\delta_0 + \frac{\sigma^2\gamma_1^2}{1-2\gamma}\phi_{2\gamma-1}(n_*)]}{2}, \quad \tilde{\Gamma}_2 = \frac{\beta_0[n_*\delta_0 + \sigma^2\gamma_1^2\log(n_*+1)]}{2} \quad \tilde{\Gamma}_3 = \frac{\beta_0 n_*[\delta_0 + \sigma^2\gamma_1^2\zeta(2\gamma)]}{2}, \\
\tilde{\sigma}^2 &= \beta_0^2 r^2 + r(\beta_0^2 + \sigma^2)/2 + \sigma^2, \\
r &= \delta_0 + \sigma^2\sum_{k=1}^{\infty}\gamma_k^2.
\end{aligned}$$

□

B. Proofs of technical lemmas

Proof of Lemma 3.1. Note from the ISGD update (3) that $\theta_{n-1} - \theta_n = \gamma_n\nabla\ell(Z_n, \theta_n)$. Therefore

$$\gamma_n[\ell(Z_n, \theta_n) - \ell(Z_n, \theta_{n-1})] \leq \gamma_n\nabla\ell(Z_n, \theta_n)^T(\theta_n - \theta_{n-1}) = -\left\|\theta_n - \theta_{n-1}\right\|^2,$$

where the first inequality follows from the convexity of $\ell(Z_n, \cdot)$, i.e., Assumptions A1. Using this convexity once more, we see $\ell(Z_n, \theta_n) - \ell(Z_n, \theta_{n-1}) \geq \nabla\ell(Z_n, \theta_{n-1})^T(\theta_n - \theta_{n-1})$. It follows that

$$\left\|\theta_n - \theta_{n-1}\right\|^2 \leq \gamma_n\nabla\ell(Z_n, \theta_{n-1})(\theta_{n-1} - \theta_n) \leq \gamma_n\left\|\nabla\ell(Z_n, \theta_{n-1})\right\|\left\|\theta_{n-1} - \theta_n\right\|$$

and the claim is proved. □

Proof of Lemma A.2. Let us first consider the case $n < n_0$. Since $c_n \downarrow 0$, $Q_{i+1}^n < (1 + c_{i+1})\cdots(1 + c_n) \leq (1 + c_1)^{n_0}$.

By expanding inequality (A.6), we see

$$\begin{aligned} y_n &\le Q_1^n y_0 + \sum_{i=1}^n Q_{i+1}^n a_i \le Q_1^n y_0 + (1+c_1)^{n_0} \sum_{i=1}^n a_i \\ &\le Q_1^n y_0 + (1+c_1)^{n_0} A \\ &\le K_0 \frac{a_n}{b_n} + Q_1^n y_0 + (1+c_1)^{n_0} A + B, \end{aligned} \tag{B.1}$$

where the last line is inequality (A.7).

Now consider the case $n \ge n_0$. Since

$$\frac{1+(1+\delta)b_1}{1+\delta-\delta_n-\zeta_n} \le \frac{1+(1+\delta)b_1}{1+\delta-\delta_{n_0}-\zeta_{n_0}} = K_0$$

and $b_n \downarrow 0$, it follows that $K_0(\delta_n+\zeta_n)+1+(1+\delta)b_n \le K_0(1+\delta)$, or

$$\frac{K_0}{a_n}\left(\frac{a_{n-1}}{b_{n-1}} - \frac{a_n}{b_n} + \frac{c_n a_{n-1}}{b_{n-1}}\right) + 1 + (1+\delta)b_n \le K_0(1+\delta).$$

This implies

$$a_n[1+(1+\delta)b_n] \le K_0\left(\frac{[1+(1+\delta)b_n]a_n}{b_n} - \frac{(1+c_n)a_{n-1}}{b_{n-1}}\right)$$

and

$$a_n \le K_0\left(\frac{a_n}{b_n} - \frac{1+c_n}{1+(1+\delta)b_n}\frac{a_{n-1}}{b_{n-1}}\right). \tag{B.2}$$

Combining inequalities (A.6) and (B.2), we obtain

$$y_n - K_0\frac{a_n}{b_n} \le \frac{1+c_n}{1+(1+\delta)b_n}\left(y_{n-1} - K_0\frac{a_{n-1}}{b_{n-1}}\right).$$

Define $s_n = y_n - K_0 a_n/b_n$. Then $|s_n| \le (1+c_n)[1+(1+\delta)b_n]^{-1}|s_{n-1}|$ and thus $|s_n| \le Q_{n_0+1}^n |s_{n_0}|$. Therefore,

$$\begin{aligned} y_n - K_0\frac{a_n}{b_n} \le |s_n| &\le Q_{n_0+1}^n \left| y_{n_0} - K_0\frac{a_{n_0}}{b_{n_0}}\right| \\ &\le Q_{n_0+1}^n \left(y_{n_0} + K_0\frac{a_{n_0}}{b_{n_0}}\right) \\ &= Q_{n_0+1}^n y_{n_0} + Q_{n_0+1}^n B \\ &\le Q_1^n y_0 + Q_{n_0+1}^n (1+c_1)^{n_0} A + Q_{n_0+1}^n B, \end{aligned}$$

where the last line follows from inequality (B.1). □

Proof of Corollary A.1. It is easy to verify that

$$1 - \eta n^{-\gamma} + \nu n^{-\beta} \le \frac{1+\nu(1+\eta)n^{-\beta}}{1+\eta n^{-\gamma}}$$

for all $n \ge 1$. Let $a_n = a_1 n^{-\alpha}$, $b_n = \eta n^{-\gamma}/(1+\delta)$, $c_n = \nu(1+\eta)n^{-\beta}$. Then inequality (A.6) in Lemma A.2 holds for (a_n, b_n, c_n).

To see if the conditions for Lemma A.2 hold, verify that $a_n \downarrow 0$, $\sum_{n=1}^\infty a_n < \infty$, $b_n \downarrow 0$, $c_n \downarrow 0$, $c_n/b_n \propto n^{-(\beta-\alpha)} \downarrow 0$, hence $c_n/b_n < 1$ for all sufficiently large n, and

$$\delta_n = \frac{1}{a_n}\left(\frac{a_{n-1}}{b_{n-1}} - \frac{a_n}{b_n}\right) = \frac{1+\delta}{\eta} n^\gamma \left(\left[\frac{n}{n-1}\right]^{\alpha-\gamma} - 1\right) \downarrow \delta,$$

$$\zeta_n = \frac{c_n}{b_{n-1}}\frac{a_{n-1}}{a_n} = \nu(1+\eta)(1+\delta)\left(\frac{n}{n-1}\right)^{\alpha-\gamma} n^{-(\beta-\gamma)} \downarrow 0.$$

Recall that $\delta = \frac{1+\delta}{\eta}(\alpha-\gamma)$ if $\gamma = 1$.

Therefore inequality (A.7) from Lemma A.2 translates to

$$y_n \le K_1 n^{-(\alpha-\gamma)} + Q_1^n y_0 + Q_{n_0+1}^n \left[(1+c_1)^{n_0} A + B\right], \tag{B.3}$$

where K_1, A, and B are given in equations (A.10a) and (A.10d); the conditions for the n_0 translates to inequalities (A.11). Furthermore,

$$Q_i^n = \begin{cases} \prod_{j=i}^n \frac{1+\nu(1+\eta)n^{-\beta}}{1+\eta n^{-\gamma}}, & n \ge i, \\ 1, & n < i. \end{cases}$$

Since $b_n, c_n \downarrow 0$, $Q_1^n = \prod_{j=1}^n \frac{1+c_j}{1+(1+\delta)b_j} \ge \prod_{j=1}^n \frac{1}{1+(1+\delta)b_j} \ge (1+\eta)^{-n}$. Hence

$$Q_{n_0+1}^n = Q_1^n / Q_1^{n_0} \le (1+\eta)^{n_0} Q_1^n = (1+\eta)^{n_0} Q_1^n. \tag{B.4}$$

In order to bound Q_1^n, take a logarithm to see

$$\log Q_1^n = \sum_{k=1}^n \log(1+\nu(1+\eta)k^{-\beta}) - \sum_{k=1}^n \log(1+\eta k^{-\gamma}).$$

For the first term, use $\log(1+x) \le x$ for $x \ge 0$ to get

$$\sum_{k=1}^n \log(1+\nu(1+\eta)k^{-\beta}) \le \nu(1+\eta)\sum_{k=1}^n \frac{1}{k^\beta} \le \begin{cases} \nu(1+\eta)\sum_{k=1}^\infty \frac{1}{k^\beta}, & \beta > 1, \\ \nu(1+\eta)\phi_\beta(n), & \beta \le 1, \end{cases},$$

since $\sum_{k=1}^n k^{-\beta} \le \phi_\beta(n)$ for $\beta \le 1$. For the second term, since $x \mapsto x^{-1}\log(1+x)$ is decreasing for $x > -1$ and $k^{-\gamma} \downarrow 0$, we have $\log(1+\eta k^{-\gamma}) \ge k^{-\gamma}\log(1+\eta)$ to get

$$-\sum_{i=1}^n \log(1+\eta k^{-\gamma}) \le -\log(1+\eta)\sum_{k=1}^n \frac{1}{k^\gamma} \le -\frac{1}{2}\log(1+\eta)\phi_\gamma(n), \tag{B.5}$$

since $\sum_{k=1}^n k^{-\gamma} \ge \frac{1}{2}\phi_\gamma(n)$ for $\gamma \le 1$. Thus

$$Q_1^n \le K_2(n)\exp\left(-\tfrac{1}{2}\log(1+\eta)\phi_\gamma(n)\right), \tag{B.6}$$

where $K_2(n)$ is given in equation (A.10b). Combining inequalities (B.3), (B.4), and (B.6), we finally obtain inequality (A.9), with D_{n_0} given in equation (A.10c). □

Proof of Lemma A.5. Recall that $\theta_n - \theta_\star = \theta_{n-1} - \theta_\star - \gamma_n \nabla\ell(Z_n, \theta_n)$. Therefore,

$$\begin{aligned} \|\theta_n - \theta_\star\|^2 &= \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n(\theta_{n-1} - \theta_\star)^T \nabla\ell(Z_n, \theta_n) + \gamma_n^2 \|\nabla\ell(Z_n, \theta_n)\|^2 \\ &\le \|\theta_{n-1} - \theta_\star\|^2 - 2\gamma_n W_n + \gamma_n^2 V_n, \end{aligned} \tag{B.7}$$

where $W_n = (\theta_{n-1} - \theta_\star)^T \nabla\ell(Z_n, \theta_n)$ and $V_n = \|\nabla\ell(Z_n, \theta_{n-1})\|^2$; the last line follows from Lemma 3.1.

We first bound the fourth moment. Squaring both sides of inequality (B.7) yields

$$\begin{aligned} \|\theta_n - \theta_\star\|^4 \le \|\theta_{n-1} - \theta_\star\|^4 &+ 4\gamma_n^2 W_n^2 + \gamma_n^4 V_n^2 \\ &- 4\gamma_n \|\theta_{n-1} - \theta_\star\|^2 W_n + 2\gamma_n^2 \|\theta_{n-1} - \theta_\star\|^2 V_n - 4\gamma_n^3 W_n V_n. \end{aligned}$$

In order to bound $\mathbf{E}\left[\|\theta_n - \theta_\star\|^4 | \mathcal{F}_{n-1}\right]$, first note that

$$\mathbf{E}\left[W_n | \mathcal{F}_{n-1}\right] = \mathbf{E}\left[(\theta_{n-1} - \theta_\star)^T \nabla\ell(Z_n, \theta_{n-1}) - \gamma_n^{-1}(\theta_{n-1} - \theta_\star)^T R_n) | \mathcal{F}_{n-1}\right]$$

$$
\begin{aligned}
&\geq (\theta_{n-1}-\theta_\star)^T \mathbf{E}\left[\nabla \ell(Z_n,\theta_{n-1})|\mathcal{F}_{n-1}\right] - \gamma_n^{-1}\left\|\theta_{n-1}-\theta_\star\right\| \mathbf{E}\left[\|R_n\||\mathcal{F}_{n-1}\right] \\
&\geq (\theta_{n-1}-\theta_\star)^T \mathbf{E}\left[\nabla \ell(Z_n,\theta_{n-1})|\mathcal{F}_{n-1}\right] - \gamma_n\beta_0^2\left\|\theta_{n-1}-\theta_\star\right\|^2 - \gamma_n(\beta_0^2+\sigma^2)\left\|\theta_{n-1}-\theta_\star\right\| \\
&= (\theta_{n-1}-\theta_\star)^T(\nabla L(\theta_{n-1}) - \nabla L(\theta_\star)) - \gamma_n\beta_0^2\left\|\theta_{n-1}-\theta_\star\right\|^2 - \gamma_n(\beta_0^2+\sigma^2)\left\|\theta_{n-1}-\theta_\star\right\| \\
&\geq \lambda\left\|\theta_{n-1}-\theta_\star\right\|^2 - \gamma_n\beta_0^2\left\|\theta_{n-1}-\theta_\star\right\|^2 - \gamma_n(\beta_0^2+\sigma^2)\left\|\theta_{n-1}-\theta_\star\right\|. \qquad \text{(B.8)}
\end{aligned}
$$

The first inequality is Cauchy-Schwarz, the second is inequality (A.3), and the last inequality is from the strong convexity of L. In addition,

$$
\mathbf{E}\left[V_n|\mathcal{F}_{n-1}\right] \leq 2\beta_0^2\left\|\theta_{n-1}-\theta_\star\right\|^2 + 2\sigma^2 \qquad \text{(B.9)}
$$

from inequality (A.4), and

$$
\begin{aligned}
\mathbf{E}\left[-W_nV_n|\mathcal{F}_{n-1}\right] &= -\mathbf{E}\left[(\theta_{n-1}-\theta_\star)^T\nabla\ell(Z_n,\theta_n)\left\|\nabla\ell(Z_n,\theta_{n-1})\right\|^2|\mathcal{F}_{n-1}\right] \\
&\leq \left\|\theta_{n-1}-\theta_\star\right\|\mathbf{E}\left[\|\nabla\ell(Z_n,\theta_n)\|\left\|\nabla\ell(Z_n,\theta_{n-1})\right\|^2|\mathcal{F}_{n-1}\right] \\
&\leq \left\|\theta_{n-1}-\theta_\star\right\|\mathbf{E}\left[\left\|\nabla\ell(Z_n,\theta_{n-1})\right\|^3|\mathcal{F}_{n-1}\right]
\end{aligned}
$$

again by Lemma 3.1. Since $\|\nabla\ell(Z_n,\theta_{n-1})\| \leq \beta_0\|\theta_{n-1}-\theta_\star\| + \|\nabla\ell(Z_n,\theta_\star)\|$ by Assumption A2′, combining with Assumption A4′ we have

$$
\begin{aligned}
\mathbf{E}\left[\left\|\nabla\ell(Z_n,\theta_{n-1})\right\|^3|\mathcal{F}_{n-1}\right] &\leq 4\beta_0^3\left\|\theta_{n-1}-\theta_\star\right\|^3 + 4\,\mathbf{E}\left\|\nabla\ell(Z_n,\theta_\star)\right\|^3 \\
&\leq 4\beta_0^3\left\|\theta_{n-1}-\theta_\star\right\|^3 + 4\sigma^3.
\end{aligned}
$$

by Proposition A.1. Therefore

$$
\mathbf{E}\left[-W_nV_n|\mathcal{F}_{n-1}\right] \leq 4\beta_0^3\left\|\theta_{n-1}-\theta_\star\right\|^4 + 4\sigma^3\left\|\theta_{n-1}-\theta_\star\right\|. \qquad \text{(B.10)}
$$

Finally,

$$
\begin{aligned}
\mathbf{E}\left[W_n^2|\mathcal{F}_{n-1}\right] &= \mathbf{E}\left[((\theta_{n-1}-\theta_\star)^T\nabla\ell(Z_n,\theta_n))^2|\mathcal{F}_{n-1}\right] \\
&\leq \left\|\theta_{n-1}-\theta_\star\right\|^2\mathbf{E}\left[\|\ell(Z_n,\theta_n)\|^2|\mathcal{F}_{n-1}\right] \\
&\leq \left\|\theta_{n-1}-\theta_\star\right\|^2\mathbf{E}\left[\|\ell(Z_n,\theta_{n-1})\|^2|\mathcal{F}_{n-1}\right] \\
&\leq 2\beta_0^2\left\|\theta_{n-1}-\theta_\star\right\|^4 + 2\sigma^2\left\|\theta_{n-1}-\theta_\star\right\|^2 \qquad \text{(B.11)}
\end{aligned}
$$

from inequality (B.9), and

$$
\begin{aligned}
\mathbf{E}\left[V_n^2|\mathcal{F}_{n-1}\right] &\leq \mathbf{E}\left[\|\ell(Z_n,\theta_{n-1})\|^4|\mathcal{F}_{n-1}\right] \\
&\leq 8\beta_0^4\left\|\theta_{n-1}-\theta_\star\right\|^4 + 8\,\mathbf{E}\left[\|\ell(Z_n,\theta_\star)\|^4|\mathcal{F}_{n-1}\right] \\
&\leq 8\beta_0^4\left\|\theta_{n-1}-\theta_\star\right\|^4 + 8\sigma^4 \qquad \text{(B.12)}
\end{aligned}
$$

from Proposition A.1 and Assumption A4′.

Combining inequalities (B.8), (B.9), (B.10), (B.11), and (B.12), we see

$$
\begin{aligned}
\mathbf{E}\left[\|\theta_n-\theta_\star\|^4|\mathcal{F}_{n-1}\right] &\leq (1-4\lambda\gamma_n+14\beta_0^2\gamma_n^2+16\beta_0^3\gamma_n^3+8\beta_0^4\gamma_n^4)\left\|\theta_{n-1}-\theta_\star\right\|^4 \\
&\quad + 4(\beta_0^2+\sigma^2)\gamma_n^2\left\|\theta_{n-1}-\theta_\star\right\|^3 + 2\sigma^2(\gamma_n^2+4\gamma_n^4)\left\|\theta_{n-1}-\theta_\star\right\|^2 \\
&\quad + 16\sigma^3\gamma_n^3\left\|\theta_{n-1}-\theta_\star\right\| + 8\sigma^4\gamma_n^4 \\
&\leq (1-4\lambda\gamma_n+14\beta_0^2\gamma_n^2+16\beta_0^3\gamma_n^3+8\beta_0^4\gamma_n^4)\left\|\theta_{n-1}-\theta_\star\right\|^4 \\
&\quad + 4(\beta_0^2+\sigma^2)\gamma_n^2\left\|\theta_{n-1}-\theta_\star\right\|^3 \\
&\quad + 2(5\sigma^2\gamma_n^2+4\gamma_n^4)\left\|\theta_{n-1}-\theta_\star\right\|^2 + 16\sigma^4\gamma_n^4,
\end{aligned} \qquad \text{(B.13)}
$$

where the last inequality follows from $16\sigma^3\gamma_n^3\|\theta_{n-1}-\theta_\star\| \leq 8\sigma^2\gamma_n^2\|\theta_{n-1}-\theta_\star\|^2 + 8\sigma^4\gamma_n^4$.

We now bound the third moment. Multiplying inequality (B.7) with

$$\begin{aligned}\|\theta_n - \theta_\star\| &\leq \|\theta_{n-1} - \theta_\star\| + \gamma_n \|\nabla \ell(Z_n, \theta_n)\| \\ &\leq \|\theta_{n-1} - \theta_\star\| + \gamma_n V_n^{1/2}\end{aligned}$$

yields

$$\begin{aligned}\|\theta_n - \theta_\star\|^3 \leq \|\theta_{n-1} - \theta_\star\|^3 &- 2\gamma_n \|\theta_n - \theta_\star\| W_n + \gamma_n^2 \|\theta_n - \theta_\star\| V_n \\ &+ \gamma_n \|\theta_n - \theta_\star\|^2 V_n^{1/2} - 2\gamma_n W_n V_n^{1/2} + \gamma_n^3 V_n^{3/2}.\end{aligned}$$

In addition to inequalities (B.9) and (B.12), we have

$$\begin{aligned}\mathbf{E}\left[V_n^{3/2}|\mathcal{F}_{n-1}\right] &= \mathbf{E}\left[\|\nabla \ell(Z_n, \theta_{n-1})\|^3 |\mathcal{F}_{n-1}\right] \leq 4\beta_0^3 \|\theta_n - \theta_\star\|^3 + 4\sigma^3 \\ \mathbf{E}\left[V_n^{1/2}|\mathcal{F}_{n-1}\right] &= \mathbf{E}\left[\|\nabla \ell(Z_n, \theta_{n-1})\| |\mathcal{F}_{n-1}\right] \leq \beta_0[\|\nabla \ell(Z_n, \theta_{n-1})\|] + \mathbf{E}\|\nabla \ell(Z_n, \theta_\star)\| \\ &\leq \beta_0[\|\nabla \ell(Z_n, \theta_{n-1})\|] + \sigma \\ \mathbf{E}\left[-W_n V_n^{1/2}|\mathcal{F}_{n-1}\right] &= -\mathbf{E}\left[(\theta_{n-1} - \theta_\star)^T \nabla \ell(Z_n, \theta_n) \|\nabla \ell(Z_n, \theta_{n-1})\| |\mathcal{F}_{n-1}\right] \\ &\leq \|\theta_{n-1} - \theta_\star\| \mathbf{E}\left[\|\nabla \ell(Z_n, \theta_n)\| \|\nabla \ell(Z_n, \theta_{n-1})\| |\mathcal{F}_{n-1}\right] \\ &\leq \|\theta_{n-1} - \theta_\star\| \mathbf{E}\left[\|\nabla \ell(Z_n, \theta_{n-1})\|^2 |\mathcal{F}_{n-1}\right] \\ &\overset{(B.9)}{\leq} 2\beta_0^2 \|\theta_{n-1} - \theta_\star\|^3 + 2\sigma^2 \|\theta_{n-1} - \theta_\star\|.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{E}\left[\|\theta_n - \theta_\star\|^3 |\mathcal{F}_{n-1}\right] &\leq [1 + (\beta_0 - 2\lambda)\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3] \|\theta_n - \theta_\star\|^3 \\ &\quad + [\sigma\gamma_n + 2(\beta_0^2 + \sigma^2)\gamma_n^2] \|\theta_n - \theta_\star\|^2 + 6\sigma^2\gamma_n^2 \|\theta_n - \theta_\star\| + 4\sigma^3\gamma_n^3 \\ &\leq [1 + (\beta_0 - 2\lambda)\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3] \|\theta_n - \theta_\star\|^3 \\ &\quad + [4\sigma\gamma_n + 2(\beta_0^2 + \sigma^2)\gamma_n^2] \|\theta_n - \theta_\star\|^2 + 7\sigma^3\gamma_n^3 \qquad \text{(B.14)}\end{aligned}$$

since $6\sigma^2\gamma_n^2 \|\theta_n - \theta_\star\| \leq 3\sigma\gamma_n \|\theta_n - \theta_\star\|^2 + 3\sigma^3\gamma_n^3$.

Now, let

$$U_n = \|\theta_n - \theta_\star\|^4 + c\beta_0\gamma_{n+1} \|\theta_n - \theta_\star\|^3, \quad c = \frac{8(\beta_0^2 + \sigma^2)}{\beta_0\lambda}.$$

Then, from inequalities (B.13) and (B.14)

$$\begin{aligned}\mathbf{E}\left[U_n|\mathcal{F}_{n-1}\right] &\leq \|\theta_{n-1} - \theta_\star\|^4 [1 - 4\lambda\gamma_n + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\ &\quad + \|\theta_{n-1} - \theta_\star\|^3 [4(\beta_0^2 + \sigma^2)\gamma_n^2 + c\beta_0\gamma_{n+1}(1 + (\beta_0 - 2\lambda)\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3)] \\ &\quad + \|\theta_{n-1} - \theta_\star\|^2 [10\sigma^2\gamma_n^2 + 8\gamma_n^4 + c\beta_0\gamma_{n+1}(4\sigma\gamma_n + 2(\beta_0^2 + \sigma^2)\gamma_n^2)] \\ &\quad + 16\sigma^4\gamma_n^4 + 7c\beta_0\sigma^3\gamma_{n+1}\gamma_n^3 \\ &\leq \|\theta_{n-1} - \theta_\star\|^4 [1 - \tfrac{1}{2}\lambda\gamma_n + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\ &\quad + \|\theta_{n-1} - \theta_\star\|^3 [4(\beta_0^2 + \sigma^2)\gamma_n^2 + c\beta_0\gamma_n(1 - 2\lambda\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3)] \\ &\quad + c\beta_0^2\gamma_n^2 \|\theta_n - \theta_\star\|^3 \\ &\quad + \|\theta_{n-1} - \theta_\star\|^2 [10\sigma^2\gamma_n^2 + 8\gamma_n^4 + c\beta_0\gamma_n(4\sigma\gamma_n + 2(\beta_0^2 + \sigma^2)\gamma_n^2)] \\ &\quad + 16\sigma^4\gamma_n^4 + 7c\beta_0\sigma^3\gamma_n^4 \\ &\leq \|\theta_{n-1} - \theta_\star\|^4 [1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{3c}{4}\beta_0^{5/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\ &\quad + \|\theta_{n-1} - \theta_\star\|^3 [4(\beta_0^2 + \sigma^2)\gamma_n^2 + c\beta_0\gamma_n(1 - 2\lambda\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3)] \\ &\quad + \|\theta_{n-1} - \theta_\star\|^2 [10\sigma^2\gamma_n^2 + 8\gamma_n^4 + c\beta_0\gamma_n(4\sigma\gamma_n + 2(\beta_0^2 + \sigma^2)\gamma_n^2)]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{c}{4}\beta_0^3\gamma_n^3 + 16\sigma^4\gamma_n^4 + 7c\beta_0\sigma^3\gamma_n^4 \\
&\leq (\|\theta_{n-1}-\theta_\star\|^4 + c\beta_0\gamma_n\|\theta_n-\theta_\star\|^3) \\
&\quad \times [1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{3c}{4}\beta_0^{5/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\
&\quad + \|\theta_{n-1}-\theta_\star\|^3 [4(\beta_0^2+\sigma^2)\gamma_n^2 + c\beta_0\gamma_n(1-\lambda\gamma_n + 8\beta_0^2\gamma_n^2 + 4\beta_0^3\gamma_n^3) \\
&\qquad - c\beta_0\gamma_n(1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{3c}{4}\beta_0^{5/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4)] \\
&\quad + \|\theta_{n-1}-\theta_\star\|^2 [10\sigma^2\gamma_n^2 + 8\gamma_n^4 + c\beta_0\gamma_n(4\sigma\gamma_n + 2(\beta_0^2+\sigma^2)\gamma_n^2)] \\
&\quad + \frac{c}{4}\beta_0^3\gamma_n^3 + 16\sigma^4\gamma_n^4 + 7c\beta_0\sigma^3\gamma_n^4 \\
&\leq (\|\theta_{n-1}-\theta_\star\|^4 + c\beta_0\gamma_n\|\theta_n-\theta_\star\|^3) \\
&\quad \times [1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{3c}{4}\beta_0^{5/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\
&\quad + \|\theta_{n-1}-\theta_\star\|^3 [(4(\beta_0^2+\sigma^2) - \tfrac{1}{2}c\beta_0\lambda)\gamma_n^2] \\
&\quad + \|\theta_{n-1}-\theta_\star\|^2 [10\sigma^2\gamma_n^2 + 8\gamma_n^4 + c\beta_0\gamma_n(4\sigma\gamma_n + 2(\beta_0^2+\sigma^2)\gamma_n^2)] \\
&\quad + \frac{c}{4}\beta_0^3\gamma_n^3 + 16\sigma^4\gamma_n^4 + 7c\beta_0\sigma^3\gamma_n^4 \\
&= U_{n-1}[1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{6(\beta_0^2+\sigma^2)}{\lambda}\beta_0^{2/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4] \\
&\quad + \|\theta_{n-1}-\theta_\star\|^2 [(10\sigma^2 + \tfrac{32\sigma(\beta_0^2+\sigma^2)}{\lambda})\gamma_n^2 + \tfrac{16(\beta_0^2+\sigma^2)^2}{\lambda}\gamma_n^3 + 8\gamma_n^4] \\
&\quad + \tfrac{2(\beta_0^2+\sigma^2)}{\lambda}\beta_0^2\gamma_n^3 + (16\sigma^4 + \tfrac{56(\beta_0^2+\sigma^2)}{\lambda}\sigma^3)\gamma_n^4,
\end{aligned}
$$

where the third inequality is due to Young's inequality

$$
c\beta_0^2\gamma_n^2\|\theta_{n-1}-\theta_\star\|^3 \leq \tfrac{3c}{4}\beta_0^{5/3}\gamma_n^{5/3}\|\theta_{n-1}-\theta_\star\|^4 + \tfrac{c}{4}\beta_0^3\gamma_n^3.
$$

Therefore,

$$
\begin{aligned}
\mathbf{E}[U_n] &\leq [1 - \tfrac{1}{2}\lambda\gamma_n + \tfrac{6(\beta_0^2+\sigma^2)}{\lambda}\beta_0^{2/3}\gamma_n^{5/3} + 14\beta_0^2\gamma_n^2 + 16\beta_0^3\gamma_n^3 + 8\beta_0^4\gamma_n^4]\,\mathbf{E}[U_{n-1}] \\
&\quad + [(10\sigma^2 + \tfrac{32\sigma(\beta_0^2+\sigma^2)}{\lambda})\gamma_n^2 + \tfrac{16(\beta_0^2+\sigma^2)^2}{\lambda}\gamma_n^3 + 8\gamma_n^4]\,\mathbf{E}\|\theta_{n-1}-\theta_\star\|^2 \\
&\quad + \tfrac{2(\beta_0^2+\sigma^2)}{\lambda}\beta_0^2\gamma_n^3 + (16\sigma^4 + \tfrac{56(\beta_0^2+\sigma^2)}{\lambda}\sigma^3)\gamma_n^4
\end{aligned}
$$

Since from Theorem 4.1

$$
\mathbf{E}\|\theta_n-\theta_\star\|^2 \leq (K_1 + \|\theta_0-\theta_\star\|^2 + D_{n_0})n^{-\gamma} = K_2\gamma_n, \quad K_2 = (K_1 + \|\theta_0-\theta_\star\|^2 + D_{n_0})/\gamma_1
$$

and $\gamma_{n-1} \leq 2\gamma_n$ (Proposition A.2),

$$
\mathbf{E}[U_n] \leq (1 - \tfrac{1}{2}\lambda\gamma_n + C_0\gamma_n^{5/3})\,\mathbf{E}[U_{n-1}] + C_1\gamma_n^3,
$$

where

$$
\begin{aligned}
C_0 &= \tfrac{6(\beta_0^2+\sigma^2)}{\lambda}\beta_0^{2/3} + 14\beta_0^2\gamma_1^{1/3} + 16\beta_0^3\gamma_1^{4/3} + 8\beta_0^4\gamma_1^{7/3} \\
&\quad + 2K_2[(10\sigma^2 + \tfrac{32\sigma(\beta_0^2+\sigma^2)}{\lambda})\gamma_1^{1/3} + \tfrac{16(\beta_0^2+\sigma^2)^2}{\lambda}\gamma_1^{4/3} + 8\gamma_1^{7/3}] \\
C_1 &= \tfrac{2(\beta_0^2+\sigma^2)}{\lambda}\beta_0^2 + (16\sigma^4 + \tfrac{56(\beta_0^2+\sigma^2)}{\lambda}\sigma^3)\gamma_1.
\end{aligned}
$$

It follows from Corollary A.1 that

$$
\mathbf{E}[U_n] \leq \tilde{K}_1 n^{-2\gamma} + \exp\Big(\nu(1+\tfrac{\lambda\gamma_1}{2})\phi_{\frac{5}{3}\gamma}(n) - \tfrac{1}{2}\log(1+\tfrac{\lambda\gamma_1}{2})\phi_\gamma(n)\Big)(U_0 + \tilde{D}_{\tilde{n}_0}),
$$

where $\nu = C_0\gamma_1^{5/3}$ and the $\tilde{n}_0$ is as given in equation (A.21). The other constants are as given in equation (A.22). Noting $\mathbf{E}\|\theta_n-\theta_\star\|^4 \leq \mathbf{E}[U_n]$ completes the proof. □