
Gaussian Approximation and Multiplier Bootstrap for Stochastic Gradient Descent

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 In this paper, we establish the non-asymptotic validity of the multiplier bootstrap
2 procedure for constructing the confidence sets using the Stochastic Gradient De-
3 scent (SGD) algorithm. Under appropriate regularity conditions, our approach
4 avoids the need to approximate the limiting covariance of Polyak-Ruppert SGD
5 iterates, which allows us to derive approximation rates in convex distance of order
6 up to $1/\sqrt{n}$. Notably, this rate can be faster than the one that can be proven in
7 the Polyak-Juditsky central limit theorem. To our knowledge, this provides the
8 first fully non-asymptotic bound on the accuracy of bootstrap approximations in
9 SGD algorithms. Our analysis builds on the Gaussian approximation results for
10 nonlinear statistics of independent random variables.

11 1 Introduction

12 Stochastic Gradient Descent (SGD) is a widely used first-order optimization method that is well
13 suited for large data sets and online learning. The algorithm has attracted significant attention; see
14 [34, 31, 27, 26, 8]. SGD aims to solve the optimization problem:

$$f(\theta) \rightarrow \min_{\theta \in \mathbb{R}^d}, \quad \nabla f(\theta) = \mathbb{E}_{\xi \sim \mathbb{P}_\xi} [F(\theta, \xi)], \quad (1)$$

15 where ξ is a random variable defined on a measurable space (Z, \mathcal{Z}) . Instead of the exact gradient
16 $\nabla f(\theta)$, the algorithm accesses only unbiased stochastic estimates $F(\theta, \xi)$.

17 Throughout this work, we focus on the case of strongly convex objective functions and denote by θ^*
18 the unique minimizer of (1). The iterates θ_k , $k \in \mathbb{N}$, generated by SGD follow the recursive update:

$$\theta_{k+1} = \theta_k - \alpha_{k+1} F(\theta_k, \xi_{k+1}), \quad \theta_0 \in \mathbb{R}^d, \quad (2)$$

19 where $\{\alpha_k\}_{k \in \mathbb{N}}$ is a sequence of step sizes (or learning rates), which may be either diminishing
20 or constant, and $\{\xi_k\}_{k \in \mathbb{N}}$ is an i.i.d. sequence sampled from \mathbb{P}_ξ . Theoretical properties of SGD,
21 particularly in the convex and strongly convex settings, have been extensively studied; see, e.g.,
22 [28, 26, 8, 23]. Many optimization algorithms build upon the recurrence (2) to accelerate the
23 convergence of the sequence θ_k to θ^* . Notable examples include momentum acceleration [32],
24 variance reduction techniques [12, 39], and averaging methods. In this work, we focus on Polyak-
25 Ruppert averaging, originally proposed in [36] and [31], which improves convergence by averaging
26 the SGD iterates (2). Specifically, the estimator is defined as

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=0}^{n-1} \theta_i, \quad n \in \mathbb{N}. \quad (3)$$

27 It has been established (see [31, Theorem 3]) that under appropriate conditions on the objective
28 function f , the noisy gradient estimates F , and the step sizes α_k , the sequence of averaged iterates

29 $\{\bar{\theta}_n\}_{n \in \mathbb{N}}$ is asymptotically normal:

$$\sqrt{n}(\bar{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_\infty), \quad (4)$$

30 where \xrightarrow{d} denotes convergence in distribution, and $\mathcal{N}(0, \Sigma_\infty)$ is a zero-mean Gaussian distribution
31 with covariance matrix Σ_∞ , defined later in Section 2.2. This result raises two key questions:

- 32 (i) what is the rate of convergence in (4)?
33 (ii) how can (4) be leveraged to construct confidence sets for θ^* , given that Σ_∞ is unknown in
34 practice?

35 In our paper we aim to answer both questions. To quantify convergence rates in (4), we employ
36 convex distance, which is defined for random vectors $X, Y \in \mathbb{R}^d$ as

$$d_C(X, Y) = \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}(X \in B) - \mathbb{P}(Y \in B)|,$$

37 where $\mathcal{C}(\mathbb{R}^d)$ denotes the collection of convex subsets of \mathbb{R}^d . The authors of [42] derive Berry-
38 Esseen-type bounds for $d_C(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*), \mathcal{N}(0, I_d))$, where Σ_n is the covariance matrix of
39 the linearized counterpart of (2), see precise definitions (13). We complement this result with the
40 rates of convergence in (4). Interestingly, we also provide the lower bounds on the convex distance
41 $d_C(\sqrt{n}(\bar{\theta}_n - \theta^*), \mathcal{N}(0, \Sigma_\infty))$, which indicates, that for some choice of step sizes α_k in (2), the
42 normal approximation by $\mathcal{N}(0, \Sigma_\infty)$ is less accurate, compared to normal approximation with other
43 covariance matrix, in particular, with Σ_n . This effect has been previously observed in the bootstrap
44 method for i.i.d. observations without the context of gradients methods, see [41, Theorem 3.11].

45 One of the popular approaches for solving (ii) is based on the *plug-in* methods [11, 9], which aim to
46 construct an estimator $\hat{\Sigma}_n$ of Σ_∞ directly. These methods often provide a non-asymptotic bounds on
47 the closeness $\hat{\Sigma}_n$ is to Σ_∞ , often in terms of $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|]$. At the same time, the analysis of this
48 methods typically bypass the item (i) and the issues related with the rate of convergence in (4). In our
49 paper we suggest, to the best of our knowledge, the first fully non-asymptotic analysis of procedure
50 for constructing the confidence intervals, based on the bootstrap approach [15, 16], which avoids the
51 direct approximation of Σ_∞ , moreover, theoretical analysis of the underlying procedure together with
52 the results on normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ from (i) shows that the same approximation rate
53 can not be achieved by the plug-in methods, at least for some range of step sizes α_k in (2). Our key
54 contributions are as follows:

- 55 • We establish the non-asymptotic validity of the multiplier bootstrap procedure introduced in
56 [16]. Under appropriate regularity conditions, our bounds imply that the quantiles of the
57 exact distribution of $\sqrt{n}(\bar{\theta}_n - \theta^*)$ can be approximated, up to logarithmic factors, at a rate of
58 $n^{-\gamma/2}$ for step sizes of the form $\alpha_k = c_0/(k + k_0)^\gamma$, $\gamma \in (1/2, 1)$. To our knowledge, this
59 provides the first fully non-asymptotic bound on the accuracy of bootstrap approximations
60 in SGD algorithms. Notably, this rate can be faster than the one that we can prove in (4).
61 Our rates improve upon recent works [38, 46], which addressed the convergence rate in
62 similar procedures for the LSA algorithm.
- 63 • Our analysis of the multiplier bootstrap procedure reveals an interesting property: unlike
64 plug-in estimators, the validity of the bootstrap method does not directly depend on approxi-
65 mating $\sqrt{n}(\bar{\theta}_n - \theta^*)$ by $\mathcal{N}(0, \Sigma_\infty)$. Instead, it requires approximating $\mathcal{N}(0, \Sigma_n)$ for some
66 matrix Σ_n . The structure of Σ_n and its associated convergence rates play a central role in
67 our present analysis, both for convergence rate in (4) and non-asymptotic bootstrap validity.
68 Precise definitions are provided in Section 2.2.
- 69 • We analyze the Polyak-Ruppert averaged SGD iterates (3) for strongly convex minimization
70 problems and establish Gaussian approximation rates in (4) in terms of the convex distance.
71 Specifically, we show that the approximation rate $d_C(\sqrt{n}(\bar{\theta}_n - \theta^*), \mathcal{N}(0, \Sigma_\infty))$ is of order
72 $n^{-1/4}$ when using the step size $\alpha_k = c_0/(k + k_0)^{3/4}$ with a suitably chosen α_0 . Our
73 result is based on the techniques of [42] and [46]. We also provide the lower bound,
74 which indicate that our rate of normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ is tight in the regime
75 $\alpha_k = c_0/(k + k_0)^\gamma$ with $\gamma \geq 3/4$.

76 **Notations.** Throughout this paper, we use the following notations. For a matrix $A \in \mathbb{R}^{d \times d}$ and a
77 vector $x \in \mathbb{R}^d$, we denote by $\|A\|$ and $\|x\|$ their spectral norm and Euclidean norm, respectively. We
78 also write $\|A\|_F$ for the Frobenius norm of matrix A . Given a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we write $\nabla f(\theta)$
79 and $\nabla^2 f(\theta)$ for its gradient and Hessian at a point θ . Additionally, we use the standard abbreviations
80 "i.i.d." for "independent and identically distributed" and "w.r.t." for "with respect to".

Literature review. Asymptotic properties of the SGD algorithm, including the asymptotic normality of the estimator $\bar{\theta}_n$ and its almost sure convergence, have been extensively studied for smooth and strongly convex minimization problems [31, 22, 7]. Optimal mean-squared error (MSE) bounds for $\theta_n - \theta^*$ and $\bar{\theta}_n - \theta^*$ were established in [27] for smooth and strongly convex objectives, and later refined in [26]. The case of constant-step size SGD for strongly convex problems has been analyzed in depth in [14]. High-probability bounds for SGD iterates were obtained in [33] and later extended in [20]. Both works address non-smooth and strongly convex minimization problems.

It is important to note that the results discussed above do not directly imply convergence rates for $\sqrt{n}(\bar{\theta}_n - \theta^*)$ to $\mathcal{N}(0, \Sigma_\infty)$ in terms of $d_C(\cdot, \cdot)$ or the Kantorovich–Wasserstein distance. Among the relevant contributions in this direction, we highlight recent works [44, 38, 46], which provide quantitative bounds on the convergence rate in (4) for iterates of the temporal difference learning algorithm and general linear stochastic approximation (LSA) schemes. However, these algorithms do not necessarily correspond to SGD with a quadratic objective f , as the system matrix in LSA is not necessarily symmetric. Non-asymptotic convergence rates of order $1/\sqrt{n}$ in a smooth Wasserstein distance were established in [2]. Recent paper [1] provide Berry-Essen bounds for last iterate of SGD for high-dimensional linear regression of order up to $n^{-1/4}$.

Bootstrap methods for i.i.d. observations were first introduced in [15]. In the context of SGD methods, [16] proposed the multiplier bootstrap approach for constructing confidence intervals for θ^* and established its asymptotic validity. The same algorithm, with non-asymptotic guarantees, was analyzed in [38] for the LSA algorithm, obtaining rate $n^{-1/4}$ when approximating quantiles of the exact distribution of $\sqrt{n}(\bar{\theta}_n - \theta^*)$.

Popular group of methods for constructing confidence sets for θ^* is based on estimating the asymptotic covariance matrix Σ_∞ . Plug-in and batch-mean estimators for Σ_∞ attracted lot of attention, see [11, 9, 10], especially in the setting when the stochastic estimates of Hessian are available. The latter two papers focused on learning with contextual bandits. Estimates for Σ_∞ based on batch-mean method and its online modification were considered in [11] and [49]. The authors in [25] considered the asymptotic validity of the plug-in estimator for Σ_∞ in the local SGD setting. [47] refined the validity guarantees for both the multiplier bootstrap and batch-mean estimates of Σ_∞ for nonconvex problems. However, these papers typically provide recovery rates Σ_∞ , but only show asymptotic validity of the proposed confidence intervals. A notable exception is the recent paper [46], where the temporal difference (TD) learning algorithm was studied. The authors of [46] provided purely non-asymptotic analysis of their procedure, obtaining the approximation rate $n^{-1/3}$ for quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^*)$.

2 Main results

This section establishes the nonasymptotic validity of the multiplier bootstrap method proposed in [16]. We focus on smooth and strongly convex minimization problems, following the framework established in [26], [2] and [42]. The underlying procedure is based on perturbing the trajectory (2). We restate the procedure for the sake of clarity. Let $\mathcal{W}^{n-1} = \{w_\ell\}_{1 \leq \ell \leq n-1}$ be i.i.d. random variables with distribution \mathbb{P}_w , each with mean $\mathbb{E}[w_1] = 1$ and variance $\text{Var}[w_1] = 1$. Assume \mathcal{W}^{n-1} is independent of $\Xi^{n-1} = \{\xi_\ell\}_{1 \leq \ell \leq n-1}$. We then use \mathcal{W}^{n-1} to construct randomly perturbed SGD trajectories, following the same recursive structure as the primary sequence

$$\begin{aligned} \theta_k^b &= \theta_{k-1}^b - \alpha_k w_k \{ \nabla f(\theta_{k-1}^b) + g(\theta_{k-1}^b, \xi_k) + \eta(\xi_k) \}, \quad k \geq 1, \quad \theta_0^b = \theta_0, \\ \bar{\theta}_n^b &= n^{-1} \sum_{k=0}^{n-1} \theta_k^b, \quad n \geq 1. \end{aligned} \tag{5}$$

Note that, when generating different weights w_k , we can draw samples from the conditional distribution of $\bar{\theta}_n^b$ given the data Ξ^{n-1} . We further denote $\mathbb{P}^b = \mathbb{P}(\cdot \mid \Xi^{n-1})$ and $\mathbb{E}^b = \mathbb{E}(\cdot \mid \Xi^{n-1})$.

The core principle behind the bootstrap procedure (5) is that the "bootstrap world" probabilities $\mathbb{P}^b(\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) \in B)$ are close to $\mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^*) \in B)$ for $B \in \mathcal{C}(\mathbb{R}^d)$. More formally, we say that the procedure (5) is asymptotically valid if

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} \left| \mathbb{P}^b(\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^*) \in B) \right| \tag{6}$$

converges to 0 in \mathbb{P} -probability as $n \rightarrow \infty$. This result was studied in [16] under assumptions close to the original paper [31]. While an analytical expression for $\mathbb{P}^b(\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) \in B)$ is unavailable, it

can be approximated via Monte Carlo simulations by generating M perturbed trajectories according to (5). Standard arguments (see, e.g., [40, Section 5.1]) suggest that the accuracy of this Monte Carlo approximation scales as $\mathcal{O}(M^{-1/2})$ when generating M parallel perturbed trajectories in (5).

Assumptions. We impose the following regularity conditions on the objective function f :

A1. The function f is two times continuously differentiable and L_1 -smooth on \mathbb{R}^d , i.e., there is a constant $L_1 > 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$,

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L_1 \|\theta - \theta'\|.$$

Moreover, we assume that f is μ -strongly convex on \mathbb{R}^d , that is, there exists a constant $\mu > 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that

$$(\mu/2)\|\theta - \theta'\|^2 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle.$$

A1 implies the following two-sided bound on the Hessian $\nabla^2 f(\theta)$, $\mu I_d \preceq \nabla^2 f(\theta) \preceq L_1 I_d$ for all $\theta \in \mathbb{R}^d$. We now formalize the assumptions on $F(\theta, \xi)$. Namely, we rewrite $F(\theta, \xi)$ as

$$F(\theta_{k-1}, \xi_k) = \nabla f(\theta_{k-1}) + \zeta_k,$$

where $\{\zeta_k\}_{k \in \mathbb{N}}$ is a sequence of d -dimensional random vectors. Then the SGD recursion takes form

$$\theta_k = \theta_{k-1} - \alpha_k (\nabla f(\theta_{k-1}) + \zeta_k), \quad \theta_0 \in \mathbb{R}^d. \quad (7)$$

We impose the following assumption on the noise sequence ζ_k :

A2. For each $k \geq 1$, ζ_k admits the decomposition $\zeta_k = \eta(\xi_k) + g(\theta_{k-1}, \xi_k)$, where

- (i) $\{\xi_k\}_{k=1}^{n-1}$ is a sequence of i.i.d. random variables on (Z, \mathcal{Z}) with distribution \mathbb{P}_ξ , $\eta : Z \rightarrow \mathbb{R}^d$ is a function such that $\mathbb{E}[\eta(\xi_1)] = 0$ and $\mathbb{E}[\eta(\xi_1)\eta(\xi_1)^\top] = \Sigma_\xi$. Moreover, $\lambda_{\min}(\Sigma_\xi) > 0$.
- (ii) The function $g : \mathbb{R}^d \times Z \rightarrow \mathbb{R}^d$ satisfies $\mathbb{E}[g(\theta, \xi_1)] = 0$ for any $\theta \in \mathbb{R}^d$. Moreover, there exists $L_2 > 0$ such that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that

$$\|g(\theta, \xi) - g(\theta', \xi)\| \leq L_2 \|\theta - \theta'\| \quad \text{and} \quad g(\theta^*, z) = 0 \quad \text{for all } z \in Z. \quad (8)$$

- (iii) There exist $C_{1,\xi}, C_{2,\xi} > 0$ such that \mathbb{P}_ξ -almost surely that $\|\eta(\xi)\| \leq C_{1,\xi}$ and $\sup_\theta \|g(\theta, \xi)\| \leq C_{2,\xi}$.

As an example of a sequence ζ_k satisfying conditions (i) and (ii) from A2, consider the case when the oracle function $F(\theta, \xi)$ satisfies:

- 1. $\mathbb{E}[F(\theta, \xi)] = \nabla f(\theta)$ for all $\theta \in \mathbb{R}^d$;
- 2. $\|F(\theta, \xi) - F(\theta', \xi)\| \leq L\|\theta - \theta'\|$ for all $\xi \in Z$, and $\sup_\theta |F(\theta, \xi) - F(\theta^*, \xi)| \leq c_\xi$ for some $c_\xi > 0$.

In this case, (i) and (ii) from A2 holds with $\eta(\xi) = F(\theta^*, \xi)$ and $g(\theta, \xi) = F(\theta, \xi) - F(\theta^*, \xi)$. Additionally, note that the identity (8) can be relaxed when one considers only last iterate bounds, such as $\mathbb{E}[\|\theta_k - \theta^*\|^2]$, see [26]. Item (ii) from A2 is often imposed when considering averaged iterates, see [26, Assumption H2*], and [14, 42].

The assumption (iii) from A2 is crucial to prove high-order moment bounds (20), see Lemma 15. In our proof, we closely follow the argument presented in [20, Theorem 4.1], which requires that the noise variables ζ_k be almost sure to be bounded. This setting can be generalized to the case where ζ_k is sub-Gaussian conditioned on \mathcal{F}_{k-1} with variance proxy which is uniformly bounded by a constant factor, that is, there is a constant M , such that $\mathbb{E}[\exp\{\|F(\theta, \xi_1)\|^2/M^2\}] \leq 2$ for any $\theta \in \mathbb{R}^d$. This assumption is widely considered in the literature; see [27, 21], and the remarks in [20]. However, when $\zeta_k = g(\theta_{k-1}, \xi_k) + \eta(\xi_k)$ and g is only Lipschitz w.r.t. θ , its moments will naturally scale with $\|\theta_{k-1} - \theta^*\|$, thus the sub-Gaussian bound with M not depending upon θ is unlikely to hold. Other authors who considered bounds of type (20), e.g. [33], made stronger assumption that $\sup_{\theta \in \mathbb{R}^d} \|F(\theta, \xi)\|$ is a.s. bounded. Another popular direction is to consider schemes for gradient clipping; see e.g. [37]. Unfortunately, employing such schemes change the key representation (12) that we use later in the proof of the main result (see Theorem 1). We leave further studies of clipped gradient schemes for future work. We further impose condition on the Hessian matrix $\nabla^2 f(\theta)$ at θ^* :

A3. There exist $L_3, \beta > 0$ such that for all θ with $\|\theta - \theta^*\| \leq \beta$, it holds

$$\|\nabla^2 f(\theta) - \nabla^2 f(\theta^*)\| \leq L_3 \|\theta - \theta^*\|.$$

A3 ensures that the Hessian of f is Lipschitz continuous in a neighborhood of θ^* . Similar assumptions have been previously considered in [42] and [2], as well as in other works on first-order optimization methods, see, e.g. [24]. Several studies on the non-asymptotic analysis of SGD impose stronger smoothness assumptions, such as bounded derivatives of f up to order four, see [14]. We additionally assume an almost sure co-coercivity of the stochastic gradient:

A4. The stochastic gradient $F(\theta, \xi) := \nabla f(\theta) + g(\theta, \xi) + \eta(\xi)$ is almost surely L_4 -co-coercive, that is, for any $\theta, \theta' \in \mathbb{R}^d$, it holds \mathbb{P}_ξ -almost surely that

$$L_4 \langle F(\theta, \xi) - F(\theta', \xi), \theta - \theta' \rangle \geq \|F(\theta, \xi) - F(\theta', \xi)\|^2.$$

In particular, A4 holds (see e.g. [48]), when there is a function $v(\theta, \xi)$, such that $F(\theta, \xi) = \nabla_\theta v(\theta, \xi)$, where $v(\theta, \xi)$ is convex \mathbb{P}_ξ -a.s. and L_4 -smooth. Co-coercivity is stronger than just requiring $F(\theta, \xi)$ to be monotone. We also impose an assumption on the bootstrap weights W_i used in the algorithm:

A5. There exist constants $0 < W_{\min} < W_{\max} < +\infty$, such that $W_{\min} \leq w_1 \leq W_{\max}$ a.s.

The original paper [16] also considered positive bootstrap weights w_i . We have to impose boundedness of w_i due to our high-probability bound on Lemma 15. A particular example of a distribution satisfying A5 is provided in Appendix E.1. We also consider the following bound for step sizes α_k and sample size n :

A6. Let $\alpha_k = c_0 \{k_0 + k\}^{-\gamma}$, where $\gamma \in (1/2, 1)$, an c_0 satisfies $c_0 W_{\max} \max(2L_4, \mu) \leq 1$ and $k_0 \geq (\frac{2\gamma}{\mu c_0 W_{\min}})^{1/(1-\gamma)}$.

A7. Number of observations n satisfies $n \geq e^3$ and $\frac{n}{\log(2dn)} \geq \max(1, \frac{(20C_{Q,\xi}C_\Sigma^2)^2}{9})$, where the constants $C_{Q,\xi}$ and C_Σ are defined in (61) and (27), respectively.

The particular bound on k_0 in A6 appears due to the high-order moment bounds (see Lemma 15 in appendix). We note that it is possible to remove the co-coercivity assumption A4, but at the price of slightly stronger constraints on c_0 above. We discuss the bound on the number of observations imposed in A7 later in the proof of Theorem 1.

2.1 Non-asymptotic multiplier bootstrap validity

Theorem 1. Assume A1 - A7. Then with \mathbb{P} -probability at least $1 - 2/n$, it holds

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^*) \in B)| \leq \frac{C_1 \sqrt{\log n}}{n^{1/2}} + \frac{C_2 \log n}{n^{\gamma-1/2}} + \frac{C_3 (\log n)^{3/2}}{n^{\gamma/2}},$$

where C_1, C_2 and C_3 are given in Appendix E.7, equation (64).

Remark 1. It is possible to prove the result of Theorem 1 for the step size $\alpha_k = c_0/(k + k_0)$. The required Gaussian approximation result with the covariance matrix Σ_n is proved in [42], and we expect that the only difference with Theorem 1 will occur in extra $\log n$ factors in the corresponding bound and slightly different conditions on c_0 and k_0 in A6.

Proof sketch of Theorem 1. The proof of non-asymptotic bootstrap validity is based on the Gaussian approximation performed both in the "real" world and bootstrap world together with an appropriate Gaussian comparison inequality:

$$\begin{array}{ccc} \text{Real world:} & \sqrt{n}(\bar{\theta}_n - \theta^*) & \xleftarrow{\text{Gaussian approximation, Th. 2}} \Sigma^{1/2} Y \sim \mathcal{N}(0, \Sigma) \\ & & \updownarrow \text{Gaussian comparison, Lem. 19} \\ \text{Bootstrap world:} & \sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) & \xleftarrow{\text{Gaussian approximation, Th. 3}} \{\Sigma^b\}^{1/2} Y^b \sim \mathcal{N}(0, \Sigma^b). \end{array}$$

Here Σ and Σ^b are some covariance matrices to be chosen later. In order to understand where the Gaussian approximation comes from, we consider the process of linearization of statistics $\sqrt{n}(\bar{\theta}_n - \theta^*)$ and $\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n)$. We provide details for $\sqrt{n}(\bar{\theta}_n - \theta^*)$, and give similar derivations for $\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n)$ in Section 2.3. Denote $G = \nabla^2 f(\theta^*)$. We expand $\sqrt{n}(\bar{\theta}_n - \theta^*)$ into a weighted sum of independent random vectors, along with the remaining terms of smaller order. By the Newton-Leibniz formula, we obtain

$$\nabla f(\theta) = G(\theta - \theta^*) + H(\theta), \quad (9)$$

where $H(\theta) = \int_0^1 (\nabla^2 f(\theta^* + t(\theta - \theta^*)) - G)(\theta - \theta^*) dt$. Note that $H(\theta)$ is of the order $\|\theta - \theta^*\|^2$ (see Lemma 5). The recursion for the SGD algorithm error (7) can be expressed as

$$\theta_k - \theta^* = (I_d - \alpha_k G)(\theta_{k-1} - \theta^*) - \alpha_k(\eta(\xi_k) + g(\theta_{k-1}, \xi_k) + H(\theta_{k-1})) . \quad (10)$$

For $i \in \{0, \dots, n-1\}$, we define the matrices

$$Q_i = \alpha_i \sum_{j=i}^{n-1} \prod_{k=i+1}^j (I_d - \alpha_k G) , \quad (11)$$

where empty products are defined to be equal to I_d by convention. Then taking average of (10) and rearranging the terms, we obtain the following expansion:

$$\sqrt{n}(\bar{\theta}_n - \theta^*) = W + D , \quad W = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i \eta(\xi_i), \quad D = \sqrt{n}(\bar{\theta}_n - \theta^*) - W . \quad (12)$$

Note that W is a weighted sum of i.i.d. random vectors with mean zero and covariance matrix

$$\Sigma_n = n^{-1} \sum_{k=1}^{n-1} Q_k \Sigma_\xi Q_k^\top , \quad (13)$$

and D is the remainder term which is defined in Appendix C, equation (40). Furthermore, in Appendix D.1 we show that Q_i may be approximated by $G^{-\top}$ and Σ_n approximates

$$\Sigma_\infty = G^{-1} \Sigma_\xi G^{-\top} .$$

We expect that the summand D does not significantly distort the asymptotic distribution for the linear statistic W , which should be Gaussian by virtue of the central limit theorem. An important question is the choice of the approximating Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = \Sigma_n$ or Σ_∞ as well their bootstrap counterpart Σ^b . This choice is instrumental in the sense that it does not change the procedure (5), but only affects the rates in (6). The authors of [16] choose the approximation with $\mathcal{N}(0, \Sigma_\infty)$ for their asymptotic analysis. A similar approach was considered in [38, Theorem 3] for the LSA algorithm setting. However, as it will be shown later in Theorem 4, this choice implies that the rate of normal approximation in (6) is not faster than $n^{-1/4}$. At the same time, Theorem 2 and Theorem 3 below demonstrate that we can achieve approximation rates of up to $n^{-1/2}$ by selecting $\Sigma = \Sigma_n$ in the diagram 2.1, and its bootstrap-world counterpart in the Gaussian approximation. To finish the proof, it remains to apply the Gaussian comparison inequality; see Lemma 19. Detailed proof of Theorem 1 is provided in Appendix E. \square

Discussion. In [38] a counterpart of Theorem 1 was established with an approximation rate of the order $n^{-1/4}$ up to logarithmic factors for the setting of the LSA algorithm. The obtained rate is suboptimal, since the authors have chosen $\mathcal{N}(0, \Sigma_\infty)$ for Gaussian approximation when showing bootstrap validity. A recent paper [46] improved this rate to $n^{-1/3}$ for the temporal learning (TD) procedure with linear function approximation. The algorithm they considered is based on the direct estimate of Σ_∞ , yielding a rate of order $n^{-1/3}$ when approximating the quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^*)$, see [46, Theorem 3.4 and 3.5]. The authors in [11] constructed a plug-in estimator $\hat{\Sigma}_n$ of Σ_∞ and showed guarantees of the form $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|] \lesssim Cn^{-\gamma/2}$, $\gamma \in (1/2, 1)$ under weaker assumptions than those considered in the current section. At the same time, approximating quantiles of $\sqrt{n}(\bar{\theta}_n - \theta^*)$ with the method of [11] would require one more step - a Berry-Esseen type bound on the rate of approximation of $\sqrt{n}(\bar{\theta}_n - \theta^*)$ with $\mathcal{N}(0, \Sigma_\infty)$. As we show in Theorem 4, this rate vanishes as $\gamma \rightarrow 1$, which introduces an additional trade-off to the potential analysis of the plug-in procedures based on estimating Σ_∞ . This effect highlights the fundamental difference between the multiplier bootstrap approach and the plug-in approach of [11].

Moreover, we highlight that in-expectation bound $\mathbb{E}[\|\hat{\Sigma}_n - \Sigma_\infty\|]$, which are typically studied in literature for plug-in estimates [11, 35], are not sufficient to prove the analogue of the Gaussian comparison result Lemma 1 for $\mathcal{N}(0, \hat{\Sigma}_n)$ and $\mathcal{N}(0, \Sigma_\infty)$ on the set with large \mathbb{P} -probability. Thus, the complete non-asymptotic analysis of the confidence sets constructed with the plug-in procedure, remains an open problem.

2.2 Gaussian approximation in the real world

For results of this section, assumptions A2 and A6 can be relaxed. We impose a family of assumptions, denoted as A8(p) with $p \geq 2$, on the noise sequence ζ_k , and A9 on the step sizes α_k :

A 8 (p). Conditions (i) and (ii) from A 2 holds. Moreover, there exists $\sigma_p > 0$ such that $\mathbb{E}^{1/p}[\|\eta(\xi_1)\|^p] \leq \sigma_p$.

255 **A9.** Suppose that $\alpha_k = c_0/(k_0 + k)^\gamma$, where $\gamma \in (1/2, 1)$, $k_0 \geq 1$, and c_0 satisfies $2c_0 L_1 \leq 1$.

256 Note that A6 implies A9, as well as A2 implies A8(p) for any $p \geq 2$. In the main result of this section
 257 we provide the Gaussian approximation result for $\sqrt{n}(\bar{\theta}_n - \theta^*)$ with $\mathcal{N}(0, \Sigma_n)$, which refines the
 258 bounds obtained in [42, Theorem 3.4] and is instrumental for further studies of normal approximation
 259 with $\mathcal{N}(0, \Sigma_\infty)$ in Section 2.4.

260 **Theorem 2.** Assume A1, A3, A8(4), A9. Then, with $Y \sim \mathcal{N}(0, I_d)$, it holds that

$$\mathrm{d}_{\mathcal{C}}(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*), Y) \leq \frac{C_4}{\sqrt{n}} + \frac{C_5}{n^{\gamma-1/2}} + \frac{C_6}{n^{\gamma/2}}, \quad (14)$$

261 where C_4, C_5, C_6 are given in Appendix C, equation (41). Moreover, since Σ_n is non-degenerate,
 262 and an image of a convex set under non-degenerate linear mapping is a convex set, we have

$$\mathrm{d}_{\mathcal{C}}(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*), Y) = \mathrm{d}_{\mathcal{C}}(\sqrt{n}(\bar{\theta}_n - \theta^*), \Sigma_n^{1/2}Y).$$

263 **Remark 2.** When $\gamma \rightarrow 1$, the correction terms above scale as $\mathcal{O}(1/\sqrt{n})$, yielding the overall
 264 approximation rate that approaches $1/\sqrt{n}$. Expressions for C_4, C_5, C_6 from Theorem 2 depend
 265 upon the problem dimension d , parameters specified in A1 - A8(4)- A3-A9. Moreover, C_5 depends
 266 upon $\|\theta_0 - \theta^*\|$. When $\gamma \in (0, 1)$, we have that $1/n^{\gamma/2} < 1/n^{\gamma-1/2}$, thus, the term $C_5/n^{\gamma-1/2}$
 267 dominates. We prefer to keep both terms in (14), since they are responsible for the moments of
 268 statistics $\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i H(\theta_{i-1})$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i g(\theta_{i-1}, \xi_i)$, respectively. The first of them has
 269 non-zero mean, since $H(\theta_{i-1})$ is quadratic in $\|\theta_i - \theta^*\|^2$. When using constant step size SGD, one
 270 can correct this term using the Richardson-Romberg technique [14, 43], however, it is unclear if this
 271 type of ideas can be generalized for diminishing step size.

272 *Proof sketch of Theorem 2.* The decomposition (12) represents a particular instance of the general
 273 problem of Gaussian approximation for nonlinear statistics of the form $\sqrt{n}(\bar{\theta}_n - \theta^*)$, where the
 274 estimator is expressed as the sum of linear and nonlinear components. To establish the Gaussian ap-
 275 proximation result, we adapt the arguments from [42], which can be stated as follows. Let X_1, \dots, X_n
 276 be independent random variables taking values in some space \mathcal{X} , and let $T = T(X_1, \dots, X_n)$ be a
 277 general d -dimensional statistic that can be decomposed as

$$W := W(X_1, \dots, X_n) = \sum_{\ell=1}^n Z_\ell, \quad D := D(X_1, \dots, X_n) = T - W.$$

278 Here, we define $Z_\ell = r_\ell(X_\ell)$, where $r_\ell : \mathcal{X} \rightarrow \mathbb{R}^d$ is a Borel-measurable function. The term D
 279 represents the nonlinear component and is treated as an error term, assumed to be "small" relative
 280 to W in an appropriate sense. Suppose that $\mathbb{E}[Z_\ell] = 0$ and that the Z_ℓ is normalized in such a way
 281 that $\sum_{\ell=1}^n \mathbb{E}[Z_\ell Z_\ell^\top] = I_d$ holds. Let $\Upsilon_n = \sum_{\ell=1}^n \mathbb{E}[\|Z_\ell\|^3]$. Then, for $Y \sim \mathcal{N}(0, I_d)$, the following
 282 bound holds:

$$\mathrm{d}_{\mathcal{C}}(T, Y) \leq 259d^{1/2}\Upsilon_n + 2\mathbb{E}[\|W\|\|D\|] + 2\sum_{\ell=1}^n \mathbb{E}[\|Z_\ell\|\|D - D^{(\ell)}\|], \quad (15)$$

283 where $D^{(\ell)} = D(X_1, \dots, X_{\ell-1}, X'_\ell, X_{\ell+1}, \dots, X_n)$ and X'_ℓ is an independent copy of X_ℓ . This
 284 result follows from [42, Theorem 2.1]. Furthermore, this bound can be extended to the case where
 285 $\sum_{\ell=1}^n \mathbb{E}[Z_\ell Z_\ell^\top] = \Sigma \succ 0$, as detailed in [42, Corollary 2.3]. In order to apply (15), we let $X_i = \xi_i$,
 286 $Z_\ell = h(X_\ell)$, ξ'_i be an i.i.d. copy of ξ_i . Then we need to upper bound $\mathbb{E}^{1/2}[\|D(\xi_1, \dots, \xi_{n-1})\|^2]$ and
 287 $\mathbb{E}^{1/2}[\|D - D'_i\|^2]$, respectively. Detailed proof is given in Appendix C. \square

288 2.3 Gaussian approximation in the bootstrap world

289 In the main result of this section, we study the Gaussian approximation result for $\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n)$ with
 290 appropriate normal distribution with respect to \mathbb{P}^b . Despite this result is similar in its nature with the
 291 one of Theorem 2, it requires to handle some significant challenges that arises when working in the
 292 "bootstrap world". Our first steps are the same as in (10) and (5):

$$\begin{aligned} \theta_k^b - \theta_k &= (I - \alpha_k G)(\theta_{k-1}^b - \theta_{k-1}) \\ &\quad - \alpha_k (H(\theta_{k-1}^b) + g(\theta_{k-1}^b, \xi_k) - H(\theta_{k-1}) - g(\theta_{k-1}, \xi_k)) \\ &\quad - \alpha_k (w_k - 1)(G(\theta_{k-1}^b - \theta^*) + \eta(\xi_k) + g(\theta_{k-1}^b, \xi_k) + H(\theta_{k-1}^b)). \end{aligned} \quad (16)$$

293 Taking an average of (16) and rearranging the terms, we obtain a counterpart of (12):

$$\begin{aligned} \sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) &= W^b + D^b, \\ W^b &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (w_i - 1) Q_i \eta(\xi_i), \quad D^b = \sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) - W^b. \end{aligned} \quad (17)$$

294 Here W^b is a weighted sum of i.i.d. random variables Ξ^{n-1} , such that $\mathbb{E}^b[W^b] = 0$ and

$$\mathbb{E}^b[W^b \{W^b\}^\top] := \Sigma_n^b = n^{-1} \sum_{i=1}^{n-1} Q_i \eta(\xi_i) \eta(\xi_i)^\top Q_i^\top,$$

295 and D^b is a non-linear statistic of Ξ^{n-1} . The principal difficulty arises when considering the
296 conditional distribution of $\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n)$ given the data Ξ^{n-1} . In fact, the approach of [42] would
297 require to control the second moments of D^b and $D^b - \{D^b\}^{(i)}$ with respect to a bootstrap measure
298 \mathbb{P}^b , on the high-probability event with respect to a measure \mathbb{P} . At the same time, we loose a martingale
299 structure of the summands in D^b , unless we condition on the extended filtration

$$\tilde{\mathcal{F}}_i = \sigma(w_1, \dots, w_i, \xi_1, \dots, \xi_i), \quad 1 \leq i \leq n-1. \quad (18)$$

300 Therefore, it is not clear if we can directly apply the approach of [42] discussed in Section 2.2. Instead,
301 we have to use the linearization approach based on the high-order moment bounds for the remainder
302 term D^b (see Proposition 3 in Appendix E). This justifies the strong bounded noise assumption A2,
303 that we had to impose. We state the main result of this section below:

304 **Theorem 3.** Assume A1 - A7. Then with \mathbb{P} -probability at least $1 - 2/n$, it holds

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\sqrt{n}\{\Sigma_n^b\}^{-1/2}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}^b(Y^b \in B)| \leq \frac{M_{3,1}^b}{n^{1/2}} + \frac{M_{3,2}^b \log n}{n^{\gamma-1/2}} + \frac{M_{3,3}^b \log^{3/2} n}{n^{\gamma/2}},$$

305 where $\{M_{3,i}^b\}_{i=1}^3$ are defined in Appendix E.6, equation (62).

306 *Proof sketch of Theorem 3.* We apply the bound

$$\begin{aligned} &\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\{\Sigma_n^b\}^{-\frac{1}{2}}(W^b + D^b) \in B) - \mathbb{P}^b(Y^b \in B)| \\ &\leq \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\{\Sigma_n^b\}^{-\frac{1}{2}}W^b \in B) - \mathbb{P}^b(Y^b \in B)| + 2c_d(\mathbb{E}^b[\|\{\Sigma_n^b\}^{-\frac{1}{2}}D^b\|^p])^{\frac{1}{1+p}}, \end{aligned} \quad (19)$$

307 where $c_d \leq 4d^{1/4}$ is the isoperimetric constant of the class of convex sets. The proof of (19) is
308 provided in Proposition 3 in Appendix E. We can control $\mathbb{E}[\|D^b\|^p]$ by Burkholder's inequality, where
309 \mathbb{E} denotes the expectation w.r.t. the product measure $\mathbb{P}_\xi^{\otimes n} \otimes \mathbb{P}_w^{\otimes n}$. Then we proceed with Markov's
310 inequality to obtain \mathbb{P} -high-probability bounds on the behavior of $\mathbb{E}^b[\|D^b\|^p]$. This result requires
311 us to provide bounds for

$$\mathbb{E}^{1/p}[\|\theta_k - \theta^*\|^p] \quad \text{and} \quad \mathbb{E}^{1/p}[\|\theta_k^b - \theta^*\|^p], \quad k \in \{1, \dots, n-1\}, \quad (20)$$

312 with $p \simeq \log n$ and polynomial dependence on p . To control the second term in the r.h.s. of (19) we
313 note that the matrix Σ_n^b concentrates around Σ_n due to the matrix Bernstein inequality (see Lemma 18
314 for details). Hence, there is a set Ω_1 such that $\mathbb{P}(\Omega_1) \geq 1 - 1/n$ and $\lambda_{\min}(\Sigma_n^b) > 0$ on Ω_1 . Moreover,
315 on this set we may use Berry-Essen-type bound for non-i.i.d. sums of random vectors. Detailed proof
316 is given in Appendix E. \square

317 2.4 Rate of convergence in the Polyak–Juditsky central limit theorem

318 In the final part of this section, we discuss the issue of transition from Σ_n to Σ_∞ and estimation of
319 convergence rates in the Polyak–Juditsky result (4). We utilize the result of Theorem 2 together with
320 the following lemma.

321 **Lemma 1.** Assume that A1 and A9 hold. Let $Y, Y' \sim \mathcal{N}(0, I_d)$. Then, the Kolmogorov distance
322 between the distributions of $\Sigma_n^{1/2}Y$ and $\Sigma_\infty^{1/2}Y'$ is bounded by

$$d_C(\Sigma_n^{1/2}Y, \Sigma_\infty^{1/2}Y') \leq C_\infty n^{\gamma-1},$$

323 where the constant C_∞ is defined in (51).

324 Theorem 2, Lemma 1, and triangle inequality imply the following result on closeness to $\mathcal{N}(0, \Sigma_\infty)$.

325 **Theorem 4.** Assume A1, A3, A8(4), A9. Then, with $Y \sim \mathcal{N}(0, I_d)$ it holds that

$$d_C(\sqrt{n}(\bar{\theta}_n - \theta^*), \Sigma_\infty^{1/2}Y) \leq \frac{C_4}{\sqrt{n}} + \frac{C_5}{n^{\gamma-1/2}} + \frac{C_6}{n^{\gamma/2}} + \frac{C_\infty}{n^{1-\gamma}}, \quad (21)$$

326 where C_4, C_5 and C_6 are given in Theorem 2.

327 **Discussion.** Theorem 2 reveals that the normal approximation through $\mathcal{N}(0, \Sigma_n)$ improves when
 328 the step sizes α_k are less aggressive, that is, as $\gamma \rightarrow 1$. However, Theorem 4 shows that there
 329 is a trade-off, since the rate at which Σ_n converges to Σ_∞ also affects the overall quality of the
 330 approximation. Optimizing the bound in (21) for γ yields an optimal value of $\gamma = 3/4$, leading to
 331 the following approximation rate:

$$d_C(\sqrt{n}(\bar{\theta}_n - \theta^*), \Sigma_\infty^{1/2}Y) \leq \frac{C'_1}{n^{1/4}} + \frac{C'_2}{\sqrt{n}}(\|\theta_0 - \theta^*\| + \|\theta_0 - \theta^*\|^2),$$

332 where C'_1 and C'_2 are instance-dependent quantities (but not depending on $\|\theta_0 - \theta^*\|$), that can be
 333 inferred from Theorem 4. Given the result of Theorem 4 one can proceed with a non-asymptotic
 334 evaluation of the methods for constructing confidence intervals based on direct estimation of Σ_∞ ,
 335 such as [11, 49].

336 **Lower bounds** We provide a lower bound indicating that the bound Theorem 4 is tight at least in
 337 some regimes of step size decay power $\gamma \in (1/2, 1)$. For this aim we consider minimization problem
 338 (1) of the following form:

$$f(\theta) = \theta^2/2 \rightarrow \min_{\theta \in \mathbb{R}}, \quad \theta_0 = 0.$$

339 In this case $\theta^* = 0$. We consider an additive noise model, that is, the stochastic gradient oracles
 340 $F(\theta, \xi)$ have a form $F(\theta, \xi) = \theta + \xi$, where $\xi \sim \mathcal{N}(0, 1)$. Enrolling (2), we get

$$\theta_k = -\sum_{j=1}^k \alpha_j \prod_{\ell=j+1}^k (1 - a\alpha_\ell) \xi_j \text{ and } \sqrt{n}\bar{\theta}_n = -\frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} Q_j \xi_j, \quad (22)$$

341 where $Q_j = \alpha_j \sum_{k=j}^{n-1} \prod_{\ell=j+1}^k (1 - \alpha_\ell)$. Note that $\sqrt{n}(\bar{\theta}_n - \theta^*)$ follows normal distribution
 342 $\mathcal{N}(0, \sigma_{n,\gamma}^2)$ with $\sigma_{n,\gamma}^2 = \frac{1}{n} \sum_{j=1}^{n-1} Q_j^2$. Due to Lemma 1 (see also equation (50) in the Appendix), we
 343 have $G = 1, \Sigma_\infty = 1$, and $\sigma_{n,\gamma}^2 \rightarrow 1$ as $n \rightarrow \infty$. Moreover, the following lower bound holds:

344 **Proposition 1.** Consider the sequence $\{\theta_k\}_{k \in \mathbb{N}}$ defined by the recurrence (22) with $\alpha_j = c_0/(1+j)^\gamma$.
 345 Then it holds, for the number of observations n sufficiently large, that

$$|\sigma_{n,\gamma}^2 - 1| > \frac{C_1(\gamma, c_0)}{n^{1-\gamma}}, \quad (23)$$

346 where the constant $C_1(\gamma, c_0)$ depends only upon c_0 and γ . Moreover, for n large enough

$$d_C(\sqrt{n}(\bar{\theta}_n - \theta^*), \mathcal{N}(0, 1)) > \frac{C_2(\gamma, c_0)}{n^{1-\gamma}}. \quad (24)$$

347 **Discussion.** Proof of Proposition 1 is provided in Appendix F, together with some simple numerical
 348 simulations which indicate the tightness of the lower bound (23). Note that the bound (24) reveals
 349 that the distribution of $\sqrt{n}(\bar{\theta}_n - \theta^*)$ can not be approximated by $\mathcal{N}(0, \Sigma_\infty)$ with the rate faster than
 350 $1/n^{1-\gamma}$. Moreover, it shows that the rate of normal approximation in Theorem 4 can not be improved
 351 when $\gamma \in [3/4, 1)$. This fact is extremely important when taking into account the bootstrap validity
 352 result of Theorem 1 and normal approximation in Theorem 2. Indeed, both results suggests that the
 353 rates of normal approximation of order up to $1/\sqrt{n}$ can be achieved when $\gamma \rightarrow 1$, but they require to
 354 consider another covariance matrix Σ_n , corresponding to the linearized recurrence in (13). At the
 355 same time, in the regime $\gamma \rightarrow 1$, the approximation by $\mathcal{N}(0, \Sigma_\infty)$ can be too slow. It is an interesting
 356 and, to the best of our knowledge, open question to provide lower bounds analogous to Proposition 1
 357 which show the tightness of other summands in Theorem 4 in the regime $1/2 < \gamma < 3/4$.

358 3 Conclusion

359 In our paper, we performed the fully non-asymptotic analysis of the multiplier bootstrap procedure for
 360 SGD applied to strongly convex minimization problems. We showed that the algorithm can achieve
 361 approximation rates in convex distances of order up to $1/\sqrt{n}$. We highlight the fact that the validity
 362 of the multiplier bootstrap procedure does not require one to consider Berry-Esseen bounds with
 363 the asymptotic covariance matrix Σ_∞ , which is in sharp contrast to the methods that require direct
 364 estimation of Σ_∞ .

References

- [1] Bhavya Agrawalla, Krishnakumar Balasubramanian, and Promit Ghosal. High-dimensional central limit theorems for linear functionals of online least-squares sgd. *arXiv preprint arXiv:2302.09727*, 2023.
- [2] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 115–137. PMLR, 25–28 Jun 2019.
- [3] G. D. Anderson and S.-L. Qiu. A monotoneity property of the gamma function. *Proceedings of the American Mathematical Society*, 125(11):3355–3362, 1997.
- [4] Keith Ball. The reverse isoperimetric problem for Gaussian measure. *Discrete & Computational Geometry*, 10(4):411–420, October 1993.
- [5] S. Barsov and V. Ulyanov. Estimates for the closeness of Gaussian measures. *Dokl. Akad. Nauk SSSR*, 291(2):273–277, 1986.
- [6] V. Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.
- [7] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.
- [8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- [9] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.
- [10] Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*, 2022.
- [11] Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251 – 273, 2020.
- [12] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.
- [13] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean. *arXiv e-prints*, page arXiv:1810.08693, October 2018.
- [14] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020.
- [15] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.
- [16] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.
- [17] M. V. Fedoryuk. *Metod perevala*. Izdat. “Nauka”, Moscow, 1977.
- [18] Wolfgang Gabcke. Neue herleitung und explizite restabschätzung der riemann-siegel-formel. 2015.
- [19] Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538–2563, 2019.

- 412 [20] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for
413 non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613.
414 PMLR, 2019.
- 415 [21] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms
416 for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*,
417 15(1):2489–2512, 2014.
- 418 [22] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and
419 applications*, volume 35. Springer Science & Business Media, 2003.
- 420 [23] Guanhui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. 01
421 2020.
- 422 [24] Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp
423 nonasymptotics and asymptotic efficiency in a single algorithm. In Po-Ling Loh and Maxim
424 Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of
425 *Proceedings of Machine Learning Research*, pages 909–981. PMLR, 02–05 Jul 2022.
- 426 [25] Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and
427 online inference via local sgd. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of
428 Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning
429 Research*, pages 1613–1661. PMLR, 02–05 Jul 2022.
- 430 [26] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algo-
431 rithms for machine learning. *Advances in neural information processing systems*, 24:451–459,
432 2011.
- 433 [27] Arkadi Nemirovski, Anatoli Juditsky, Guanhui Lan, and Alexander Shapiro. Robust stochastic
434 approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–
435 1609, 2009.
- 436 [28] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Opti-
437 mization. Springer, 2004.
- 438 [29] Frank W. J. Olver. *Asymptotics and special functions*. AKP Classics. A K Peters, Ltd., Wellesley,
439 MA, 1997. Reprint of the 1974 original [Academic Press, New York; MR0435697 (55 #8655)].
- 440 [30] Adam Osekowski. *Sharp martingale and semimartingale inequalities*, volume 72. Springer
441 Science & Business Media, 2012.
- 442 [31] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging.
443 *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- 444 [32] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*,
445 12(1):145–151, 1999.
- 446 [33] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for
447 strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on
448 International Conference on Machine Learning*, pages 1571–1578, 2012.
- 449 [34] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of
450 mathematical statistics*, pages 400–407, 1951.
- 451 [35] Abhishek Roy and Krishnakumar Balasubramanian. Online covariance estimation for stochastic
452 gradient descent under markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.
- 453 [36] David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Tech-
454 nical report, Cornell University Operations Research and Industrial Engineering, 1988.
- 455 [37] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel,
456 Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds
457 for stochastic optimization and variational inequalities: the case of unbounded variance. In
458 *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

- 459 [38] Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian
460 Approximation and Multiplier Bootstrap for Polyak-Ruppert Averaged Linear Stochastic
461 Approximation with Applications to TD Learning. In *Advances in Neural Information Processing
462 Systems*, volume 37, pages 12408–12460. Curran Associates, Inc., 2024.
- 463 [39] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic
464 average gradient. *Mathematical Programming*, 162:83–112, 2017.
- 465 [40] Jun Shao. *Mathematical statistics*. Springer Science & Business Media, 2003.
- 466 [41] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics.
467 Springer New York, NY, 1 edition, 1995. Springer Book Archive, © Springer Science+Business
468 Media New York 1995.
- 469 [42] Qi-Man Shao and Zhuo-Song Zhang. Berry–Esseen bounds for multivariate nonlinear statis-
470 tics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*,
471 28(3):1548–1576, 2022.
- 472 [43] Marina Sheshukova, Denis Belomestny, Alain Durmus, Eric Moulines, Alexey Naumov, and
473 Sergey Samsonov. Nonasymptotic analysis of stochastic gradient descent with the richardson-
474 romberg extrapolation. *arXiv preprint arXiv:2410.05106*, 2024.
- 475 [44] R Srikant. Rates of Convergence in the Central Limit Theorem for Markov Chains, with an
476 Application to TD learning. *arXiv preprint arXiv:2401.15719*, 2024.
- 477 [45] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and
478 Trends® in Machine Learning*, 8(1-2):1–230, 2015.
- 479 [46] Weichen Wu, Gen Li, Yuting Wei, and Alessandro Rinaldo. Statistical Inference for Temporal
480 Difference Learning with Linear Function Approximation. *arXiv preprint arXiv:2410.16106*,
481 2024.
- 482 [47] Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Online Bootstrap Inference with Noncon-
483 vex Stochastic Gradient Descent Estimator. *arXiv preprint arXiv:2306.02205*, 2023.
- 484 [48] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes
485 for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.
- 486 [49] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online Covariance Matrix Estimation in Stochastic
487 Gradient Descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract clearly lays out the paper’s main contributions—namely, the non-asymptotic validity of the multiplier bootstrap for constructing confidence sets in SGD, with convex-distance approximation rates up to $O(n^{-\gamma/2})$ (and up to $O(n^{-1/2})$ as $\gamma \rightarrow 1$), constituting the first fully non-asymptotic bound of this kind in SGD algorithms. The introduction then explicitly enumerates these same contributions (non-asymptotic bootstrap validity, the role of the linearized covariance Σ_n , and Gaussian approximation rates with matching lower bounds) in its “Our key contributions” section, directly reflecting the theorems and scope developed later in the paper.

Main results are, respectively, the ones of Theorem 1 and Theorem 3, their statements are complete and supported by the proofs in the Appendix section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Section 2, we explicitly acknowledge key limitations of our analysis, including the almost-sure boundedness requirements on the noise and bootstrap weights (Assumptions A2(ii)–(iii) and A5) as well as Lipschitz and co-coercivity conditions on the gradient. We furthermore note that relaxing these via gradient clipping schemes is left as future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are stated with explicit pointers to the underlying assumptions (see Sections 2 and 2.2), and every theorem and corollary is accompanied by a full, detailed proof in the supplementary paper, with clear references to the corresponding sections. To help the reader, we also include the proofs of some non-standard results (on which we do not claim originality); we always carefully state sources.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Numerical results for Proposition 1 are given in Appendix F with a full description of the sets of hyperparameters we used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

- (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Even though the experiment is toy-like - most of the paper is theoretical-, we include a link to an anonymous GitHub repository.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental section in Appendix F explicitly provides the settings (step size, specific equations, number of observations, etc.) necessary for understanding and reproducing the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We omit error bars since the experiment contains no stochastic components.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All necessary information to reproduce experiments is provided in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This paper is of purely theoretical nature, and the proposed methods do not deal with sensitive attributes that could induce unfairness or privacy issues.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper is of purely theoretical nature. We do not foresee any societal harm from the proof of non-asymptotic bootstrap validity.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper contains no models or datasets with potential for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No external datasets, software, or assets were used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets were introduced in this research.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research on human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

804 Answer: [NA]

805 Justification: Paper does not involve crowdsourcing nor research on human subjects.

806 Guidelines:

- 807 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 808 human subjects.
- 809 • Depending on the country in which research is conducted, IRB approval (or equivalent)
- 810 may be required for any human subjects research. If you obtained IRB approval, you
- 811 should clearly state this in the paper.
- 812 • We recognize that the procedures for this may vary significantly between institutions
- 813 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
- 814 guidelines for their institution.
- 815 • For initial submissions, do not include any information that would break anonymity (if
- 816 applicable), such as the institution conducting the review.

817 **16. Declaration of LLM usage**

818 Question: Does the paper describe the usage of LLMs if it is an important, original, or

819 non-standard component of the core methods in this research? Note that if the LLM is used

820 only for writing, editing, or formatting purposes and does not impact the core methodology,

821 scientific rigorousness, or originality of the research, declaration is not required.

822 Answer: [NA]

823 Justification: The paper does not involve the use of LLMs or related methods.

824 Guidelines:

- 825 • The answer NA means that the core method development in this research does not
- 826 involve LLMs as any important, original, or non-standard components.
- 827 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
- 828 for what should or should not be described.

Below we provide proofs of the main results of this submission. The appendix is structured as follows:

- Appendix **A** contains technical statements about the properties of the step sizes α_k and the matrices Q_i , which are instrumental to our subsequent analysis;
- Appendix **B** contains auxiliary moment bounds on the last-iterate error $\mathbb{E}[\|\theta_k - \theta^*\|^p]$ for $p \geq 2$. These results will be crucial for our further analysis, especially for the Gaussian approximation results in the real world (Theorem 2);
- Appendix **C** provides the proof of Theorem 2;
- Appendix **D** contains the proof of the Gaussian approximation in the "real world" using the asymptotic covariance matrix Σ_∞ instead of Σ_n (Theorem 4), along with an appropriate Gaussian comparison result;
- Appendix **E** contains the results required for the Gaussian approximation of "bootstrap-world" statistics and the proof of the main result of this submission (see Theorem 1);
- Finally, Appendix **F** contains the proof of the lower bounds on the accuracy of the normal approximation for $\sqrt{n}(\hat{\theta}_n - \theta^*)$ with $\mathcal{N}(0, \Sigma_\infty)$.

A Technical bounds

We begin this section with useful technical bounds on sums of coefficients

$$\sum_{i=m}^k \alpha_i^p,$$

where the step sizes α_i have a form

$$\alpha_i = \frac{c_0}{(k_0 + i)^\gamma}, \quad 1/2 < \gamma < 1, \quad k_0 \geq 1.$$

We also bound other related quantities, which are instrumental to our further analysis.

Lemma 2. Assume A9. Then

(a) for any $p \geq 2$, it holds that

$$\sum_{i=1}^k \alpha_i^p \leq \frac{c_0^p}{p\gamma - 1},$$

(b) for any $m \in \{0, \dots, k\}$, it holds that

$$\sum_{i=m+1}^k \alpha_i \geq \frac{c_0}{2(1-\gamma)} ((k + k_0)^{1-\gamma} - (m + k_0)^{1-\gamma}),$$

Proof. To prove (a), observe that

$$\sum_{i=1}^k \alpha_i^p \leq c_0^p \int_1^{+\infty} \frac{dx}{x^{p\gamma}} \leq \frac{c_0^p}{p\gamma - 1},$$

To prove (b), note that for any $i \geq 1$ and $k_0 \geq 1$, we have $2(i + k_0)^{-\gamma} \geq (i + k_0 - 1)^{-\gamma}$. Hence,

$$\sum_{i=m+1}^k \alpha_i \geq \frac{1}{2} \sum_{i=m}^{k-1} \alpha_i \geq \frac{c_0}{2} \int_{m+k_0}^{k+k_0} \frac{dx}{x^\gamma} = \frac{c_0}{2(1-\gamma)} ((k + k_0)^{1-\gamma} - (m + k_0)^{1-\gamma}).$$

□

Lemma 3. For any $A > 0$, any $0 \leq i \leq n-1$, and any $\gamma \in (1/2, 1)$ it holds

$$\sum_{j=i}^{n-1} \exp\left\{-A(j^{1-\gamma} - i^{1-\gamma})\right\} \leq \begin{cases} 1 + \exp\left\{\frac{1}{1-\gamma}\right\} \frac{1}{A^{1/(1-\gamma)(1-\gamma)}} \Gamma\left(\frac{1}{1-\gamma}\right), & \text{if } Ai^{1-\gamma} \leq \frac{1}{1-\gamma} \text{ and } i \geq 1; \\ 1 + \frac{1}{A(1-\gamma)^2} i^\gamma, & \text{if } Ai^{1-\gamma} > \frac{1}{1-\gamma} \text{ and } i \geq 1; \\ 1 + \frac{1}{A^{1/(1-\gamma)(1-\gamma)}} \Gamma\left(\frac{1}{1-\gamma}\right), & \text{if } i = 0. \end{cases}$$

855 *Proof.* Note that

$$\begin{aligned} \sum_{j=i}^{n-1} \exp\left\{-A(j^{1-\gamma} - i^{1-\gamma})\right\} &\leq 1 + \exp\left\{Ai^{1-\gamma}\right\} \int_i^{+\infty} \exp\left\{-Ax^{1-\gamma}\right\} dx \\ &= 1 + \exp\left\{Ai^{1-\gamma}\right\} \frac{1}{A^{1/(1-\gamma)}(1-\gamma)} \int_{Ai^{1-\gamma}}^{+\infty} e^{-u} u^{\frac{1}{1-\gamma}-1} du \end{aligned}$$

856 Applying [18, Theorem 4.4.3], we get

$$\int_{Ai^{1-\gamma}}^{+\infty} e^{-u} u^{\frac{1}{1-\gamma}-1} du \leq \begin{cases} \Gamma\left(\frac{1}{1-\gamma}\right), & \text{if } Ai^{1-\gamma} < \frac{1}{1-\gamma}; \\ \frac{1}{1-\gamma} \exp\{-Ai^{1-\gamma}\} A^{\gamma/(1-\gamma)} i^{\gamma}, & \text{otherwise.} \end{cases}$$

857 Combining inequities above, for $i \geq 1$ we obtain

$$\sum_{j=i}^{n-1} \exp\left\{-A(j^{1-\gamma} - i^{1-\gamma})\right\} \leq \begin{cases} 1 + \exp\left\{\frac{1}{1-\gamma}\right\} \frac{1}{A^{1/(1-\gamma)}(1-\gamma)} \Gamma\left(\frac{1}{1-\gamma}\right), & \text{if } Ai^{1-\gamma} < \frac{1}{1-\gamma}; \\ 1 + \frac{1}{A(1-\gamma)^2} i^{\gamma}, & \text{otherwise.} \end{cases},$$

858 and for $i = 0$, we have

$$\sum_{j=0}^{n-1} \exp\left\{-Aj^{1-\gamma}\right\} \leq 1 + \frac{1}{A^{1/(1-\gamma)}(1-\gamma)} \Gamma\left(\frac{1}{1-\gamma}\right).$$

859 □

860 Based on Lemma 2 and Lemma 3 above, we can prove that the matrices Q_i defined in (11) are
861 bounded for any i . This fact will be used later in Appendix C. Moreover, $\lambda_{\min}(Q_i)$ is bounded from
862 below, which allows us to prove that $\lambda_{\min}(\Sigma_n)$ is also bounded away from below.

863 **Lemma 4.** Assume A1 and A9. Then for any $i \in \{0, \dots, n-1\}$ it holds that

$$\lambda_{\max}(Q_i) \leq C_Q,$$

864 where the constant C_Q is given by

$$C_Q = \left[1 + \max\left(\exp\left\{\frac{1}{1-\gamma}\right\} \left(\frac{2(1-\gamma)}{\mu c_0}\right)^{1/(1-\gamma)} \frac{1}{1-\gamma} \Gamma\left(\frac{1}{1-\gamma}\right), \frac{2}{\mu c_0(1-\gamma)}\right) \right] c_0. \quad (25)$$

865 Moreover,

$$\lambda_{\min}(Q_i) \geq \frac{1}{L_1} (1 - (1 - \alpha_i L_1)^{n-i}), \text{ and } \|\Sigma_n^{-1/2}\| \leq C_{\Sigma}, \quad (26)$$

866 where the matrix Σ_n is defined in (13), and

$$C_{\Sigma} = \frac{\sqrt{2}L_1}{(1 - \exp\{-\frac{\mu c_0 L_1}{2(k_0+1)}\}) \sqrt{\lambda_{\min}(\Sigma_{\xi})}}. \quad (27)$$

867 *Proof.* Note that using Lemma 2(b), for $i \geq 0$, it holds that

$$\begin{aligned} \lambda_{\max}(Q_i) &\leq \alpha_i \sum_{j=i}^{n-1} \prod_{k=i+1}^j (1 - \alpha_k \mu) \leq \alpha_i \sum_{j=i}^{n-1} \exp\left\{-\mu \sum_{k=i+1}^j \alpha_k\right\} \\ &\leq \alpha_i \sum_{j=i+k_0}^{n-1+k_0} \exp\left\{-\frac{\mu c_0}{2(1-\gamma)} (j^{1-\gamma} - (i+k_0)^{1-\gamma})\right\}. \end{aligned}$$

868 Using Lemma 3, we complete the first part with the constant C_Q defined in (25). In order to prove
869 (26), we note that

$$\lambda_{\min}(Q_i) \geq \alpha_i \sum_{j=i}^{n-1} (1 - \alpha_i L_1)^{j-i} = \frac{1}{L_1} (1 - (1 - \alpha_i L_1)^{n-i}).$$

870 Then for $i \leq n/2$, we have

$$\lambda_{\min}(Q_i) \geq \frac{1}{L_1}(1 - (1 - \alpha_i L_1)^{n/2}) \geq \frac{1}{L_1}(1 - \exp\{-\mu \alpha_i L_1 n/2\}) \geq \frac{1}{L_1}(1 - \exp\{-\frac{\mu c_0 L_1}{2(k_0 + 1)}\}) ,$$

871 where the last inequality holds, since $\alpha_i n \geq \alpha_n n \geq \frac{c_0 n}{k_0 + n} \geq \frac{c_0}{1 + k_0}$. Combining previous inequalities,
872 we get

$$\lambda_{\min}(\Sigma_n) \geq \lambda_{\min}\left(n^{-1} \sum_{i=1}^{n/2} Q_i \Sigma_{\xi} Q_i^{\top}\right) \geq \frac{\lambda_{\min}(\Sigma_{\xi})}{2L_1^2}(1 - \exp\{-\frac{\mu c_0 L_1}{2(k_0 + 1)}\})^2 ,$$

873 and (26) follows. \square

874 **Lemma 5.** Assume A1 and A3. Then

$$\|H(\theta)\| \leq L_H \|\theta - \theta^*\|^2 ,$$

875 where $L_H = \max(L_3, 2L_1/\beta)$.

876 *Proof.* Using A3 and the definition of $H(\theta)$ in (9), we get

$$\|H(\theta)\| \mathbb{1}(\|\theta - \theta^*\| \leq \beta) \leq L_3 \|\theta - \theta^*\|^2 .$$

877 Since $\mu I \preceq \nabla^2 f(\theta) \preceq L_1 I$, we also obtain

$$\|H(\theta)\| \mathbb{1}(\|\theta - \theta^*\| > \beta) \leq 2L_1 \mathbb{1}(\|\theta - \theta^*\| > \beta) \|\theta - \theta^*\| \leq \frac{2L_1}{\beta} \|\theta - \theta^*\|^2 .$$

878 This concludes the proof. \square

879 B Last iterate bound

880 **Lemma 6.** Assume A1, A3, A8(2), and A9. Then for any $k \in \mathbb{N}$ it holds that

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq C_1 \exp\left\{-\frac{\mu c_0}{4}(k + k_0)^{1-\gamma}\right\} [\|\theta_0 - \theta^*\|^2 + \sigma_2^2] + C_2 \sigma_2^2 \alpha_k ,$$

881 where σ_2^2 is defined in A8(2), and the constants C_1 and C_2 are given by

$$C_1 = \exp\left\{\frac{3\mu c_0}{4(1-\gamma)} k_0^{1-\gamma}\right\} \left((1 + L_2^{-2}) \exp\left\{\frac{6c_0^2 L_2^2}{2\gamma - 1}\right\} + \frac{2c_0^2}{2\gamma - 1} \right) ,$$

$$C_2 = \frac{2^{1+\gamma}}{\mu} .$$

882 *Proof.* From (2) and A8 it follows that

$$\|\theta_k - \theta^*\|^2 = \|\theta_{k-1} - \theta^*\|^2 - 2\alpha_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle + \alpha_k^2 \|\nabla f(\theta_{k-1}) + \zeta_k\|^2 .$$

883 Using A1 and A8(2), we obtain

$$2\alpha_k \mathbb{E}[\langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle | \mathcal{F}_{k-1}] = 2\alpha_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) - \nabla f(\theta^*) \rangle .$$

884 Using A8(2) and A1, we get

$$\begin{aligned} \mathbb{E}[\|\nabla f(\theta_{k-1}) + \zeta_k\|^2 | \mathcal{F}_{k-1}] &= \|\nabla f(\theta_{k-1}) - \nabla f(\theta^*)\|^2 + \mathbb{E}[\|\eta(\xi_k) + g(\theta_{k-1}, \xi_k)\|^2 | \mathcal{F}_{k-1}] \\ &\leq L_1 \langle \nabla f(\theta_{k-1}) - \nabla f(\theta^*), \theta_{k-1} - \theta^* \rangle + 2L_2^2 \|\theta_{k-1} - \theta^*\|^2 + 2\sigma_2^2 . \end{aligned}$$

885 Combining the above inequalities, we obtain

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq (1 - \mu \alpha_k (2 - \alpha_k L_1) + 2\alpha_k^2 L_2^2) \mathbb{E}[\|\theta_{k-1} - \theta^*\|^2] + 2\alpha_k^2 \sigma_2^2 . \quad (28)$$

886 By applying the recurrence (28), we obtain that

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq A_{1,k} \|\theta_0 - \theta^*\|^2 + 2\sigma_2^2 A_{2,k} ,$$

887 where we have set

$$\begin{aligned} A_{1,k} &= \prod_{i=1}^k (1 - (3/2)\alpha_i\mu + 2\alpha_i^2 L_2^2), \\ A_{2,k} &= \sum_{i=1}^k \prod_{j=i+1}^k (1 - (3/2)\alpha_j\mu + 2\alpha_j^2 L_2^2) \alpha_i^2. \end{aligned} \quad (29)$$

888 Using the elementary bound $1 + t \leq e^t$ for any $t \in \mathbb{R}$, we get

$$A_{1,k} \leq \exp\left\{-(3/2)\mu \sum_{i=1}^k \alpha_i\right\} \exp\left\{2L_2^2 \sum_{i=1}^k \alpha_i^2\right\}.$$

889 Using Lemma 2, we obtain

$$A_{1,k} \leq c_1 \exp\left\{-\frac{3\mu c_0}{4(1-\gamma)}(k + k_0)^{1-\gamma}\right\},$$

890 where we have set

$$c_1 = \exp\left\{\frac{2c_0^2 L_2^2}{2\gamma - 1} + \frac{3\mu c_0}{4(1-\gamma)} k_0^{1-\gamma}\right\}. \quad (30)$$

891 Now we estimate $A_{2,k}$. Let k_1 be the largest index k such that $4\alpha_k^2 L_2^2 \geq \alpha_k\mu$. Then, for $i > k_1$, we
892 have that

$$1 - (3/2)\alpha_i\mu + 2\alpha_i^2 L_2^2 \leq 1 - \alpha_i\mu.$$

893 Thus, using the definition of $A_{2,k}$ in (29), we obtain that

$$A_{2,k} \leq \sum_{i=1}^k \alpha_i^2 \prod_{j=i+1}^k (1 - \alpha_j\mu) + \sum_{i=1}^{k_1} \alpha_i^2 \left\{ \prod_{j=i+1}^{k_1} (1 + 2\alpha_j^2 L_2^2) \right\} \left\{ \prod_{j=k_1+1}^k (1 - \alpha_j\mu) \right\}.$$

894 Note that

$$\begin{aligned} \sum_{i=1}^{k_1} \alpha_i^2 \prod_{j=i+1}^{k_1} (1 + 2\alpha_j^2 L_2^2) &= \frac{1}{2L_2^2} \sum_{i=1}^{k_1} \left(\prod_{j=i}^{k_1} (1 + 2\alpha_j^2 L_2^2) - \prod_{j=i+1}^{k_1} (1 + 2\alpha_j^2 L_2^2) \right) \\ &\leq \frac{1}{2L_2^2} \prod_{j=1}^{k_1} (1 + 2\alpha_j^2 L_2^2) \leq \frac{1}{2L_2^2} \exp\left\{2L_2^2 \sum_{j=1}^{k_1} \alpha_j^2\right\}. \end{aligned}$$

895 Note, that for $k \leq k_1$, $\alpha_k \geq \mu/(4L_2^2)$, hence, we have

$$\prod_{j=k_1+1}^k (1 - \alpha_j\mu) \leq \exp\left\{-\mu \sum_{i=1}^k \alpha_i\right\} \exp\left\{\mu \sum_{i=1}^{k_1} \alpha_i\right\} \leq \exp\left\{-\mu \sum_{i=1}^k \alpha_i\right\} \exp\left\{4L_2^2 \sum_{i=1}^{k_1} \alpha_i^2\right\}.$$

896 Moreover, for any $m \in \{1, \dots, k\}$, we obtain

$$\begin{aligned} \sum_{i=1}^k \alpha_i^2 \prod_{j=i+1}^k (1 - \alpha_j\mu) &= \sum_{i=1}^m \prod_{j=i+1}^k (1 - \alpha_j\mu) \alpha_i^2 + \sum_{i=m+1}^k \prod_{j=i+1}^k (1 - \alpha_j\mu) \alpha_i^2 \\ &\leq \prod_{j=m+1}^k (1 - \alpha_j\mu) \sum_{i=1}^m \alpha_i^2 + \alpha_m \sum_{i=m+1}^k \prod_{j=i+1}^k (1 - \alpha_j\mu) \alpha_i \\ &\leq \exp\left\{-\mu \sum_{j=m+1}^k \alpha_j\right\} \sum_{i=1}^m \alpha_i^2 + \frac{\alpha_m}{\mu} \sum_{i=m+1}^k \left(\prod_{j=i+1}^k (1 - \alpha_j\mu) - \prod_{j=i}^k (1 - \alpha_j\mu) \right) \\ &\leq \exp\left\{-\mu \sum_{j=m+1}^k \alpha_j\right\} \sum_{i=1}^m \alpha_i^2 + \frac{\alpha_m}{\mu} \left(1 - \prod_{j=m+1}^k (1 - \alpha_j\mu) \right) \\ &\leq \exp\left\{-\mu \sum_{j=m+1}^k \alpha_j\right\} \sum_{i=1}^m \alpha_i^2 + \frac{\alpha_m}{\mu}. \end{aligned}$$

Thus, setting $m = \lfloor k/2 \rfloor$, and using the definition of $A_{2,k}$ in (29), we obtain that

$$A_{2,k} \leq \exp \left\{ \frac{-\mu c_0}{2(1-\gamma)} ((k+k_0)^{1-\gamma} - (\lfloor k/2 \rfloor + k_0)^{1-\gamma}) \right\} \frac{c_0^2}{2\gamma-1} + \frac{c_0}{\mu(k_0 + \lfloor k/2 \rfloor)^\gamma} \\ + c_2 \exp \left\{ -\frac{\mu c_0}{2(1-\gamma)} (k+k_0)^{1-\gamma} \right\},$$

where we have set

$$c_2 = \frac{1}{2L_2^2} \exp \left\{ \frac{6c_0^2 L_2^2}{2\gamma-1} + \frac{\mu c_0}{2(1-\gamma)} k_0^{1-\gamma} \right\}. \quad (31)$$

Using that $\lfloor k/2 \rfloor \leq k/2$ together with the elementary inequality

$$\frac{x^\beta}{\beta} - \frac{(x/2)^\beta}{\beta} \geq \frac{x^\beta}{2},$$

which is valid for $\beta \in (0, 1]$, and $\frac{c_0}{\mu(k_0 + \lfloor k/2 \rfloor)^\gamma} \leq \frac{2^\gamma c_0}{\mu(k+k_0)^\gamma}$, we obtain that

$$A_{2,k} \leq \exp \left\{ -\frac{\mu c_0}{4} (k+k_0)^{1-\gamma} \right\} \exp \left\{ \frac{\mu c_0}{2(1-\gamma)} k_0^{1-\gamma} \right\} \frac{c_0^2}{2\gamma-1} + \frac{2^\gamma c_0}{\mu(k+k_0)^\gamma} \\ + c_2 \exp \left\{ \frac{-\mu c_0}{2(1-\gamma)} (k+k_0)^{1-\gamma} \right\}.$$

Combining the bounds for $A_{1,k}$ and $A_{2,k}$, we obtain that

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq c_1 \exp \left\{ -\frac{\mu c_0}{(1-\gamma)} (k+k_0)^{1-\gamma} \right\} \|\theta_0 - \theta^*\|^2 \\ + \exp \left\{ -\frac{\mu c_0}{4} (k+k_0)^{1-\gamma} \right\} \frac{2c_0^2 \sigma_2^2}{2\gamma-1} \exp \left\{ \frac{\mu c_0}{2(1-\gamma)} k_0^{1-\gamma} \right\} + \frac{2^{1+\gamma} c_0 \sigma_2^2}{\mu(k+k_0)^\gamma} \\ + 2c_2 \sigma_2^2 \exp \left\{ \frac{-\mu c_0}{2(1-\gamma)} (k+k_0)^{1-\gamma} \right\} \\ \leq C_1 \exp \left\{ -\frac{\mu c_0}{4} (k+k_0)^{1-\gamma} \right\} [\|\theta_0 - \theta^*\|^2 + \sigma_2^2] + C_2 \alpha_k,$$

where we have set constants C_1 and C_2 using the definitions of c_1 and c_2 from (30) and (31). \square

Now we provide the following corollary:

Corollary 1. Under the assumptions of Lemma 6, it holds that

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq D_1 (\|\theta_0 - \theta^*\|^2 + \sigma_2^2) \alpha_k,$$

where

$$D_1 = C_1 (1/c_0 + C_2) \left(\frac{4\gamma}{(1-\gamma)\mu c_0 e} \right)^{\gamma/(1-\gamma)}.$$

Proof. Define $C_3 = (\frac{4\gamma}{(1-\gamma)\mu c_0 e})^{\gamma/(1-\gamma)} > 1$, then $\exp\{-\mu c_0 (k+k_0)^{1-\gamma}/4\} \leq C_3 (k+k_0)^{-\gamma}$, and the statement follows. \square

Now we provide bound for p-moment of last iterate.

Proposition 2. Assume A1, A3, A8(2p), and A9. Then for any $k \in \mathbb{N}$ it holds that

$$\mathbb{E}[\|\theta_k - \theta^*\|^{2p}] \leq C_{2p,1} \exp \left\{ -\frac{p\mu c_0}{4} (k+k_0)^{1-\gamma} \right\} (\|\theta_0 - \theta^*\|^{2p} + \sigma_{2p}^{2p}) + C_{2p,2} \sigma_{2p}^{2p} \alpha_k^p,$$

where

$$C_{2p,1} = 2^{2p-1} (D_{2(p-1)} C_4^p c_0^p + 1) c_4,$$

$$C_{2p,2} = 2^{2p-1} D_{2(p-1)} C_4^p \frac{2^{1+\gamma p}}{\mu p c_0},$$

constants $D_{2(p-1)}$ are defined in (38), and

$$C_4^p = (4c_0^{1/2}2^{\gamma/2} + 2^\gamma + 4c_0)^p$$

$$c_4 = \left(\exp \left\{ \exp \left\{ 5pc_0(L_1 + L_2) \right\} \frac{4p^2(L_1 + L_2)^2}{2\gamma - 1} \right\} + 1 \right) \exp \left\{ \frac{p\mu c_0}{1 - \gamma} k_0^{1-\gamma} \right\} \frac{1}{\gamma(p+1) - 1}$$

Proof. We prove the statement by induction in p . We first assume that $\theta_0 = \theta^*$ and then provide a result for arbitrary initial condition. The result for $p = 1$ is provided in Corollary 1. Assume that for any $t \leq p - 1$ and all $k \in \mathbb{N}$ we proved that

$$\mathbb{E}[\|\theta_k - \theta^*\|^{2t}] \leq D_{2t} \sigma_{2t}^{2t} \alpha_k^t, \quad (32)$$

and the sequence of constants $\{D_{2t}\}$ is non-decreasing in t . Inequality (32) implies that, since $\sigma_{2t} \leq \sigma_{2p}$ for $t \leq p - 1$,

$$\mathbb{E}[\|\theta_k - \theta^*\|^{2t}] \leq D_{2t} \sigma_{2p}^{2t} \alpha_k^t.$$

For any $k \in \mathbb{N}$ we denote $\delta_k = \|\theta_k - \theta^*\|$. Using (2), we get

$$\begin{aligned} \delta_k^{2p} &= (\delta_{k-1}^2 - 2\alpha_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle + \alpha_k^2 \|\nabla f(\theta_{k-1}) + \zeta_k\|^2)^p \\ &= \sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}}} \frac{p!}{i!j!l!} \delta_{k-1}^{2i} (-2\alpha_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle)^j \alpha_k^{2l} \|\nabla f(\theta_{k-1}) + \zeta_k\|^{2l}. \end{aligned}$$

Now we bound each term in the sum above.

1. First, for $i = p, j = 0, l = 0$, the corresponding term in the sum equals δ_{k-1}^{2p} .

2. Second, for $i = p - 1, j = 1, l = 0$, we obtain, applying A1, that

$$\begin{aligned} 2p\alpha_k \mathbb{E}[\langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle \delta_{k-1}^{2(p-1)} | \mathcal{F}_{k-1}] &= 2p\alpha_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) - \nabla f(\theta^*) \rangle \delta_{k-1}^{2(p-1)} \\ &\geq 2p\mu\alpha_k \delta_{k-1}^{2p}. \end{aligned}$$

3. Third, for $l \geq 1$ or $j \geq 2$ (that is, $2l + j \geq 2$), we use Cauchy-Schwartz inequality

$$|\langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1}) + \zeta_k \rangle^j| \leq \|\theta_{k-1} - \theta^*\|^j \|\nabla f(\theta_{k-1}) + \zeta_k\|^j,$$

moreover, applying A1 and A8(2p) together with the Lyapunov inequality, we get

$$\begin{aligned} \mathbb{E}[\|\nabla f(\theta_{k-1}) + \zeta_k\|^{2l+j} | \mathcal{F}_{k-1}] &= \mathbb{E}[\|\nabla f(\theta_{k-1}) + g(\theta_{k-1}, \xi_k) + \eta(\xi_k)\|^{2l+j} | \mathcal{F}_{k-1}] \\ &\leq 2^{2l+j-1} ((L_1 + L_2)^{2l+j} \delta_{k-1}^{2l+j} + \sigma_{2p}^{2l+j}). \end{aligned}$$

Combining inequalities above, we get

$$\begin{aligned} \mathbb{E}[\delta_k^{2p} | \mathcal{F}_{k-1}] &\leq \left(1 - 2p\mu\alpha_k + \sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}: \\ j+2l \geq 2}} \frac{p!}{i!j!l!} \alpha_k^{j+2l} 2^{2l+2j-1} (L_1 + L_2)^{2l+j} \right) \delta_{k-1}^{2p} \\ &\quad + \sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}: \\ j+2l \geq 2}} \frac{p!}{i!j!l!} \delta_{k-1}^{2i+j} \alpha_k^{j+2l} 2^{2l+2j-1} \sigma_{2p}^{2l+j}. \end{aligned}$$

Consider the first term above, and note that

$$\begin{aligned} &\sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}: \\ j+2l \geq 2}} \frac{p!}{i!j!l!} \alpha_k^{j+2l} 2^{2l+2j-1} (L_1 + L_2)^{2l+j} \quad (33) \\ &\leq 2\alpha_k^2 (L_1 + L_2)^2 \left(\sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}: \\ l \geq 1}} \frac{p!}{i!j!l!} (4\alpha_k (L_1 + L_2))^j (4\alpha_k^2 (L_1 + L_2)^2)^{l-1} + \sum_{\substack{i+j+l=p \\ i,j,l \in \{0, \dots, p\}: \\ l=0; j \geq 2}} \frac{p!}{i!j!l!} (4\alpha_k (L_1 + L_2))^{j-2} \right) \\ &\leq 2p^2 \alpha_k^2 (L_1 + L_2)^2 (1 + 5\alpha_k (L_1 + L_2))^p. \end{aligned}$$

926 Hence,

$$\mathbb{E}[\delta_k^{2p}] \leq (1 - 2p\mu\alpha_k + 2p^2\alpha_k^2(L_1 + L_2)^2(1 + 5\alpha_k(L_1 + L_2))^p) \delta_{k-1}^{2p} + T_1,$$

927 where we have defined

$$T_1 = \sum_{\substack{i+j+l=p; \\ i,j,l \in \{0, \dots, p\}; \\ j+2l \geq 2; \\ i+j=p}} \frac{p!}{i!j!l!} \mathbb{E}[\delta_{k-1}^{2i+j}] \alpha_k^{j+2l} 2^{2l+2j-1} \sigma_{2p}^{2l+j}.$$

928 For the last term we apply Hölder's inequality together with induction assumption (32) and $(k + k_0 -$
929 $1)^{-\gamma} \leq 2^\gamma (k + k_0)^{-\gamma}$ and obtain

$$\begin{aligned} T_1 &\leq \sum_{\substack{i+j+l=p; \\ i,j,l \in \{0, \dots, p\}; \\ j+2l \geq 2}} \frac{p!}{i!j!l!} \mathbb{E}^{1/2}[\delta_{k-1}^{2(i+\lceil j/2 \rceil)}] \mathbb{E}^{1/2}[\delta_{k-1}^{2(i+\lceil j/2 \rceil)}] \alpha_k^{j+2l} 2^{2l+2j-1} \sigma_{2p}^{2l+j} \\ &\leq D_{2(p-1)} \frac{(4c_0^{1/2} 2^{\gamma/2} + 2^\gamma + 4c_0)^p}{2} c_0^p \sigma_{2p}^{2p} (k + k_0)^{-\gamma(p+1)}. \end{aligned}$$

930 Hence, combining the above bounds, we obtain that

$$\mathbb{E}[\delta_k^{2p}] \leq (1 - 2p\mu\alpha_k + 16\alpha_k^2(L_1 + L_2)^2 3^p) \mathbb{E}[\delta_{k-1}^{2p}] + D_{2(p-1)} C_4^p c_0^p \sigma_{2p}^{2p} k^{-\gamma(p+1)}, \quad (34)$$

931 where we have defined

$$C_4^p = (4c_0^{1/2} 2^{\gamma/2} + 2^\gamma + 4c_0)^p.$$

932 Note that

$$1 - 2p\mu\alpha_k + 2p^2\alpha_k^2(L_1 + L_2)^2(1 + 5\alpha_k(L_1 + L_2))^p > 1 - 2p\mu\alpha_k + \alpha_k^2\mu^2 p^2 \geq 0.$$

933 Unrolling the recurrence (34), we get

$$\mathbb{E}[\delta_k^{2p}] \leq A'_{2,k} D_{2(p-1)} C_4^p c_0^p \sigma_{2p}^{2p},$$

934 where we have set

$$A'_{2,k} = \sum_{t=1}^k \prod_{i=t+1}^k (1 - 2p\mu\alpha_i + 2p^2\alpha_i^2(L_1 + L_2)^2(1 + 5\alpha_i(L_1 + L_2))^p) (t + k_0)^{-\gamma(p+1)}. \quad (35)$$

935 For simplicity, we define $C_5 = 2p^2(L_1 + L_2)^2$. Let k_1 is the largest k such that $\alpha_k^2 C_5 (1 + 5\alpha_i(L_1 +$
936 $L_2))^p \geq p\mu\alpha_k$. Then, for $i > k_1$, we have

$$1 - 2p\mu\alpha_i + C_5\alpha_i^2(1 + 5\alpha_i(L_1 + L_2))^p \leq 1 - p\mu\alpha_i.$$

937 Hence, using the definition of $A'_{2,k}$ in (35), we get

$$\begin{aligned} A'_{2,k} &= \sum_{t=k_1+1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t + k_0)^{-\gamma(p+1)} \\ &\quad + \prod_{t=k_1+1}^k \exp\left\{-p\mu\alpha_t\right\} \sum_{t=1}^{k_1} \prod_{i=t+1}^{k_1} \exp\left\{C_5\alpha_i^2(1 + 5\alpha_i(L_1 + L_2))^p\right\} (t + k_0)^{-\gamma(p+1)} \\ &\leq \sum_{t=1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t + k_0)^{-\gamma(p+1)} \\ &\quad + \prod_{t=1}^k \exp\left\{-p\mu\alpha_t\right\} \prod_{t=1}^{k_1} \exp\{p\mu\alpha_t\} \prod_{i=1}^{k_1} \exp\left\{C_5\alpha_i^2(1 + 5\alpha_i(L_1 + L_2))^p\right\} \sum_{t=1}^{k_1} (t + k_0)^{-\gamma(p+1)} \\ &\leq \sum_{t=1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t + k_0)^{-\gamma(p+1)} \\ &\quad + \prod_{i=1}^{k_1} \exp\left\{2C_5\alpha_i^2(1 + 5\alpha_i(L_1 + L_2))^p\right\} \prod_{t=1}^k \exp\left\{-p\mu\alpha_t\right\} \sum_{t=1}^{k_1} (t + k_0)^{-\gamma(p+1)} \end{aligned}$$

938 For any $m \in \{1, \dots, k\}$ we have

$$\begin{aligned}
& \sum_{t=1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t+k_0)^{-\gamma(p+1)} \\
&= \sum_{t=1}^m \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t+k_0)^{-\gamma(p+1)} + \sum_{t=m+1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t+k_0)^{-\gamma(p+1)} \\
&\leq \prod_{i=m+1}^k \exp\left\{-p\mu\alpha_i\right\} \sum_{t=1}^m (t+k_0)^{-\gamma(p+1)} + \sum_{t=m+1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (m+k_0)^{-\gamma p} (t+k_0)^{-\gamma} \\
&\leq \prod_{i=m+1}^k \exp\left\{-p\mu\alpha_i\right\} \sum_{t=1}^k (t+k_0)^{-\gamma(p+1)} + (m+k_0)^{-\gamma p} \sum_{t=1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} (t+k_0)^{-\gamma}
\end{aligned}$$

939 Applying Lemma 2(b), we have

$$\begin{aligned}
& \sum_{t=1}^k \prod_{i=t+1}^k \exp\left\{-p\mu\alpha_i\right\} t^{-\gamma} \leq \sum_{t=1}^k \exp\left\{\frac{-p\mu c_0}{2(1-\gamma)}((k+k_0)^{1-\gamma} - (t+k_0)^{1-\gamma})\right\} (t+k_0)^{-\gamma} \\
&\leq \exp\left\{\frac{-p\mu c_0}{2(1-\gamma)}(k+k_0)^{1-\gamma}\right\} \frac{2}{p\mu c_0} \int_0^{\frac{p\mu c_0}{2(1-\gamma)}(k+k_0)^{1-\gamma}} e^u du \leq \frac{2}{p\mu c_0}.
\end{aligned}$$

940 Applying Lemma 2(a), we get

$$\sum_{i=1}^k (i+k_0)^{-\gamma(p+1)} \leq \frac{1}{(p+1)\gamma-1},$$

941 and

$$\sum_{i=1}^k 2C_5 \alpha_i^2 (1+5\alpha_i(L_1+L_2))^p \leq 2C_5 (1+5c_0(L_1+L_2))^p \sum_{i=1}^{+\infty} \alpha_k^2 \leq \exp\left\{5pc_0(L_1+L_2)\right\} \frac{2C_5}{2\gamma-1}$$

942 Substituting $m = \lfloor k/2 \rfloor$ and applying (b), we get

$$\begin{aligned}
A'_{2,k} &\leq \exp\left\{-\frac{p\mu c_0}{2(1-\gamma)}((k+k_0)^{1-\gamma} - (\lfloor k/2 \rfloor + k_0)^{1-\gamma})\right\} \frac{1}{\gamma(p+1)-1} + \frac{2(\lfloor k/2 \rfloor + k_0)^{-\gamma p}}{p\mu c_0} \\
&+ c_3 \exp\left\{-\frac{p\mu c_0}{2(1-\gamma)}(k+k_0)^{1-\gamma}\right\},
\end{aligned}$$

943 where we have set

$$c_3 = \exp\left\{\exp\left\{5pc_0(L_1+L_2)\right\} \frac{2C_5}{2\gamma-1} + \frac{p\mu c_0}{2(1-\gamma)} k_0^{1-\gamma}\right\} \frac{1}{\gamma(p+1)-1}.$$

944 Using that $\lfloor k/2 \rfloor \leq k/2$ together with the elementary inequality

$$\frac{x^\beta}{\beta} - \frac{(x/2)^\beta}{\beta} \geq \frac{x^\beta}{2},$$

945 which is valid for $\beta \in (0, 1]$, and $\frac{2}{\mu pc_0(\lfloor k/2 \rfloor + k_0)^{\gamma p}} \leq \frac{2^{1+\gamma p}}{\mu pc_0(k+k_0)^{\gamma p}}$, we obtain that

$$\begin{aligned}
A'_{2,k} &\leq \exp\left\{-\frac{p\mu c_0}{4}(k+k_0)^{1-\gamma}\right\} \exp\left\{\frac{p\mu c_0}{2(1-\gamma)} k_0^{1-\gamma}\right\} \frac{1}{\gamma(p+1)-1} \\
&+ \frac{2^{1+\gamma p}}{\mu pc_0(k+k_0)^{\gamma p}} + c_3 \exp\left\{-\frac{p\mu c_0}{2(1-\gamma)}(k+k_0)^{1-\gamma}\right\} \\
&\leq c_4 \exp\left\{-\frac{p\mu c_0}{4}(k+k_0)^{1-\gamma}\right\} + c_5 (k+k_0)^{-\gamma p},
\end{aligned}$$

946 where we have set

$$c_4 = \left(\exp \left\{ \exp \left\{ 5p c_0 (L_1 + L_2) \right\} \frac{4p^2 (L_1 + L_2)^2}{2\gamma - 1} \right\} + 1 \right) \exp \left\{ \frac{p\mu c_0}{1 - \gamma} k_0^{1-\gamma} \right\} \frac{1}{\gamma(p+1) - 1}$$

$$c_5 = \frac{2^{1+\gamma p}}{\mu p c_0}$$

947 Finally, we get

$$\mathbb{E}[\delta_k^{2p}] \leq C'_{2p,1} \exp \left\{ -\frac{p\mu c_0}{4} (k + k_0)^{1-\gamma} \right\} \sigma_{2p}^{2p} + C'_{2p,2} \sigma_{2p}^{2p} \alpha_k^p,$$

948 where

$$C'_{2p,1} = D_{2(p-1)} C_4^p c_4$$

$$C'_{2p,2} = D_{2(p-1)} C_4^p c_5.$$

949 To provide the result for arbitrary start point $\theta_0 = \theta$ we consider the synchronous coupling construc-
950 tion defined by the recursions

$$\begin{aligned} \theta_k &= \theta_{k-1} - \alpha_k (\nabla f(\theta_{k-1}) + g(\theta_{k-1}, \xi_k) + \eta(\xi_k)), & \theta_0 &= \theta \\ \theta'_k &= \theta'_{k-1} - \alpha_k (\nabla f(\theta'_{k-1}) + g(\theta'_{k-1}, \xi_k) + \eta(\xi_k)), & \theta'_0 &= \theta^* \end{aligned} \quad (36)$$

951 For any $k \in \mathbb{N}$ we denote $\delta'_k = \|\theta_k - \theta'_k\|$. Using (36) together with A1 and A8(2p), we get

$$\begin{aligned} \delta_k'^{2p} &= (\delta_{k-1}'^2 - 2\alpha_k \langle \theta_{k-1} - \theta'_{k-1}, \nabla f(\theta_{k-1}) - \nabla f(\theta'_{k-1}) + g(\theta_{k-1}, \xi_k) - g(\theta'_{k-1}, \xi_k) \rangle + \alpha_k^2 (L_1 + L_2)^2 \delta_{k-1}'^2)^p \\ &\leq \sum_{\substack{i+j+l=p; \\ i,j,l \in \{0, \dots, p\}}} \frac{p!}{i!j!l!} \delta_{k-1}'^{2i} (-2\alpha_k \langle \theta_{k-1} - \theta'_{k-1}, \nabla f(\theta_{k-1}) - \nabla f(\theta'_{k-1}) + g(\theta_{k-1}, \xi_k) - g(\theta'_{k-1}, \xi_k) \rangle)^j (\alpha_k (L_1 + L_2) \delta_{k-1}')^{2l} \end{aligned}$$

952 Now we bound each term in the sum above.

953 1. First, for $i = p, j = 0, l = 0$, the corresponding term in the sum equals $\delta_{k-1}'^{2p}$.

954 2. Second, for $i = p-1, j = 1, l = 0$, we obtain, applying A1, that

$$\begin{aligned} &2p\alpha_k \mathbb{E}[\langle \theta_{k-1} - \theta'_{k-1}, \nabla f(\theta_{k-1}) - \nabla f(\theta'_{k-1}) + g(\theta_{k-1}, \xi_k) - g(\theta'_{k-1}, \xi_k) \rangle \delta_{k-1}'^{2(p-1)} | \mathcal{F}_{k-1}] \\ &= 2p\alpha_k \langle \theta_{k-1} - \theta'_{k-1}, \nabla f(\theta_{k-1}) - \nabla f(\theta'_{k-1}) \rangle \delta_{k-1}'^{2(p-1)} \geq 2p\mu\alpha_k \delta_{k-1}'^{2p}. \end{aligned}$$

955 3. Third, for $l \geq 1$ or $j \geq 2$ (that is, $2l + j \geq 2$), we use Cauchy-Schwartz inequality together
956 with A8 and A1

$$|\langle \theta_{k-1} - \theta'_{k-1}, \nabla f(\theta_{k-1}) - \nabla f(\theta'_{k-1}) + g(\theta_{k-1}, \xi_k) - g(\theta'_{k-1}, \xi_k) \rangle^j| \leq \|\theta_{k-1} - \theta'_{k-1}\|^{2j} (L_1 + L_2)^j,$$

957 Combining inequalities above, we obtain

$$\mathbb{E}[\delta_k'^{2p} | \mathcal{F}_{k-1}] \leq (1 - 2p\mu\alpha_k + \sum_{\substack{i+j+l=p; \\ i,j,l \in \{0, \dots, p\}; \\ j+2l \geq 2}} \frac{p!}{i!j!l!} 2^j \alpha_k^{j+2l} (L_1 + L_2)^{j+2l}) \delta_{k-1}'^{2p} \quad (37)$$

958 Similar to (33), we have

$$\sum_{\substack{i+j+l=p; \\ i,j,l \in \{0, \dots, p\}; \\ j+2l \geq 2}} \frac{p!}{i!j!l!} 2^j \alpha_k^{j+2l} (L_1 + L_2)^{j+2l} \delta_{k-1}'^{2p} \leq \alpha_k^2 p^2 (L_1 + L_2)^2 (1 + 3\alpha_k (L_1 + L_2))^p$$

959 Enrolling recurrence (37), we get

$$\begin{aligned} \mathbb{E}[\delta_k'^{2p}] &\leq \exp \left\{ -2p\mu \sum_{i=1}^k \alpha_i \right\} \exp \left\{ p^2 (L_1 + L_2)^2 \sum_{i=1}^k \alpha_i^2 (1 + 3\alpha_i (L_1 + L_2))^p \right\} \|\theta_0 - \theta^*\|^{2p} \\ &\leq c_6 \exp \left\{ -\frac{p\mu c_0}{1 - \gamma} (k + k_0)^{1-\gamma} \right\} \|\theta_0 - \theta^*\|^{2p}, \end{aligned}$$

960 where we have set

$$c_6 = \exp \left\{ \exp \left\{ 3pc_0(L_1 + L_2) \right\} \frac{p^2(L_1 + L_2)^2}{2\gamma - 1} + \frac{p\mu c_0}{1 - \gamma} k_0^{1-\gamma} \right\}.$$

961 It remains to note that

$$\begin{aligned} \mathbb{E}[\|\theta_k - \theta^*\|^{2p}] &\leq 2^{2p-1} \mathbb{E}[\|\theta'_k - \theta^*\|^{2p}] + 2^{2p-1} \mathbb{E}[\|\theta_k - \theta'_k\|^{2p}] \\ &\leq C_{2p,1} \exp \left\{ -\frac{p\mu c_0}{4} (k + k_0)^{1-\gamma} \right\} (\|\theta_0 - \theta^*\|^{2p} + \sigma_{2p}^{2p}) + C_{2p,2} \sigma_{2p}^{2p} \alpha_k^p. \end{aligned}$$

962

□

963 For validity of induction in Proposition 2, we need the following corollary.

964 **Corollary 2.** *Under the assumptions of Proposition 2, it holds that*

$$\mathbb{E}[\|\theta_k - \theta^*\|^{2p}] \leq D_{2p} (\|\theta_0 - \theta^*\|^{2p} + \sigma_{2p}^{2p}) \alpha_k^p,$$

965 where

$$D_{2p} = C_{2p,1} (1/c_0^p + C_{2p,2}) \left(\frac{4\gamma}{(1-\gamma)\mu pc_0 e} \right)^{\gamma p/(1-\gamma)}. \quad (38)$$

966 *Proof.* Define $C_5 = \left(\frac{4\gamma}{(1-\gamma)\mu pc_0 e} \right)^{\gamma p/(1-\gamma)} > 1$, then $\exp\{-\mu pc_0(k + k_0)^{1-\gamma}/4\} \leq C_5(k + k_0)^{-p\gamma}$,
967 and the statement follows. □

968 **Corollary 3.** *Assume A1, A3, A8(4) and A9. Then for any $k \in \mathbb{N}$ it holds that*

$$\mathbb{E}[\|\theta_k - \theta^*\|^4] \leq C_{4,1} \exp \left\{ -\frac{2\mu c_0}{4} k^{1-\gamma} \right\} (\|\theta_0 - \theta^*\|^4 + \sigma_4^4) + C_{4,2} \sigma_4^4 \alpha_k^2,$$

969 with

$$C_{4,1} = 2^3 \left(C_1(1/c_0 + C_2) \left(\frac{4\gamma}{(1-\gamma)\mu c_0 e} \right)^{\gamma/(1-\gamma)} (4c_0^{1/2} 2^{\gamma/2} + 2^\gamma + 4c_0)^2 c_0^2 + 1 \right) c_{2,4}$$

970 and

$$C_{4,2} = 2^3 C_1(1/c_0 + C_2) \left(\frac{4\gamma}{(1-\gamma)\mu c_0 e} \right)^{\gamma/(1-\gamma)} (4c_0^{1/2} 2^{\gamma/2} + 2^\gamma + 4c_0)^2 c_{2,5}.$$

971 Here C_1 and C_2 are defined in Lemma 6 and

$$\begin{aligned} c_{2,4} &= \left(\exp \left\{ \exp \left\{ 10c_0(L_1 + L_2) \right\} \frac{16(L_1 + L_2)^2}{2\gamma - 1} \right\} + 1 \right) \exp \left\{ \frac{2\mu c_0}{1 - \gamma} k_0^{1-\gamma} \right\} \frac{1}{3\gamma - 1}, \\ c_{2,5} &= \frac{2^{1+2\gamma}}{2\mu c_0}. \end{aligned}$$

972 *Proof.* The proof follows directly from Proposition 2 and Corollary 1. □

973 C Proof of Theorem 2

974 We first provide details of the expansion (12). Recall that the error of SGD approximation may be
975 rewritten as follows

$$\theta_k - \theta^* = (I - \alpha_k G)(\theta_{k-1} - \theta^*) - \alpha_k (H(\theta_{k-1}) + \eta(\xi_k) + g(\theta_{k-1}, \xi_k)). \quad (39)$$

976 Iteratively spinning this expression out we get

$$\theta_k - \theta^* = \prod_{j=1}^k (I - \alpha_j G)(\theta_0 - \theta^*) - \sum_{j=1}^k \alpha_j \prod_{i=j+1}^k (I - \alpha_i G)(H(\theta_{j-1}) + \eta(\xi_j) + g(\theta_{j-1}, \xi_j)).$$

977 Taking average of (39) and changing the order of summation, we obtain

$$\sqrt{n}(\bar{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}\alpha_0}Q_0(\theta_0 - \theta^*) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i(H(\theta_{i-1}) + \eta(\xi_i) + g(\theta_{i-1}, \xi_i)),$$

978 where Q_i is defined in (11). Finally, we obtain

$$\begin{aligned}\sqrt{n}(\bar{\theta}_n - \theta^*) &= W + D, \\ D &= \frac{1}{\sqrt{n}\alpha_0}Q_0(\theta_0 - \theta^*) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i g(\theta_{i-1}, \xi_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i H(\theta_{i-1}), \\ W &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i \eta(\xi_i).\end{aligned}\tag{40}$$

979 *Proof of Theorem 2.* We normalize the both parts of (12) by $\Sigma_n^{1/2}$ and obtain

$$\sqrt{n}\Sigma_n^{-\frac{1}{2}}(\bar{\theta}_n - \theta^*) = \sum_{i=1}^{n-1} \underbrace{\frac{\Sigma_n^{-\frac{1}{2}}}{\sqrt{n}}Q_i \eta(\xi_i)}_{w_i} + D_{n,1} + D_{n,2} + D_{n,3},$$

980 where we have set

$$\begin{aligned}D_{n,1} &= \frac{\Sigma_n^{-\frac{1}{2}}}{\sqrt{n}\alpha_0}Q_0(\theta_0 - \theta^*), \\ D_{n,2} &= -\frac{\Sigma_n^{-\frac{1}{2}}}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i H(\theta_{i-1}), \\ D_{n,3} &= -\frac{\Sigma_n^{-\frac{1}{2}}}{\sqrt{n}}\sum_{i=1}^{n-1}Q_i g(\theta_{i-1}, \xi_i).\end{aligned}$$

981 Also, for any $1 \leq i \leq n-1$ we construct

$$\begin{aligned}D_{n,1}^{(i)} &= \frac{\Sigma_n^{-1/2}}{\sqrt{n}\alpha_0}Q_0(\theta_0^{(i)} - \theta^*), \\ D_{n,2}^{(i)} &= -\frac{\Sigma_n^{-1/2}}{\sqrt{n}}\sum_{j=1}^{n-1}Q_j H(\theta_{j-1}^{(i)}), \\ D_{n,3}^{(i)} &= -\frac{\Sigma_n^{-1/2}}{\sqrt{n}}\sum_{j=1}^{n-1}Q_j g(\theta_{j-1}^{(i)}, \tilde{\xi}_j^{(i)}),\end{aligned}$$

982 where we set

$$\tilde{\xi}_j^{(i)} = \begin{cases} \xi_j, & \text{if } j \neq i \\ \xi'_j, & \text{if } j = i. \end{cases}$$

983 Define $D_n = D_{n,1} + D_{n,2} + D_{n,3}$, $D_n^{(i)} = D_{n,1}^{(i)} + D_{n,2}^{(i)} + D_{n,3}^{(i)}$, $W_n = \sum_{i=1}^{n-1} w_i$ and $\Upsilon_n =$
984 $\sum_{i=1}^n \mathbb{E}[\|\omega_i\|^3]$ (we keep the same notations as in the unnormalized setting for simplicity). Let
985 $Y \sim \mathcal{N}(0, I_d)$. Then, using [42, Theorem 2.1], we have

$$d_C(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*), Y) \leq 259d^{1/2}\Upsilon_n + 2\mathbb{E}\{\|W_n\|\|D_n\|\} + 2\sum_{i=1}^{n-1}\mathbb{E}[\|\omega_i\|\|D_n - D_n^{(i)}\|].$$

Note that $\mathbb{E}^{1/2}[\|W_n\|^2] = \sqrt{d}$. Applying Lemma 4, we get $\mathbb{E}^{1/2}\|w_i\|^2 \leq \frac{1}{\sqrt{n}}C_\Sigma C_Q \sigma_2$ and

$$\Upsilon_n \leq \frac{1}{\sqrt{n}}(C_\Sigma C_Q \sigma_4)^3.$$

986 Applying Hölder's inequality together with Lemma 7 and Lemma 10, we obtain

$$d_C(\sqrt{n}\Sigma_n^{-1/2}(\bar{\theta}_n - \theta^*), Y) \leq \frac{\sqrt{d}M_{3,1}}{\sqrt{n}} + \frac{M_{3,2}}{\sqrt{n}}(\|\theta_0 - \theta^*\| + \|\theta_0 - \theta^*\|^2 + \sigma_2 + \sigma_4^2) + M_{3,3}n^{1/2-\gamma} + M_{3,4}n^{-\gamma/2},$$

987 where

$$\begin{aligned} M_{3,1} &= 259(C_\Sigma C_Q \sigma_4)^3, \\ M_{3,2} &= 2\sqrt{d}M_{1,1} + C_\Sigma C_Q \sigma_2 M_{2,1}, \\ M_{3,3} &= 2\sqrt{d}M_{1,2}\sigma_4^2, \\ M_{3,4} &= (2\sqrt{d}M_{1,3} + M_{2,3}C_\Sigma C_Q \sigma_2)\sigma_2 + C_\Sigma C_Q M_{2,2}\sigma_4^2\sigma_2. \end{aligned}$$

988 Constants $M_{1,1}, M_{1,2}, M_{1,3}$ are defined in (43) and $M_{2,1}, M_{2,2}, M_{3,3}$ are defined in (47). We simplify
989 the last inequality and get the statement of the theorem with

$$\begin{aligned} C_4 &= \sqrt{d}M_{3,1} + M_{3,2}\|\theta_0 - \theta^*\| + \|\theta_0 - \theta^*\|^2 + \sigma_2 + \sigma_4^2, \\ C_5 &= M_{3,3}, \\ C_6 &= M_{3,4}. \end{aligned} \tag{41}$$

990

□

991 Define

$$\begin{aligned} T_1(A) &= 1 + \frac{1}{A^{1/(1-\gamma)}(1-\gamma)}\Gamma\left(\frac{1}{1-\gamma}\right), \\ T_2(A) &= 1 + \max\left(\exp\left\{\frac{1}{1-\gamma}\right\}\frac{1}{A^{1/(1-\gamma)}(1-\gamma)}\Gamma\left(\frac{1}{1-\gamma}\right), \frac{1}{A(1-\gamma)^2}\right). \end{aligned} \tag{42}$$

992 **Lemma 7.** Assume A1, A3, A8(4) and A9. Then it holds that

$$\mathbb{E}^{1/2}[\|D_n\|^2] \leq \frac{M_{1,1}}{\sqrt{n}}(\|\theta_0 - \theta^*\| + \|\theta_0 - \theta^*\|^2 + \sigma_2 + \sigma_4^2) + M_{1,2}\sigma_4^2n^{1/2-\gamma} + M_{1,3}\sigma_2n^{-\gamma/2},$$

993 where

$$\begin{aligned} M_{1,1} &= C_\Sigma C_Q \left(T_1\left(\frac{\mu c_0}{4}\right)(L_2 + L_H) \max(\sqrt{C_{4,1}}, \sqrt{C_1}) + k_0^\gamma/c_0 \right) \\ M_{1,2} &= C_\Sigma C_Q L_H \sqrt{C_{4,2}} c_0 \frac{1}{1-\gamma} \\ M_{1,3} &= C_\Sigma C_Q L_2 \sqrt{C_2} \sqrt{c_0} \sqrt{\frac{1}{1-\gamma}}, \end{aligned} \tag{43}$$

994 where $C_{4,1}$ and $C_{4,2}$ are defined in Corollary 3, C_1 and C_2 are defined in Lemma 6 and $T_1(\cdot)$ is
995 defined in eq. (42).

996 *Proof.* Using Minkowski's inequality and the definition of D_n , we obtain

$$\mathbb{E}^{1/2}[\|D_n\|^2] \leq \mathbb{E}^{1/2}[\|D_{n,1}\|^2] + \mathbb{E}^{1/2}[\|D_{n,2}\|^2] + \mathbb{E}^{1/2}[\|D_{n,3}\|^2],$$

997 and consider each of the terms $D_{n,1}, D_{n,2}, D_{n,3}$ separately. Applying Lemma 4, we get

$$\mathbb{E}^{1/2}[\|D_{n,1}\|^2] \leq \frac{C_\Sigma C_Q k_0^\gamma}{\sqrt{n}c_0} \|\theta_0 - \theta^*\|.$$

998 Now we consider the term $D_{n,2}$. Applying Minkowski's inequality, Lemma 4 and Lemma 5, we have

$$\mathbb{E}^{1/2}[\|D_{n,2}\|^2] \leq \frac{C_\Sigma C_Q}{\sqrt{n}} \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|H(\theta_{i-1})\|^2] \leq \frac{C_\Sigma C_Q L_H}{\sqrt{n}} \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|\theta_{i-1} - \theta^*\|^4].$$

999 For $D_{n,3}$ we note that $\{g(\theta_{i-1}, \xi_i)\}_{i=1}^{n-1}$ is a martingale difference with respect to \mathcal{F}_i . Hence, using
 1000 Lemma 4 and A8, we get

$$\mathbb{E}^{1/2}[\|D_{n,3}\|^2] \leq \frac{C_\Sigma C_Q}{\sqrt{n}} \left(\sum_{i=1}^{n-1} \mathbb{E}[\|g(\theta_{i-1}, \xi_i)\|^2] \right)^{1/2} \leq \frac{C_\Sigma C_Q L_2}{\sqrt{n}} \left(\mathbb{E} \left[\sum_{i=1}^{n-1} \|\theta_{i-1} - \theta^*\|^2 \right] \right)^{1/2}.$$

1001 Hence, it is enough to upper bound $\mathbb{E}[\|\theta_i - \theta^*\|^{2p}]$ for $p = 1$ and $p = 2$ and $i \in \{0, \dots, n-2\}$.
 1002 Using Lemma 6 and Lemma 3, we obtain

$$\begin{aligned} \left(\sum_{i=0}^{n-2} \mathbb{E}[\|\theta_i - \theta^*\|^2] \right)^{1/2} &\leq \left(\sum_{i=0}^{n-2} C_1 \exp \left\{ -\frac{\mu c_0}{4} (i + k_0)^{1-\gamma} \right\} [\|\theta_0 - \theta^*\|^2 + \sigma_2^2] + C_2 \sigma_2^2 \alpha_i \right)^{1/2} \\ &\leq \sqrt{C_1} \sqrt{T_1 \left(\frac{\mu c_0}{4} \right)} [\|\theta_0 - \theta^*\| + \sigma_2] + \sqrt{C_2} \sigma_2 \sqrt{c_0} \left(\frac{(n-2+k_0)^{1-\gamma} - (k_0-1)^{1-\gamma}}{1-\gamma} \right)^{1/2}, \end{aligned}$$

1003 where $T_1(\cdot)$ is defined in (42). Using Corollary 3 and Lemma 3, we get

$$\begin{aligned} \sum_{i=0}^{n-2} \mathbb{E}^{1/2}[\|\theta_i - \theta^*\|^4] &\leq \sum_{i=0}^{n-2} \sqrt{C_{4,1}} \exp \left\{ -\frac{\mu c_0}{4} i^{1-\gamma} \right\} [\|\theta_0 - \theta^*\|^2 + \sigma_4^2] + \sqrt{C_{4,2}} \sigma_4^2 \alpha_i \\ &\leq \sqrt{C_{4,1}} T_1 \left(\frac{\mu c_0}{4} \right) [\|\theta_0 - \theta^*\|^2 + \sigma_4^2] + \sqrt{C_{4,2}} \sigma_4^2 c_0 \left(\frac{(n-2+k_0)^{1-\gamma} - (k_0-1)^{1-\gamma}}{1-\gamma} \right). \end{aligned}$$

1004 We finish the proof, using simple inequality $(n-2+k_0)^{1-\gamma} - (k_0-1)^{1-\gamma} \leq n^{1-\gamma}$ \square

1005 Let $(\xi'_1, \dots, \xi'_{n-1})$ be an independent copy of $(\xi_1, \dots, \xi_{n-1})$. For each $1 \leq i \leq n-1$, we construct
 1006 the sequence $\theta_k^{(i)}$, $1 \leq k \leq n-1$, as follows:

$$\theta_k^{(i)} = \begin{cases} \theta_k, & \text{if } k < i \\ \theta_{k-1}^{(i)} - \alpha_k (\nabla f(\theta_{k-1}^{(i)}) + g(\theta_{k-1}^{(i)}, \xi'_k) + \eta(\xi'_k)), & \text{if } k = i \\ \theta_{k-1}^{(i)} - \alpha_k (\nabla f(\theta_{k-1}^{(i)}) + g(\theta_{k-1}^{(i)}, \xi_k) + \eta(\xi_k)), & \text{if } k > i. \end{cases} \quad (44)$$

1007 **Lemma 8.** Assume A1, A3, A8(2) and A9. Then for any $k \in \mathbb{N}$ and $1 \leq i \leq n-1$ it holds

$$\mathbb{E}[\|\theta_k^{(i)} - \theta_k\|^2] \leq \alpha_i^2 R_1 \exp \left\{ -2\mu \sum_{j=i+1}^k \alpha_j \right\} \left(R_2 \exp \left\{ -\frac{\mu c_0}{4} (i+k_0-1)^{1-\gamma} \right\} (\|\theta_0 - \theta^*\|^2 + \sigma_2^2) + R_3 \sigma_2^2 \right),$$

1008 where we have set

$$R_1 = 4 \exp \left\{ \frac{2c_0^2 (L_1 + L_2)^2}{2\gamma - 1} \right\}, \quad R_2 = L_2^2 C_1, \quad R_3 = (1 + C_2 L_2). \quad (45)$$

1009 And constant C_1 and C_2 are defined in Lemma 6.

1010 *Proof.* By construction (44), we have

$$\theta_k^{(i)} - \theta_k = \begin{cases} 0, & \text{if } k < i \\ -\alpha_k (g(\theta_{k-1}, \xi'_k) + \eta(\xi'_k) - g(\theta_{k-1}, \xi_k) - \eta(\xi_k)), & \text{if } k = i \\ \theta_{k-1}^{(i)} - \theta_{k-1} - \alpha_k (\nabla f(\theta_{k-1}^{(i)}) - \nabla f(\theta_{k-1}) + g(\theta_{k-1}^{(i)}, \xi_k) - g(\theta_{k-1}, \xi_k)), & \text{if } k > i \end{cases}$$

1011 Since ξ'_i is independent copy of ξ_i , we obtain

$$\begin{aligned} \mathbb{E}[\|\theta_k^{(i)} - \theta_k\|^2] &\stackrel{(a)}{\leq} 4\alpha_i^2 (L_2^2 \mathbb{E}[\|\theta_{i-1} - \theta^*\|^2] + \sigma_2^2) \\ &\stackrel{(b)}{\leq} 4\alpha_i^2 \left(L_2^2 C_1 \exp \left\{ -\frac{\mu c_0}{4} (i + k_0 - 1)^{1-\gamma} \right\} (\|\theta_0 - \theta^*\|^2 + \sigma_2^2) + (1 + C_2 L_2) \sigma_2^2 \right), \end{aligned}$$

1012 where in (a) we used A8, and in (b) we used Lemma 6 and $\alpha_{k-1}L_2 \leq 1$. For $k > i$, applying A8 and
 1013 A1, we have

$$\begin{aligned} \mathbb{E}[\|\theta_k^{(i)} - \theta_k\|^2 | \mathcal{F}_{k-1}] &\leq \|\theta_{k-1}^{(i)} - \theta_{k-1}\|^2 - 2\alpha_k \langle \theta_{k-1}^{(i)} - \theta_{k-1}, \nabla f(\theta_{k-1}^{(i)}) - \nabla f(\theta_{k-1}) \rangle \\ &\quad + 2\alpha_k^2 (L_1 + L_2)^2 \|\theta_{k-1}^{(i)} - \theta_{k-1}\|^2. \end{aligned}$$

1014 Taking expectation from both sides and applying A1 with Lemma 2(a), we obtain

$$\begin{aligned} \mathbb{E}[\|\theta_k^{(i)} - \theta_k\|^2] &\leq (1 - 2\alpha_k\mu + 2\alpha_k^2(L_1 + L_2)^2) \mathbb{E}[\|\theta_{k-1}^{(i)} - \theta_{k-1}\|^2] \\ &\leq \exp\left\{\frac{2c_0^2(L_1 + L_2)^2}{2\gamma - 1}\right\} \exp\left\{-2\mu \sum_{j=i+1}^k \alpha_j\right\} \mathbb{E}[\|\theta_i^{(i)} - \theta_i\|^2]. \end{aligned}$$

1015 Combining the above inequalities completes the proof. \square

1016 **Lemma 9.** Assume A1, A3, A8(4) and A9. Then for any $k \in \mathbb{N}$ and $1 \leq i \leq n - 1$ it holds

$$\mathbb{E}[\|\theta_k^{(i)} - \theta_k\|^4] \leq \alpha_i^4 R_{4,1} \exp\left\{-4\mu \sum_{j=i+1}^k \alpha_j\right\} \left(R_{4,2} \exp\left\{-\frac{2\mu c_0}{4}(i+k_0-1)^{1-\gamma}\right\} (\|\theta_0 - \theta^*\|^4 + \sigma_4^4) + R_{4,3} \sigma_4^4\right)$$

1017 where we have set

$$R_{4,1} = 64 \exp\left\{\frac{4(L_1 + L_2)^2(1 + 3c_0(L_1 + L_2))^2}{2\gamma - 1}\right\}, \quad R_{4,2} = L_2^4 C_{4,1}, \quad R_{4,3} = 1 + L_2^2 C_{4,2}. \quad (46)$$

1018 And constant $C_{4,1}$, $C_{4,2}$ are defined in Corollary 3.

1019 *Proof.* Repeating the proof of the Lemma 8 for $k = i$, we get

$$\begin{aligned} \mathbb{E}[\|\theta_i^{(i)} - \theta_i\|^4] &\leq 64\alpha_i^4 (L_2^4 \mathbb{E}[\|\theta_{i-1} - \theta^*\|^4] + \sigma_4^4) \\ &\leq 64\alpha_i^4 \left(L_2^4 C_{4,1} \exp\left\{-\frac{2\mu c_0}{4}(i+k_0-1)^{1-\gamma}\right\} (\|\theta_0 - \theta^*\|^4 + \sigma_4^4) + (1 + L_2^2 C_{4,2}) \sigma_4^4\right). \end{aligned}$$

1020 For $k > i$ we denote $\delta_k^{(i)} = \|\theta_k^{(i)} - \theta_k\|$, similar to (37), we obtain

$$\mathbb{E}[\{\delta_k^{(i)}\}^4 | \mathcal{F}_{k-1}] \leq (1 - 4\mu\alpha_k + 4\alpha_k^2(L_1 + L_2)^2(1 + 3c_0(L_1 + L_2))^2) \{\delta_{k-1}^{(i)}\}^4.$$

1021 Using Lemma 2(a), we obtain

$$\mathbb{E}[\{\delta_k^{(i)}\}^4] \leq \exp\left\{\frac{4(L_1 + L_2)^2(1 + 3c_0(L_1 + L_2))^2}{2\gamma - 1}\right\} \exp\left\{-4\mu \sum_{j=i+1}^k \alpha_j\right\} \mathbb{E}[\|\theta_i^{(i)} - \theta_i\|^4].$$

1022 Combining the above inequalities completes the proof. \square

1023 **Lemma 10.** Assume A1, A3, A8(4) and A9. Then it holds that

$$\sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_n - D_n^{(i)}\|^2] \leq \frac{M_{2,1}}{\sqrt{n}} (\|\theta_0 - \theta^*\| + \|\theta_0 - \theta^*\|^2 + \sigma_2 + \sigma_4^2) + M_{2,2} \sigma_4^2 n^{1/2-\gamma} + M_{2,3} \sigma_2 n^{1/2-\gamma/2},$$

1024 where

$$\begin{aligned} M_{2,1} &= C_\Sigma C_Q T_1 \left(\frac{\mu c_0}{8}\right) T_2 \left(\frac{\mu c_0}{1-\gamma}\right) (L_2 + L_H) \max(\sqrt{2(C_1 + c_0^2 k_0^{-\gamma} R_1 R_2)}, c^2 k_0^{-\gamma} \sqrt{R_{4,1} R_{4,2}}) \\ M_{2,2} &= C_\Sigma C_Q L_H c_0 \sqrt{R_{4,1} R_{4,3}} T_2 \left(\frac{\mu c_0}{1-\gamma}\right) \frac{1}{1-\gamma} \\ M_{2,3} &= \sqrt{2} C_\Sigma C_Q L_2 \sqrt{C_2 + R_1 R_3 c_0} T_2 \left(\frac{\mu c_0}{1-\gamma}\right) \frac{1}{1-\gamma/2}. \end{aligned} \quad (47)$$

1025 Constants R_1, R_2, R_3 are defined in (45) and constants $R_{4,1}, R_{4,2}, R_{4,3}$ are defined (46).

1026 *Proof.* Using Minkowski's inequality and the definition of D_n and $D_n^{(i)}$, we obtain

$$\sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_n - D_n^{(i)}\|^2] \leq \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_{n,2} - D_{n,2}^{(i)}\|^2] + \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_{n,3} - D_{n,3}^{(i)}\|^2]$$

1027 Define $\mathcal{F}_j^{(i)} = \mathcal{F}_j$ if $j \leq i$ and $\mathcal{F}_j^{(i)} = \sigma(\mathcal{F}_j \vee \sigma(\xi'_i))$ otherwise. Then $\{g(\theta_{j-1}, \xi_j) - g(\theta_{j-1}^{(i)}, \tilde{\xi}_j)\}_{j=1}^{n-1}$
 1028 is a martingale difference with respect to $\mathcal{F}_j^{(i)}$. Hence, we have, using Lemma 4 and the fact that
 1029 $\theta_{j-1} = \theta_{j-1}^{(i)}$ for $j \leq i$, we obtain that

$$\begin{aligned} \mathbb{E}[\|D_{n,3} - D_{n,3}^{(i)}\|^2] &= \mathbb{E}\left\|\frac{\Sigma_n^{-1/2}}{\sqrt{n}} \sum_{j=1}^{n-1} Q_j(g(\theta_{j-1}, \xi_j) - g(\theta_{j-1}^{(i)}, \tilde{\xi}_j))\right\|^2 \\ &\leq \frac{C_\Sigma^2 C_Q^2}{n} \mathbb{E}[\|g(\theta_{i-1}, \xi_i) - g(\theta_{i-1}, \xi'_i)\|^2] + \frac{C_\Sigma^2 C_Q^2}{n} \sum_{j=i+1}^{n-1} \mathbb{E}[\|g(\theta_{j-1}, \xi_j) - g(\theta_{j-1}^{(i)}, \xi_j)\|^2]. \end{aligned}$$

1030 Using A8 and Lemma 4, we get

$$\mathbb{E}[\|D_{n,3} - D_{n,3}^{(i)}\|^2] \leq \frac{2C_\Sigma^2 C_Q^2 L_2^2}{n} \mathbb{E}[\|\theta_{i-1} - \theta^*\|^2] + \frac{C_\Sigma^2 C_Q^2 L_2^2}{n} \sum_{j=i+1}^{n-1} \mathbb{E}[\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^2].$$

1031 Using Lemma 8 and Lemma 3, we obtain

$$\begin{aligned} \sum_{j=i+1}^{n-1} \mathbb{E}[\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^2] &\leq R_1 R_2 \exp\left\{-\frac{\mu c_0}{4}(i + k_0 - 1)^{1-\gamma}\right\} \alpha_i^2 (\|\theta_0 - \theta^*\|^2 + \sigma_2^2) T_2\left(\frac{\mu c_0}{1-\gamma}\right) (i + k_0)^\gamma \\ &\quad + R_1 R_3 \sigma_2^2 \alpha_i^2 T_2\left(\frac{\mu c_0}{1-\gamma}\right) (i + k_0)^\gamma \\ &\leq R_1 R_3 \sigma_2^2 c_0 T_2\left(\frac{\mu c_0}{1-\gamma}\right) \alpha_i + R_1 R_2 c_0^2 k_0^{-\gamma} T_2\left(\frac{\mu c_0}{1-\gamma}\right) \exp\left\{-\frac{\mu c_0}{4}(i + k_0 - 1)^{1-\gamma}\right\} (\|\theta_0 - \theta^*\|^2 + \sigma_2^2). \end{aligned}$$

1032 Combining inequalities above, we get

$$\begin{aligned} \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_{n,3} - D_{n,3}^{(i)}\|^2] &\leq \frac{\sqrt{2} C_\Sigma C_Q L_2}{\sqrt{n}} \sqrt{C_1 + c_0^2 k_0^{-\gamma} R_1 R_2 T_2\left(\frac{\mu c_0}{1-\gamma}\right) T_1\left(\frac{\mu c_0}{8}\right) (\|\theta_0 - \theta^*\| + \sigma_2)} \\ &\quad + \frac{\sqrt{2} C_\Sigma C_Q L_2}{\sqrt{n}} \sqrt{C_2 + R_1 R_3 c_0 T_2\left(\frac{\mu c_0}{1-\gamma}\right) \sigma_2^2} \left(\frac{(n + k_0 - 2)^{1-\gamma/2} - (k_0 - 1)^{1-\gamma/2}}{1 - \gamma/2}\right). \end{aligned}$$

1033 We now proceed with $\sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_{n,2} - D_{n,2}^{(i)}\|^2]$. Using Minkowski's inequality together with
 1034 Lemma 4 and Lemma 5, we get

$$\mathbb{E}^{1/2}[\|D_{n,2} - D_{n,2}^{(i)}\|^2] \leq \frac{C_\Sigma C_Q L_H}{\sqrt{n}} \sum_{j=i+1}^{n-1} \mathbb{E}^{1/2}[\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^4].$$

1035 Applying Lemma 9 and Lemma 3, we get using that $\alpha_i^2(i + k_0)^\gamma \leq \alpha_0^2 k_0^{-\gamma}$ that

$$\begin{aligned} \sum_{j=i+1}^{n-1} \mathbb{E}^{1/2}[\|\theta_{j-1} - \theta_{j-1}^{(i)}\|^4] &\leq c_0^2 k_0^{-\gamma} \sqrt{R_{4,1} R_{4,2}} T_2\left(\frac{\mu c_0}{1-\gamma}\right) \exp\left\{-\frac{\mu c_0}{4}(i + k_0 - 1)^{1-\gamma}\right\} (\|\theta_0 - \theta^*\|^2 + \sigma_4^2) \\ &\quad + \alpha_i c_0 \sqrt{R_{4,1} R_{4,3}} T_2\left(\frac{\mu c_0}{1-\gamma}\right) \sigma_4^2. \end{aligned}$$

1036 Finally, applying Lemma 3, we get

$$\begin{aligned} \sum_{i=1}^{n-1} \mathbb{E}^{1/2}[\|D_{n,2} - D_{n,2}^{(i)}\|^2] &\leq \frac{C_\Sigma C_Q L_H}{\sqrt{n}} c_0^2 k_0^{-\gamma} \sqrt{R_{4,1} R_{4,2}} T_2\left(\frac{\mu c_0}{1-\gamma}\right) T_1\left(\frac{\mu c_0}{4}\right) (\|\theta_0 - \theta^*\|^2 + \sigma_4^2) \\ &\quad + \frac{C_\Sigma C_Q L_H}{\sqrt{n}} c_0 \sqrt{R_{4,1} R_{4,3}} T_2\left(\frac{\mu c_0}{1-\gamma}\right) \sigma_4^2 \left(\frac{(n + k_0 - 2)^{1-\gamma} - (k_0 - 1)^{1-\gamma}}{1 - \gamma}\right). \end{aligned}$$

1037 We finish the proof, using that $(n - 2 + k_0)^\beta - (k_0 - 1)^\beta \leq n^\beta$ for $\beta \in (0, 1)$ \square

1038 D Proof of quantitative Polyak-Juditsky CLT

1039 D.1 Proof of Lemma 1

1040 By definition of Σ_n and Σ_∞ we may write

$$\begin{aligned} \Sigma_n - \Sigma_\infty &= \underbrace{\frac{1}{n} \sum_{t=1}^{n-1} (Q_t - G^{-1}) \Sigma_\xi G^{-\top}}_{D_1} + \underbrace{\frac{1}{n} \sum_{t=1}^{n-1} G^{-1} \Sigma_\xi (Q_t - G^{-1})^\top}_{D_2} + \\ &\quad + \underbrace{\frac{1}{n} \sum_{t=1}^{n-1} (Q_t - G^{-1}) \Sigma_\xi (Q_t - G^{-1})^\top}_{D_2} - \frac{1}{n} \Sigma_\infty . \end{aligned}$$

1041 The following lemma is an analogue of [46, pp. 26-30].

1042 **Lemma 11.** *The following identities hold*

$$Q_i - G^{-1} = S_i - G^{-1} G_{i:n-1}^{(\alpha)}, \quad S_i = \sum_{j=i+1}^{n-1} (\alpha_i - \alpha_j) G_{i+1:j-1}^{(\alpha)}, \quad (48)$$

1043 and

$$\sum_{i=1}^{n-1} (Q_i - G^{-1}) = -G^{-1} \sum_{j=1}^{n-1} G_{1:j}^{(\alpha)}, \quad (49)$$

1044 where

$$G_{i:j}^{(\alpha)} = \prod_{k=i}^j (I - \alpha_k G)$$

1045 *Proof.* To prove (49) we first change the order of summation and then use the properties of the
1046 telescopic sums we get

$$\begin{aligned} \sum_{i=1}^{n-1} Q_i &= \sum_{i=1}^{n-1} \alpha_i \sum_{j=i}^{n-1} \prod_{k=i+1}^j (I - \alpha_k G) = \sum_{j=1}^{n-1} \sum_{i=1}^j \alpha_i \prod_{k=i+1}^j (I - \alpha_k G) \\ &= \sum_{j=1}^{n-1} \sum_{i=1}^j G^{-1} \left(\prod_{k=i+1}^j (I - \alpha_k G) - \prod_{k=i}^j (I - \alpha_k G) \right) = G^{-1} \sum_{j=1}^{n-1} \left(I - \prod_{k=1}^j (I - \alpha_k G) \right). \end{aligned}$$

1047 The proof of (48) could be obtained by the following arguments. Note that

$$\begin{aligned} \alpha_i G Q_i &= Q_i - (I - \alpha_i G) Q_i = \\ &= \alpha_i I + \alpha_i \sum_{j=i+1}^{n-1} \prod_{k=i+1}^j (I - \alpha_k G) - \alpha_i \sum_{j=i+1}^{n-1} \prod_{k=i}^j (I - \alpha_k G) - \alpha_i \prod_{k=i}^{n-1} (I - \alpha_k G). \end{aligned}$$

It remains to note that

$$\prod_{k=i+1}^j (I - \alpha_k G) - \prod_{k=i}^{j-1} (I - \alpha_k G) = (\alpha_i - \alpha_j) G \prod_{k=i+1}^{j-1} (I - \alpha_k G).$$

1048 The last two equations imply (48). □

1049 **Lemma 12.** *It holds that*

(a)

$$\|S_i\| \leq C_S (i + k_0)^{\gamma-1},$$

where

$$C_S = 2c_0 \exp \left\{ \frac{\mu c_0}{k_0^\gamma} \right\} \left(2^{\gamma/(1-\gamma)} \frac{1}{\mu c_0} + \left(\frac{1}{\mu c_0} \right)^{1/(1-\gamma)} \Gamma \left(\frac{1}{1-\gamma} \right) \right).$$

(b)

$$\sum_{i=1}^{n-1} \|G_{i:n-1}^{(\alpha)}\|^2 \leq \frac{1}{1 - (1 - c_0\mu(n + k_0 - 2)^{-\gamma})^2}$$

(c)

$$\left\| \sum_{i=1}^{n-1} G_{i:n-1}^{(\alpha)} \right\| \leq \frac{k_0^\gamma n^\gamma}{c_0\mu}$$

1050 *Proof.* For simplicity we define $m_i^j = \sum_{k=i}^j (k + k_0)^{-\gamma}$. Note that

$$\left\| \sum_{j=i+1}^{n-1} (\alpha_i - \alpha_j) G_{i+1:j-1}^{(\alpha)} \right\| \leq \sum_{j=i}^{n-2} \frac{c_0}{(j + k_0 + 1)^\gamma} \left(\left(\frac{j + k_0 + 1}{i + k_0} \right)^\gamma - 1 \right) \exp\{-\mu c_0 m_{i+1}^j\}$$

1051 Following the proof of [46, Lemma A.5], we have

$$\left(\frac{j + k_0 + 1}{i + k_0} \right)^\gamma - 1 \leq (i + k_0)^{\gamma-1} \left(1 + (1 - \gamma) m_i^j \right)^{\gamma/(1-\gamma)}$$

1052 Hence, we obtain

$$\begin{aligned} \|S_i\| &\leq c_0 (i + k_0)^{\gamma-1} \sum_{j=i}^{n-2} \frac{1}{(j + k_0 + 1)^\gamma} \left(1 + (1 - \gamma) m_i^j \right)^{\gamma/(1-\gamma)} \exp\{-\mu c_0 m_{i+1}^j\} \\ &\leq c_0 (i + k_0)^{\gamma-1} \sum_{j=i}^{n-2} \frac{1}{(j + k_0)^\gamma} \left(1 + (1 - \gamma) m_i^j \right)^{\gamma/(1-\gamma)} \exp\{\mu c_0 (k_0 + i)^{-\gamma}\} \exp\{-\mu c_0 m_i^j\} \\ &\leq c_0 \exp\left\{\frac{\mu c_0}{k_0^\gamma}\right\} (i + k_0)^{\gamma-1} \sum_{j=i}^{n-2} (m_i^j - m_i^{j-1}) \left(1 + (1 - \gamma) m_i^j \right)^{\gamma/(1-\gamma)} \exp\{-\mu c_0 m_i^j\} \\ &\leq 2c_0 \exp\left\{\frac{\mu c_0}{k_0^\gamma}\right\} (i + k_0)^{\gamma-1} \int_0^{+\infty} \left(1 + (1 - \gamma) m \right)^{\gamma/(1-\gamma)} \exp\{-\mu c_0 m\} dm \\ &\leq 2c_0 \exp\left\{\frac{\mu c_0}{k_0^\gamma}\right\} (i + k_0)^{\gamma-1} \left(2^{\gamma/(1-\gamma)} \frac{1}{\mu c_0} + \left(\frac{1}{\mu c_0} \right)^{1/(1-\gamma)} \Gamma\left(\frac{1}{1-\gamma}\right) \right). \end{aligned}$$

1053 Note that

$$\begin{aligned} \left\| \sum_{i=1}^{n-1} G_{i:n-1}^{(\alpha)} \right\| &\leq \sum_{i=1}^{n-1} \prod_{k=i}^{n-1} (1 - \alpha_k \mu) = \sum_{i=1}^{n-1} \prod_{k=i}^{n-1} \alpha_{i-1}^{-1} \alpha_{i-1} (1 - \alpha_k \mu) \\ &\leq \frac{(k_0 + n - 2)^\gamma}{c_0 \mu} \sum_{i=1}^{n-1} \left(\prod_{k=i}^{n-1} (1 - \alpha_k \mu) - \prod_{k=i-1}^{n-1} (1 - \alpha_k \mu) \right) \leq \frac{k_0^\gamma n^\gamma}{\mu c_0}, \end{aligned}$$

1054 where in the last inequality we use that $(k_0 + n - 2)^\gamma \leq (k_0 n)^\gamma$. Bound for $\sum_{i=1}^{n-1} \|G_{i:n-1}^{(\alpha)}\|^2$ is
1055 obtained similarly to $\left\| \sum_{i=1}^{n-1} G_{i:n-1}^{(\alpha)} \right\|$. \square

1056 To finish the proof of Lemma 1 we need to bound D_1, D_2 . By (49) we obtain

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n-1} (Q_i - G^{-1}) \Sigma_\xi G^{-\top} \right\| &= \left\| -\frac{1}{n} G^{-1} \sum_{j=1}^{n-1} G_{1:j}^{(\alpha)} \Sigma_\xi G^{-\top} \right\| \\ &= \|n^{-1} \Sigma_\infty \sum_{j=1}^{n-1} G_{1:j}^{(\alpha)}\| \leq n^{-1} \|\Sigma_\infty\| \cdot \left\| \sum_{j=1}^{n-1} G_{1:j}^{(\alpha)} \right\|. \end{aligned}$$

1057 It remains to apply Lemma 4 which gives

$$\left\| \frac{1}{n} \sum_{i=1}^{n-1} (Q_i - G^{-1}) \Sigma_\xi G^{-\top} \right\| \leq \|\Sigma_\infty\| C_Q \frac{k_0^\gamma n^{\gamma-1}}{c_0}$$

1058 Hence,

$$\|D_1\| \leq 2 \|\Sigma_\infty\| C_Q \frac{k_0^\gamma n^{\gamma-1}}{c_0}$$

1059 To bound D_2 we use (48) which gives

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n-1} (Q_i - G^{-1}) \Sigma_\xi (Q_i - G^{-1})^\top \\ &= n^{-1} \sum_{i=1}^{n-1} (S_i - G^{-1} \prod_{k=i}^{n-1} (I - \alpha_k G)) \Sigma_\xi (S_i - G^{-1} \prod_{k=i}^{n-1} (I - \alpha_k G))^\top \\ &= n^{-1} \underbrace{\sum_{i=1}^{n-1} S_i \Sigma_\xi S_i^\top}_{D_{21}} + n^{-1} \underbrace{\sum_{i=1}^{n-1} G^{-1} \prod_{k=i}^{n-1} (I - \alpha_k G) \Sigma_\xi G^{-\top} \prod_{k=i}^{n-1} (I - \alpha_k G)^\top}_{D_{22}} \\ &\quad - n^{-1} \underbrace{\sum_{i=1}^{n-1} G^{-1} \prod_{k=i}^{n-1} (I - \alpha_k G) \cdot \Sigma_\xi S_i^\top}_{D_{23}} - n^{-1} \underbrace{\sum_{i=1}^{n-1} S_i \Sigma_\xi G^{-\top} \prod_{k=i}^{n-1} (I - \alpha_k G)^\top}_{D_{24}}. \end{aligned}$$

1060 To bound D_{21} we use Lemma 12, and obtain

$$\begin{aligned} \|D_{21}\| &= \|n^{-1} \sum_{i=1}^{n-1} S_i \Sigma_\xi S_i^\top\| \leq n^{-1} \sum_{i=1}^{n-1} \|\Sigma_\xi\| \|S_i\|^2 \\ &\leq n^{-1} \|\Sigma_\xi\| C_S^2 \sum_{i=1}^{n-1} (i + k_0)^{2(\gamma-1)} \\ &\leq n^{-1} \|\Sigma_\xi\| C_S^2 \frac{(n + k_0 - 1)^{2\gamma-1} - k_0^{2\gamma-1}}{2\gamma - 1} \\ &\leq \|\Sigma_\xi\| C_S^2 \frac{n^{2(\gamma-1)}}{2\gamma - 1} \end{aligned}$$

1061 The bound for D_{22} follows from Lemma 12

$$\begin{aligned} \|D_{22}\| &= \|n^{-1} \sum_{i=1}^{n-1} \prod_{k=i}^{n-1} (I - \alpha_k G) G^{-1} \Sigma_\xi G^{-\top} \prod_{k=i}^{n-1} (I - \alpha_k G)^\top\| \leq n^{-1} \|\Sigma_\infty\| \sum_{i=1}^{n-1} \|G_{i:n-1}^{(\alpha)}\|^2 \\ &\leq n^{-1} \frac{\|\Sigma_\infty\|}{2c_0\mu(n + k_0 - 2)^{-\gamma} - c_0^2\mu^2(n + k_0 - 2)^{-2\gamma}} \leq \|\Sigma_\infty\| k_0^\gamma \frac{n^{\gamma-1}}{c_0\mu}. \end{aligned}$$

1062 Since $D_{23} = D_{24}^\top$, we concentrate on $\|D_{24}\|$. Lemma 12 immediately imply

$$\begin{aligned} \|D_{24}\| &\leq n^{-1} \|\Sigma_\xi G^{-\top}\| \sum_{i=1}^{n-1} \|S_i\| \left\| \prod_{k=i}^{n-1} (I - \alpha_k G)^\top \right\| \\ &\leq n^{-1} \|\Sigma_\xi\| \frac{1}{\mu} C_S \sum_{i=1}^{n-1} (i + k_0)^{\gamma-1} \prod_{k=i}^{n-1} (1 - \mu \frac{c_0}{(k + k_0)^\gamma}) \\ &\leq n^{-1} \|\Sigma_\xi\| \frac{1}{\mu} C_S \sum_{i=1}^{n-1} (i + k_0)^{2\gamma-1} (i + k_0)^{-\gamma} \prod_{k=i+1}^{n-1} (1 - \mu \frac{c_0}{(k + k_0)^\gamma}) \\ &\leq \|\Sigma_\xi\| C_S k_0^{2\gamma-1} \frac{n^{2(\gamma-1)}}{\mu^2 c_0} \end{aligned}$$

1063 Combining all inequalities above, we obtain

$$\|\Sigma_n - \Sigma_\infty\| \leq C'_\infty n^{\gamma-1}, \quad (50)$$

1064 where

$$C'_\infty = \left(\frac{k_0^\gamma}{c_0 \mu} + 2C_Q \frac{k_0^\gamma}{c_0} + 1 \right) \|\Sigma_\infty\| + \left(C_S^2 \frac{1}{2\gamma - 1} + C_S \frac{k_0^{2\gamma-1}}{\mu^2 c_0} \right) \|\Sigma_\xi\|.$$

1065 To finish the proof it remains to apply Lemma 13, since

$$3/2 \|\Sigma_n^{-1/2} \Sigma_\infty \Sigma_n^{-1/2} - I\|_F \leq C_\infty n^{\gamma-1}, \text{ where } C_\infty = 3/2 \sqrt{d} C_\Sigma^2 C'_\infty. \quad (51)$$

1066 D.2 Gaussian comparison lemma

1067 There are quite a lot of works devoted to the comparison of Gaussian measures with different
 1068 covariance matrices and means. Among others we note the works [5], [19], [13]. In this work we
 1069 will use the result from [13, Theorem 1.1], which performs comparison in terms of the total variation
 1070 distance. Recall that the total variation distance between probability measures μ and ν , defined on a
 1071 measurable space (X, \mathcal{X}) , is defined as

$$d_{TV}(\mu, \nu) = \sup_{B \in \mathcal{X}} |\mu(B) - \nu(B)|.$$

1072 With a slight abuse of notation, when X and Y are random vectors with distributions μ and ν ,
 1073 respectively, we write $d_{TV}(X, Y)$ instead of $d_{TV}(\mu, \nu)$. The following lemma holds:

1074 **Lemma 13.** *Let Σ_1 and Σ_2 be positive definite covariance matrices in $\mathbb{R}^{d \times d}$. Let $X \sim \mathcal{N}(0, \Sigma_1)$
 1075 and $Y \sim \mathcal{N}(0, \Sigma_2)$. Then*

$$d_{TV}(X, Y) \leq \frac{3}{2} \|\Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2} - I_d\|_F.$$

1076 Recall that our primary aim in this paper is to obtain convergence bounds in the convex distance

$$d_C(X, Y) = \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}(X \in B) - \mathbb{P}(Y \in B)|,$$

1077 where $\mathcal{C}(\mathbb{R}^d)$ is a set of convex sets on \mathbb{R}^d . We can immediately obtain the result for convex distance
 1078 from Lemma 13, since

$$d_C(X, Y) \leq d_{TV}(X, Y).$$

1079 For this purpose Lemma 13 is sufficient. At the same time, this inequality can be significantly
 1080 improved if instead of the set of all convex sets we take the set of rectangles or the set of all balls (in
 1081 Euclidean metric).

1082 E Bootstrap validity proof

1083 E.1 Example of distribution satisfying A5

1084 To construct examples of distributions satisfying the above assumption, one can use the beta distribu-
 1085 tion, which is defined on $[0, 1]$, and then shift and scale it. Set $W = a + bX$ where $X \sim \text{Beta}(\alpha, \beta)$
 1086 and $a, b > 0$. We have $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ and $a \leq W \leq a+b$ a.s. By solving
 1087 (for a and b) the equations $\mathbb{E}[W] = a + b\mathbb{E}[X] = 1$ and $\text{Var}(W) = b^2\text{Var}(X) = 1$, we derive
 1088 $b = 1/\sqrt{\text{Var}(X)}$ and $a = 1 - \mathbb{E}[X]/\sqrt{\text{Var}(X)}$. Note that $a > 0$ provided $\alpha + \beta + 1 < \beta/\alpha$.

1089 E.2 From non-linear to linear statistics

1090 In this section we prove (19). We start from the definition of an isoperimetric constant. Define

$$A^\varepsilon = \{x \in \mathbb{R}^d : \rho_A(x) \leq \varepsilon\} \quad \text{and} \quad A^{-\varepsilon} = \{x \in A : B_\varepsilon(x) \subset A\},$$

1091 where $\rho_A(x) = \inf_{y \in A} \|x - y\|$ is the distance between $A \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, and

$$B_\varepsilon(x) = \{y \in \mathbb{R}^d : \|x - y\| \leq \varepsilon\}.$$

1092 For some class \mathcal{A} of subsets of \mathbb{R}^d we define its isoperimetric constant $a_d(\mathcal{A})$ (depending only on d
 1093 and \mathcal{A}) as follows: for all $A \in \mathcal{A}$ and $\varepsilon > 0$,

$$\mathbb{P}\{Y \in A^\varepsilon \setminus A\} \leq a_d \varepsilon, \quad \mathbb{P}\{Y \in A \setminus A^{-\varepsilon}\} \leq a_d \varepsilon$$

1094 where Y follows the standard Gaussian distribution on \mathbb{R}^d . [4] has proved that

$$e^{-1}\sqrt{\ln d} \leq \sup_{A \in \mathcal{C}} \int_{\partial A} p(x) \, ds \leq 4d^{1/4}, \quad (52)$$

1095 where $p(x)$ is the standard normal d -dimensional density and ds is the surface measure on the
1096 boundary ∂A of A . Using (52) one can show that for the class of convex sets

$$e^{-1}\sqrt{\ln d} \leq a_d(\mathcal{C}(\mathbb{R}^d)) \leq 4d^{1/4}.$$

1097 We denote $c_d = a_d(\mathcal{C}(\mathbb{R}^d))$.

1098 **Proposition 3.** *Let ν be a standard Gaussian measure in \mathbb{R}^d . Then for any random vectors X, Y
1099 taking values in \mathbb{R}^d , and any $p \geq 1$,*

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}(X + Y \in B) - \nu(B)| \leq \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}(X \in B) - \nu(B)| + 2c_d^{p/(p+1)} \mathbb{E}^{1/(p+1)}[\|Y\|^p],$$

1100 where c_d is the isoperimetric constant of class $\mathcal{C}(\mathbb{R}^d)$.

1101 *Proof.* Let $\varepsilon \geq 0$. Define $\rho(B) = \mathbb{P}(X + Y \in B) - \nu(B)$. Let B be such that $\rho(B) \geq 0$. By
1102 Markov's inequality

$$\begin{aligned} \rho(B) &\leq \mathbb{P}(X + Y \in B, |Y| \leq \varepsilon) + \frac{1}{\varepsilon^p} \mathbb{E}[\|Y\|^p] - \nu(B) \\ &\leq \sup_A |\mathbb{P}(X \in A) - \nu(A)| + \mathbb{P}(Y \in B^\varepsilon \setminus B) + \frac{1}{\varepsilon^p} \mathbb{E}[\|Y\|^p]. \end{aligned}$$

1103 Choosing

$$\varepsilon = \frac{1}{c_d^{1/(p+1)}} \mathbb{E}^{1/(p+1)}[\|Y\|^p] \quad (53)$$

1104 we obtain

$$\sup_B |\mathbb{P}(X + Y \in B) - \nu(B)| \leq \sup_B |\mathbb{P}(X \in B) - \nu(B)| + 2c_d^{p/(p+1)} \mathbb{E}^{1/(p+1)}[\|Y\|^p].$$

Assume now that $\rho(B) < 0$. We distinguish between $B^{-\varepsilon} = \emptyset$ or $B^{-\varepsilon} \neq \emptyset$. In the first case,
 $\mathbb{P}(Y \in B^{-\varepsilon}) = 0$ and

$$-\rho(B) \leq \gamma(B) = \mathbb{P}(Y \in B) - \mathbb{P}(Y \in B^{-\varepsilon}) = \mathbb{P}(Y \in B \setminus B^{-\varepsilon}) \leq c_d \varepsilon.$$

Finally, in the case $B^{-\varepsilon} \neq \emptyset$,

$$-\rho(B) \leq \sup_A |\mathbb{P}(X \in A) - \nu(A)| + \mathbb{P}(Y \in B \setminus B^{-\varepsilon}) + \frac{1}{\varepsilon^p} \mathbb{E}[\|Y\|^p].$$

1105 Taking ε as in (53) we conclude the proof. \square

1106 E.3 High probability bounds on the last iterate

1107 **Lemma 14.** *Assume A1, A2, A4, A5, A6. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ for
1108 any $k \in \{1, \dots, n\}$ it holds*

$$\|\theta_k^b - \theta^*\|^2 \leq \alpha_k K_1 \log\left(\frac{en}{\delta}\right),$$

1109 where

$$K_1 = \max\left(\frac{8W_{\max}^2(C_{1,\xi} + 2C_{2,\xi})^2}{W_{\min}\mu}, \frac{k_0^\gamma \|\theta_0 - \theta^*\|^2}{c_0}\right) \quad (54)$$

1110 *Proof.* Using (5), we have

$$\begin{aligned} \|\theta_k^b - \theta^*\|^2 &= \|\theta_{k-1}^b - \theta^*\|^2 - 2\alpha_k w_k \langle F(\theta_{k-1}^b, \xi_k), \theta_{k-1}^b - \theta^* \rangle + \alpha_k^2 w_k^2 \|\nabla F(\theta_{k-1}^b, \xi_k)\|^2 \\ &\leq \|\theta_{k-1}^b - \theta^*\|^2 - 2\alpha_k w_k \langle F(\theta_{k-1}^b, \xi_k) - F(\theta^*, \xi_k), \theta_{k-1}^b - \theta^* \rangle \\ &\quad - 2\alpha_k w_k \langle \eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle + 2\alpha_k^2 w_k^2 \|F(\theta_{k-1}^b, \xi_k) - F(\theta^*, \xi_k)\|^2 + 2\alpha_k^2 w_k^2 \|\eta(\xi_k)\|^2. \end{aligned}$$

1111 Using A4 and A6, we obtain

$$\begin{aligned}\|\theta_k^b - \theta^*\|^2 &\leq \|\theta_{k-1}^b - \theta^*\|^2 - 2\alpha_k w_k (1 - \alpha_k w_k L_4) \langle F(\theta_{k-1}^b, \xi_k) - F(\theta^*, \xi_k), \theta_{k-1}^b - \theta^* \rangle \\ &\quad - 2\alpha_k w_k \langle \eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle + 2\alpha_k^2 w_k^2 \|\eta(\xi_k)\|^2 \\ &\leq \|\theta_{k-1}^b - \theta^*\|^2 - \alpha_k w_k \langle F(\theta_{k-1}^b, \xi_k) - F(\theta^*, \xi_k), \theta_{k-1}^b - \theta^* \rangle \\ &\quad - 2\alpha_k w_k \langle \eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle + 2\alpha_k^2 w_k^2 \|\eta(\xi_k)\|^2\end{aligned}$$

1112 Using A1, we have

$$\|\theta_k^b - \theta^*\|^2 \leq (1 - \mu\alpha_k w_k) \|\theta_{k-1}^b - \theta^*\|^2 - \alpha_k w_k \langle g(\theta_{k-1}^b, \xi_k) + 2\eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle + 2\alpha_k^2 w_k^2 \|\eta(\xi_k)\|^2.$$

1113 Using A5, we get

$$\begin{aligned}\|\theta_k^b - \theta^*\|^2 &\leq (1 - \mu\alpha_k W_{\min}) \|\theta_{k-1}^b - \theta^*\|^2 - \alpha_k w_k \langle g(\theta_{k-1}^b, \xi_k) + 2\eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle \\ &\quad + 2\alpha_k^2 W_{\max}^2 \|\eta(\xi_k)\|^2.\end{aligned}\tag{55}$$

1114 Define $Y_k = \alpha_k^{-1} \|\theta_k^b - \theta^*\|^2$, and $\hat{X}_{k-1} = \frac{w_k \langle g(\theta_{k-1}^b, \xi_k) + 2\eta(\xi_k), \theta_{k-1}^b - \theta^* \rangle}{W_{\max}(C_{2,\xi} + 2C_{1,\xi}) \|\theta_{k-1}^b - \theta^*\|}$, then using (55), we obtain

$$Y_k \leq \alpha_k^{-1} \alpha_{k-1} (1 - \mu W_{\min} \alpha_k) Y_{k-1} - \sqrt{\alpha_{k-1}} W_{\max} (C_{2,\xi} + 2C_{1,\xi}) \hat{X}_{k-1} \sqrt{Y_{k-1}} + 2W_{\max}^2 \alpha_k C_{1,\xi}^2.$$

1115 Note that

$$\begin{aligned}\frac{\alpha_{k-1}}{\alpha_k} (1 - \mu W_{\min} \alpha_k) &= \left(\frac{k_0 + k}{k_0 + k - 1} \right)^\gamma - \frac{\mu W_{\min} c_0}{(k_0 + k - 1)^\gamma} \\ &\leq 1 + \frac{c_0(\gamma/c_0)}{k_0 + k - 1} - \frac{\mu W_{\min} c_0}{(k_0 + k - 1)^\gamma} \\ &= 1 - \alpha_{k-1} \left(\mu W_{\min} - \frac{(\gamma/c_0)}{(k_0 + k - 1)^{1-\gamma}} \right).\end{aligned}$$

1116 Since $k_0 \geq \left(\frac{2\gamma}{c_0 \mu W_{\min}} \right)^{1/(1-\gamma)}$, we have

$$Y_k \leq (1 - \frac{\mu W_{\min}}{2} \alpha_{k-1}) Y_{k-1} - \sqrt{\alpha_{k-1}} W_{\max} (C_{2,\xi} + 2C_{1,\xi}) \hat{X}_{k-1} \sqrt{Y_{k-1}} + 2W_{\max}^2 \alpha_k C_{1,\xi}^2.$$

1117 Note that using A8 and A2, we have

$$\begin{aligned}\mathbb{E}[\hat{X}_{k-1} | \tilde{\mathcal{F}}_{k-1}] &= 0 \\ \|\hat{X}_{k-1}\| &\leq \frac{\|w_i\| (\|g(\theta_{k-1}^b, \xi_k)\| + 2\|\eta(\xi_k)\|) \|\theta_{k-1}^b - \theta^*\|}{W_{\max}(C_{2,\xi} + 2C_{1,\xi}) \|\theta_{k-1}^b - \theta^*\|} \leq 1,\end{aligned}$$

1118 where $\tilde{\mathcal{F}}_{k-1}$ is defined in (18). Then using [20, Theorem 4.1], with probability at least $1 - \delta$ for
1119 $\forall k \in \{1, \dots, n\}$

$$Y_k \leq K_1 \log\left(\frac{en}{\delta}\right),$$

1120 where K_1 is given in (54), and the statement follows. \square

1121 **Corollary 4.** Under the assumptions of Lemma 14 for any $k \in \{1, \dots, n\}$ and any $p \geq 2$ it holds

$$\mathbb{E}^{2/p}[\|\theta_k^b - \theta^*\|^p] \leq p\alpha_k (en)^{2/p} K_1/2,$$

1122 where K_1 is defined in (54).

1123 *Proof.* Note that from Lemma 14 for $\forall k \in \{1, \dots, n\}$ and for any $t \geq 0$ it holds

$$\mathbb{P}[\|\theta_k^b - \theta^*\|^2 \geq t] \leq f(t),$$

1124 where

$$f(t) = en \exp\left\{-\frac{t}{K_1 \alpha_k}\right\}.$$

1125 Then, we have

$$\begin{aligned}\mathbb{E}[\|\theta_k^b - \theta^*\|^p] &= \int_0^{+\infty} \mathbb{P}[\|\theta_k^b - \theta^*\|^p > u] du \leq \int_0^{+\infty} en \exp\left\{-\frac{u^{2/p}}{K_1 \alpha_k}\right\} du \\ &= en(p/2) \left(K_1 \alpha_k\right)^{p/2} \int_0^{+\infty} e^{-x} x^{p/2-1} dx \leq en \left((p/2) K_1 \alpha_k\right)^{p/2},\end{aligned}$$

1126 where in the last inequality we use that $\Gamma(p/2) \leq (p/2)^{p/2-1}$ (see [3, Theorem 1.5]). \square

1127 **Lemma 15.** Assume A1, A2, A4, A6. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ for any
1128 $k \in \{1, \dots, n\}$ it holds

$$\|\theta_k - \theta^*\|^2 \leq \alpha_k K_2 \log\left(\frac{en}{\delta}\right),$$

1129 where

$$K_2 = \max\left(\frac{8(C_{1,\xi} + 2C_{2,\xi})^2}{\mu}, \frac{k_0^\gamma \|\theta_0 - \theta^*\|^2}{\alpha_0}\right) \quad (56)$$

1130 Moreover, it holds for any $k \in \{1, \dots, n\}$ and any $p \geq 2$ that

$$\mathbb{E}^{2/p}[\|\theta_k - \theta^*\|^p] \leq p \alpha_k (en)^{2/p} K_2 / 2.$$

1131 *Proof.* The proof is similar to the proof of lemma 14 and Corollary 4. \square

1132 E.4 Bounds for D^b

1133 Recall that the term D^b defined in (17), has a form:

$$\begin{aligned}D^b &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (w_i - 1) Q_i \left(G(\theta_{i-1}^b - \theta^*) + g(\theta_{i-1}^b, \xi_i) + H(\theta_{i-1}^b) \right) \\ &\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i \left(H(\theta_{i-1}^b) + g(\theta_{i-1}^b, \xi_i) - H(\theta_{i-1}) - g(\theta_{i-1}, \xi_i) \right).\end{aligned}$$

1134 The following proposition estimates the moments of D^b .

1135 **Proposition 4.** Assume A1- A6. Then it holds for any $p \geq 2$ that

$$\mathbb{E}^{1/p}[\|D^b\|^p] \leq M_{1,1}^b e^{1/p} p^{3/2} n^{1/p-\gamma/2} + M_{2,1}^b e^{2/p} p n^{1/2+2/p-\gamma}, \quad (57)$$

1136 where the constants are given by

$$\begin{aligned}M_{1,1}^b &= 4C_Q \max(L_1, L_2) \frac{\max(\sqrt{K_2}, \sqrt{K_1}) \sqrt{c_0 k_0^{1-\gamma} (W_{\max} + 1)}}{\sqrt{2}(1-\gamma)}, \\ M_{2,1}^b &= 3C_Q L_H \frac{c_0 k_0^{1-\gamma} \max(K_2, K_1) (W_{\max} + 1)}{2(1-\gamma)},\end{aligned} \quad (58)$$

1137 and K_1, K_2 are defined in (54), (56), respectively. Moreover, there is a set $\Omega_0 \in \mathcal{F}_{n-1} =$
1138 $\sigma(\xi_1, \dots, \xi_{n-1})$, such that $\mathbb{P}(\Omega_0) \geq 1 - 1/n$, and on Ω_0 it holds that

$$\{\mathbb{E}[\|D^b\|^p]\}^{1/p} \leq M_{1,1}^b e^{1/p} p^{3/2} n^{2/p-\gamma/2} + M_{2,1}^b e^{2/p} p n^{1/2+3/p-\gamma}. \quad (59)$$

1139 *Proof.* We first show (57). We split

$$D^b = D_1^b + D_2^b,$$

1140 where

$$\begin{aligned}D_1^b &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (w_i - 1) Q_i (G(\theta_{i-1}^b - \theta^*) + g(\theta_{i-1}^b, \xi_i)) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i (g(\theta_{i-1}^b, \xi_i) - g(\theta_{i-1}, \xi_i)), \\ D_2^b &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (w_i - 1) Q_i H(\theta_{i-1}^b) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i (H(\theta_{i-1}^b) - H(\theta_{i-1})).\end{aligned}$$

1141 Applying Minkowski's inequality together with Lemma 16 and Lemma 17 we get (57).

1142 To proof (59) we consider

$$\Omega_0 = \{ \{ \mathbb{E}^b[\|D^b\|^p] \}^{1/p} \leq M_{1,1}^b e^{1/p} p^{3/2} n^{2/p-\gamma/2} + M_{2,1}^b e^{2/p} p n^{1/2+3/p-\gamma} \}.$$

1143 Note that by Markov's inequality

$$\begin{aligned} \mathbb{P}(\Omega_0^c) &\leq \frac{\mathbb{E}[\{ \mathbb{E}^b[\|D^b\|^p] \}]}{n(M_{1,1}^b e^{2/p} p^{3/2} n^{2/p-\gamma/2} + M_{2,1}^b e^{1/p} p n^{1/2+1/p-\gamma})^p} \\ &= \frac{\mathbb{E}[\|D^b\|^p]}{n(M_{1,1}^b e^{1/p} p^{3/2} n^{1/p-\gamma/2} + M_{2,1}^b e^{2/p} p n^{1/2+2/p-\gamma})^p} \leq \frac{1}{n}. \end{aligned}$$

1144

□

1145 **Lemma 16.** Assume A1-A6. Then for any $p \geq 2$ it holds

$$\mathbb{E}^{1/p}[\|D_1^b\|^p] \leq M_{1,1}^b e^{1/p} p^{3/2} n^{1/p-\gamma/2},$$

1146 where

$$M_{1,1}^b = 4C_Q \max(L_1, L_2) \frac{\max(\sqrt{K_2}, \sqrt{K_1}) \sqrt{c_0} (W_{\max} + 1)}{\sqrt{2}(1-\gamma)},$$

1147 and K_1, K_2 are defined in (54), (56), respectively.

1148 *Proof.* We split D_1^b into four parts, where each part is a sum of martingale differences. Note that
 1149 $\{Q_i(g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i))\}_{i=1}^n$ is a martingale difference with respect to \mathcal{F}_{i-1} . Then applying
 1150 Burkholder's inequality [30, Theorem 8.6] together with Minkowski's inequality and Lemma 4, we
 1151 obtain that

$$\begin{aligned} \mathbb{E}^{1/p} \left[\left\| \sum_{i=1}^{n-1} Q_i(g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i)) \right\|^p \right] \\ \leq p \left(\mathbb{E}^{2/p} \left[\left(\sum_{i=1}^{n-1} \|Q_i(g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i))\|^2 \right)^{p/2} \right] \right)^{1/2} \\ \leq C_Q p \left(\mathbb{E}^{2/p} \left[\left(\sum_{i=1}^{n-1} \|g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i)\|^2 \right)^{p/2} \right] \right)^{1/2} \\ \leq C_Q p \left(\sum_{i=1}^{n-1} \mathbb{E}^{2/p} [\|g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i)\|^p] \right)^{1/2}. \end{aligned}$$

1152 Finally, using A8 and Lemma 15, we obtain

$$\begin{aligned} \mathbb{E}^{1/p} \left[\left\| \sum_{i=1}^{n-1} Q_i(g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i)) \right\|^p \right] &\leq p C_Q L_2 \left(\sum_{i=1}^{n-1} \mathbb{E}^{2/p} [\|\theta_{i-1} - \theta^*\|^p] \right)^{1/2} \\ &\leq C_Q L_2 (en)^{1/p} p^{3/2} \frac{\sqrt{K_2}}{\sqrt{2}} \left(\sum_{i=0}^{n-2} \alpha_i \right)^{1/2} \\ &\leq C_Q L_2 (en)^{1/p} p^{3/2} \frac{\sqrt{K_2}}{\sqrt{2}} \left(c_0 \frac{(k_0 + n - 2)^{1-\gamma} - (k_0 - 1)^{1-\gamma}}{1 - \gamma} \right)^{1/2}. \end{aligned}$$

Since $k_0 \geq 1$ and $(k_0 + n - 2)^{1-\gamma} - (k_0 - 1)^{1-\gamma} \leq n^{1-\gamma}$ we complete the proof for

$$\mathbb{E}^{1/p} \left[\left\| \sum_{i=1}^{n-1} Q_i(g(\theta_{i-1}, \xi_i) - g(\theta^*, \xi_i)) \right\|^p \right].$$

1153 The proof for other three terms is analogous, since each of the terms

$$\{Q_i(g(\theta_{i-1}^b, \xi_i) - g(\theta^*, \xi_i))\}_{i=1}^{n-1}, \{(w_i - 1)Q_i(g(\theta_{i-1}^b, \xi_i) - g(\theta^*, \xi_i))\}_{i=1}^{n-1}, \{(w_i - 1)Q_i G(\theta_{i-1}^b - \theta^*)\}_{i=1}^{n-1},$$

1154 are martingale differences with respect to $\tilde{\mathcal{F}}_{i-1}$ (see definition in (18)). We finish the proof applying
 1155 Minkowski's inequality. □

1156 **Lemma 17.** Assume A1- A6. Then for any $p \geq 2$ it holds

$$\begin{aligned}\mathbb{E}^{1/p}[\|D_2^b\|^p] &\leq M_{2,1}^b e^{2/p} p n^{1/2+2/p-\gamma}, \\ M_{2,1}^b &= 3C_Q L_H \frac{c_0 \max(K_2, K_1)(W_{\max} + 1)}{2(1-\gamma)},\end{aligned}$$

1157 and K_1, K_2 are defined in (54), (56), respectively.

1158 *Proof.* Using Minkowski's inequality, we get

$$\begin{aligned}\mathbb{E}^{1/p}[\|D_2^b\|^p] &\leq \frac{1}{\sqrt{n}} \mathbb{E}^{1/p}[\|\sum_{i=1}^{n-1} Q_i H(\theta_{i-1})\|^p] \\ &\quad + \frac{1}{\sqrt{n}} \mathbb{E}^{1/p}[\|\sum_{i=1}^{n-1} (w_i - 1) Q_i \left(H(\theta_{i-1}^b)\right)\|^p] \\ &\quad + \frac{1}{\sqrt{n}} \mathbb{E}^{1/p}[\|\sum_{i=1}^{n-1} Q_i H(\theta_{i-1}^b)\|^p].\end{aligned}\tag{60}$$

1159 We will now consider each term separately. Using Minkowski's inequality together with Lemma 5,
1160 we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}} \mathbb{E}^{1/p}[\|\sum_{i=1}^{n-1} Q_i H(\theta_{i-1})\|^p] &\leq \frac{C_Q L_H}{\sqrt{n}} \sum_{i=0}^{n-2} \mathbb{E}^{1/p}[\|\theta_i - \theta^*\|^{2p}] \\ &\leq \frac{C_Q L_H p}{\sqrt{n}} (en)^{2/p} (K_2/2) \sum_{i=0}^{n-1} \alpha_i \\ &\leq \frac{C_Q L_H p}{\sqrt{n}} (en)^{2/p} (K_2/2) \left(c_0 \frac{(k_0 + n - 2)^{1-\gamma} - (k_0 - 1)^{1-\gamma}}{1 - \gamma} \right).\end{aligned}$$

1161 Since $k_0 \geq 1$ and $(k_0 + n - 2)^{1-\gamma} - (k_0 - 1)^{1-\gamma} \leq n^{1-\gamma}$ we complete the proof for the first term
1162 in the r.h.s. of (60). The proof for other two terms is analogous. \square

1163 E.5 Matrix Bernstein inequality for Σ_n^b and Gaussian comparison

1164 **Lemma 18.** Under assumptions A1, A2, A6, A7, there is a set $\Omega_1 \in \mathcal{F}_{n-1}$, such that $\mathbb{P}(\Omega_1) \geq 1 - 1/n$
1165 and on Ω_1 it holds that

$$\|\Sigma_n^b - \Sigma_n\| \leq \frac{10C_{Q,\xi} \sqrt{\log(2dn)}}{3\sqrt{n}}$$

1166 where the constant $C_{Q,\xi}$ is given by

$$C_{Q,\xi} := C_Q^2 (C_{1,\xi}^2 + \lambda_{\max}(\Sigma_\xi)),\tag{61}$$

1167 and $C_{1,\xi}, C_Q$ are defined in A2 and Lemma 4, respectively.

1168 *Proof.* Note that

$$\Sigma_n^b - \Sigma_n = \frac{1}{n} \sum_{i=1}^{n-1} Q_i (\eta(\xi_i) \eta(\xi_i)^\top - \Sigma_\xi) Q_i^\top.$$

1169 For simplicity we denote $A_i = Q_i (\eta(\xi_i) \eta(\xi_i)^\top - \Sigma_\xi) Q_i^\top$. Note that for any $i \in \{1, \dots, n-1\}$ it
1170 holds that

$$\mathbb{E}[A_i] = 0, \quad \|A_i\| \leq C_{Q,\xi}, \quad \left\| \sum_{i=1}^{n-1} \mathbb{E}[A_i A_i^\top] \right\| \leq n C_{Q,\xi}^2.$$

1171 Then, using matrix Bernstein inequality [45, Chapter 6], we obtain

$$\mathbb{P}\left(\frac{1}{n} \left\| \sum_{i=1}^{n-1} A_i \right\| \geq t\right) \leq 2d \exp\left\{ \frac{-t^2 n^2 / 2}{n C_{Q,\xi}^2 + n C_{Q,\xi} t / 3} \right\}.$$

1172 Taking $t_\delta = \frac{4C_{Q,\xi} \log(2d/\delta)}{3n} + \frac{2C_{Q,\xi} \sqrt{\log(2d/\delta)}}{\sqrt{n}}$, we obtain that with probability at least $1 - \delta$, it holds

$$\frac{1}{n} \left\| \sum_{i=1}^{n-1} A_i \right\| \leq t_\delta .$$

1173 Setting $\delta = 1/n$ and applying A7 completes the proof. \square

1174 **Corollary 5.** Under assumptions A1, A2, A6, A7, on Ω_1 it holds that

$$\lambda_{\min}(\Sigma_n^b) \geq \frac{1}{2C_\Sigma^2} .$$

1175 *Proof.* Using eigenvalue stability (Lidski's) inequality, we obtain

$$\lambda_{\min}(\Sigma_n^b) \geq \lambda_{\min}(\Sigma_n) - \|\Sigma_n - \Sigma_n^b\| .$$

1176 Note that on Ω_1 , we have

$$\|\Sigma_n - \Sigma_n^b\| \leq \frac{10C_{Q,\xi} \sqrt{\log(2dn)}}{3\sqrt{n}} \leq \frac{1}{2C_\Sigma^2} ,$$

1177 where in the last inequality we use A7. \square

Lemma 19. Assume Under assumptions A1, A2, A6, A7. Then on Ω_1 , it holds that

$$d_C(\{\Sigma_n^b\}^{-1/2} \eta^b, \Sigma_n^{1/2} \eta) \leq \frac{5C_{Q,\xi} C_\Sigma^2 \sqrt{d \log(2dn)}}{\sqrt{n}} .$$

1178 *Proof.* By Lemma 4, $\|\Sigma_n^{-1/2}\| \leq C_\Sigma$. Hence, due to Lemma 18, we have

$$\text{Tr}\{(\Sigma_m^{-1/2} \Sigma_n^b \Sigma_n^{-1/2} - I_p)^2\} \leq d \|(\Sigma_n^{-1/2} \Sigma_n^b \Sigma_n^{-1/2} - I_p)^2\|^2 \leq d C_\Sigma^2 \|\Sigma_n^b - \Sigma_n\|^2 \leq \delta^2 .$$

where we have set

$$\delta = \frac{10C_{Q,\xi} C_\Sigma^2 \sqrt{d \log(2dn)}}{3\sqrt{n}}$$

1179 We finish the proof applying Lemma 13. \square

1180 E.6 GAR in the bootstrap world

1181 **Theorem 5.** Assume A1 - A7. Then with \mathbb{P} -probability at least $1 - 2/n$, it holds

$$\begin{aligned} & \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\sqrt{n}\{\Sigma_n^b\}^{-1/2}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}^b(Y^b \in B)| \\ & \leq \frac{M_{3,1}^b}{n^{1/2}} + \frac{M_{3,2}^b \log n}{n^{\gamma-1/2}} + \frac{M_{3,3}^b \log^{3/2} n}{n^{\gamma/2}} , \end{aligned}$$

1182 where

$$\begin{aligned} M_{3,1}^b &= 259(\sqrt{2}C_\Sigma C_Q C_{1,\xi})^3 W_{\max} \sqrt{d} , \\ M_{3,2}^b &= 2^{3/2} c_d C_\Sigma M_{2,1}^b e^{3/2+\gamma} , \\ M_{3,3}^b &= 2^{3/2} c_d C_\Sigma M_{1,1}^b e^{3/2+\gamma/2} , \end{aligned} \tag{62}$$

1183 and $M_{1,1}^b, M_{2,1}^b$ are defined in (58).

Proof. Since the matrix Σ_n^b concentrates around Σ_n due to Lemma 18, there is a set Ω_1 such that $\mathbb{P}(\Omega_1) \geq 1 - 1/n$ and $\lambda_{\min}(\Sigma_n^b) > 0$ on Ω_1 . Moreover, on this set Applying Lemma 3 with

$$X = \{\Sigma_n^b\}^{-1/2} W^b, \quad Y = \{\Sigma_n^b\}^{-1/2} D^b,$$

1184 we get

$$\begin{aligned} & \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\sqrt{n}\{\Sigma_n^b\}^{-1/2}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}^b(Y^b \in B)| \\ & \leq \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\{\Sigma_n^b\}^{-1/2}W^b \in B) - \mathbb{P}^b(Y^b \in B)| + 2c_d(\mathbb{E}^b[\|\{\Sigma_n^b\}^{-1/2}D^b\|^p])^{1/(1+p)}. \end{aligned}$$

1185 By [42] (with $D = 0$) we may estimate

$$\begin{aligned} & \sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\{\Sigma_n^b\}^{-1/2}W^b \in B) - \mathbb{P}^b(Y^b \in B)| \\ & \leq \frac{259d^{1/2}}{n^{3/2}} \sum_{i=1}^n \mathbb{E}^b[|w_i - 1|^3] \|(\{\Sigma_n^b\}^{-1/2}Q_i\eta(\xi_i))\|^3. \end{aligned} \quad (63)$$

1186 Applying Lemma 4 and Corollary 5 we get

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\{\Sigma_n^b\}^{-1/2}W^b \in B) - \mathbb{P}^b(Y^b \in B)| \leq \frac{259d^{1/2}(\sqrt{2}C_\Sigma C_Q C_{1,\xi})^3 W_{\max}}{n^{1/2}}.$$

1187 From Proposition 4 and Corollary 5 it follows that on the set $\Omega_0 \cap \Omega_1$ the following bound is
1188 satisfied

$$(\mathbb{E}^b[\|\{\Sigma_n^b\}^{-1/2}D^b\|^p])^{1/(p+1)} \leq \sqrt{2}C_\Sigma(M_{1,1}^b e^{1/p} p^{3/2} n^{2/p-\gamma/2} + M_{2,1}^b e^{2/p} p n^{1/2+3/p-\gamma})^{p/(p+1)}.$$

1189 Since $p \geq 2$, $M_{1,1}^b, M_{2,1}^b \geq 1$, we obtain

$$(\mathbb{E}^b[\|\{\Sigma_n^b\}^{-1/2}D^b\|^p])^{1/(p+1)} \leq \sqrt{2}C_\Sigma(e^{1/2}M_{1,1}^b p^{3/2} n^{\frac{2}{p+1}} n^{-\gamma/2} n^{\frac{\gamma/2}{(p+1)}} + eM_{2,1}^b p n^{\frac{3}{p+1}} n^{1/2-\gamma} n^{-\frac{1/2-\gamma}{p+1}}).$$

1190 Setting $p = \log n - 1$, we get

$$(\mathbb{E}^b[\|\{\Sigma_n^b\}^{-1/2}D^b\|^p])^{1/(p+1)} \leq \sqrt{2}C_\Sigma(M_{1,1}^b (\log n)^{3/2} e^{3/2+\gamma/2} n^{-\gamma/2} + M_{2,1}^b (\log n) e^{3/2+\gamma} n^{1/2-\gamma}).$$

1191 By combining the above inequalities, we complete the proof. \square

1192 **Remark 3.** We use [42] with $D = 0$ to prove (63) since we are not aware of Berry-Esseen results for
1193 non i.i.d. random vectors in dimension d with precise constants and dependence on d . The result [6]
1194 may be applied for i.i.d. vectors only.

1195 E.7 Proof of Theorem 1

1196 Collecting bounds of Theorem 2, Theorem 5, we get that with \mathbb{P} -probability at least $1 - 2/n$, it
1197 holds:

$$\sup_{B \in \mathcal{C}(\mathbb{R}^d)} |\mathbb{P}^b(\sqrt{n}(\bar{\theta}_n^b - \bar{\theta}_n) \in B) - \mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^*) \in B)| \leq \frac{C_1 \sqrt{\log n}}{n^{1/2}} + \frac{C_2 \log n}{n^{\gamma-1/2}} + \frac{C_3 \log^{3/2} n}{n^{\gamma/2}},$$

1198 where

$$C_1 = C_4 + M_{3,1}^b + 5C_{Q,\xi} C_\Sigma^2 \sqrt{d \log(2d)}, \quad C_2 = C_5 + M_{3,2}^b, \quad C_3 = C_6 + M_{3,3}^b. \quad (64)$$

1199 F Lower bounds

1200 In the following computations we provide a lower bound on the quantity $|\frac{1}{n} \sum_{j=1}^{n-1} Q_j^2 - 1|$, provided
1201 that the number of observations n is large enough. For simplicity in this bound we consider $k_0 = 1$.
1202 We first note that

$$\frac{1}{n} \sum_{j=1}^{n-1} Q_j^2 - 1 = \frac{1}{n} \sum_{j=1}^{n-1} (Q_j - 1)(Q_j + 1) - \frac{1}{n} = \frac{T_1}{n} + \frac{T_2}{n},$$

1203 where

$$T_1 = \sum_{j=1}^{n-1} (Q_j - 1)^2, \quad T_2 = -2 \sum_{j=1}^{n-1} (Q_j - 1) - 1,$$

1204 and treat the terms T_1 and T_2 separately. Using the identity (49), we get, since $G = 1$, that

$$\sum_{j=1}^{n-1} (Q_j - 1) = - \sum_{j=1}^{n-1} \prod_{\ell=1}^j (1 - \alpha_\ell).$$

1205 Hence, with Lemma 3,

$$\left| \sum_{i=1}^{n-1} (Q_i - 1) \right| \leq \frac{C_Q}{c_0}.$$

1206 Hence, we can conclude that

$$|T_2| \leq \left(\frac{2C_Q}{c_0} + 1 \right),$$

1207 and proceed with T_1 . Here we notice that, applying (48),

$$Q_i - 1 = S_i - \prod_{\ell=i}^{n-1} (1 - \alpha_\ell), \quad S_i = \sum_{j=i+1}^{n-1} (\alpha_i - \alpha_j) \prod_{\ell=i+1}^{j-1} (1 - \alpha_\ell).$$

1208 Thus, the term T_1 can be represented as

$$T_1 = \sum_{j=1}^{n-1} S_j^2 - 2 \sum_{j=1}^{n-1} S_j \prod_{\ell=j}^{n-1} (1 - \alpha_\ell) + \sum_{j=1}^{n-1} \prod_{\ell=j}^{n-1} (1 - \alpha_\ell)^2. \quad (65)$$

1209 Due to item (a) from Lemma 12, it holds that $|S_j| \leq C_S / (j+1)^{1-\gamma}$. Hence, similarly to the proof
1210 of Lemma 1 we can show that

$$\frac{1}{n} \left| \sum_{j=1}^{n-1} S_j^2 \right| \leq C_S^2 n^{2(\gamma-1)} / (2\gamma - 1),$$

1211 and

$$\frac{1}{n} \left| \sum_{j=1}^{n-1} S_j \prod_{\ell=j}^{n-1} (1 - \alpha_\ell) \right| \leq C_S n^{2(\gamma-1)} / c_0.$$

1212 Now, we proceed with the last term in (65), and provide a lower bound on the last remaining
1213 component of T_1 in (65), that is,

$$T_3 = \sum_{j=1}^{n-1} \prod_{\ell=j}^{n-1} (1 - \alpha_\ell)^2.$$

1214 Since $\alpha_j = \frac{c_0}{(1+j)^\gamma}$, we get, using an elementary inequality $1 - x \geq \exp\{-2x\}$, valid for $0 \leq x \leq$
1215 $1/2$, we get that

$$\begin{aligned} \sum_{j=1}^{n-1} \prod_{\ell=j}^{n-1} (1 - \alpha_\ell)^2 &\geq \sum_{j=1}^{n-1} \exp \left\{ - \sum_{\ell=j}^{n-1} \frac{4c_0}{(1+\ell)^\gamma} \right\} \\ &\geq \sum_{j=1}^{n-1} \exp \left\{ - \frac{4c_0}{1-\gamma} (n^{1-\gamma} - j^{1-\gamma}) \right\} \\ &= \exp \left\{ - \frac{4c_0}{1-\gamma} n^{1-\gamma} \right\} \sum_{j=1}^{n-1} \exp \left\{ \frac{4c_0}{1-\gamma} j^{1-\gamma} \right\} \end{aligned}$$

1216 Now we get that

$$\begin{aligned} \sum_{j=1}^{n-1} \exp\left\{\frac{4c_0}{1-\gamma} j^{1-\gamma}\right\} &\geq \int_0^{n-1} \exp\left\{\frac{4c_0}{1-\gamma} y^{1-\gamma}\right\} dy \\ &= (n-1) \int_0^1 \exp\left\{\frac{4c_0}{1-\gamma} ((n-1)z)^{1-\gamma}\right\} dz . \end{aligned}$$

1217 Now we proceed with Laplace approximation (see e.g. [17] or [29]) for the inner integral:

$$\int_0^1 \exp\left\{\frac{4c_0}{1-\gamma} ((n-1)z)^{1-\gamma}\right\} dz = \exp\left\{\frac{4c_0}{1-\gamma} (n-1)^{1-\gamma}\right\} \frac{(n-1)^{\gamma-1}}{4c_0} [1 + \mathcal{O}(n^{\gamma-1})]$$

1218 Since $n^{1-\gamma} - (n-1)^{1-\gamma} \leq 1$ and $\frac{n-1}{n} \geq 1/2$ for $n \geq 2$, we get

$$\frac{1}{n} \sum_{j=1}^{n-1} \prod_{\ell=j}^{n-1} (1 - \alpha_\ell)^2 \geq \frac{1}{4c_0} \exp\left\{-\frac{8c_0}{1-\gamma}\right\} \frac{1}{(n-1)^{1-\gamma}} + \mathcal{O}(n^{2(\gamma-1)}) .$$

1219 Hence, we conclude that for n large enough,

$$|\sigma_{n,\gamma}^2 - 1| > \frac{C_1(\gamma, c_0)}{n^{1-\gamma}} ,$$

1220 and the statement follows. To prove the second part, it remains to apply the lower bound on the total
1221 variation distance between Gaussian random vectors given in [13, Theorem 1.1].

1222 F.1 Numerical demonstration

1223 In order to illustrate numerically the tightness of bounds provided in Proposition 1, we consider the
1224 following simple experiment. We consider the statistics

$$|\sigma_{n,\gamma}^2 - 1| \cdot n^{1-\gamma} , \quad n \in \{2^{10}, \dots, 2^{27}\} .$$

1225 We illustrate numerically the tightness of our bound in the Figure 1 below by calculating

$$n^{1-\gamma} \cdot |\sigma_{n,\gamma}^2 - 1|$$

1226 for different values of $\gamma \in \{0.5, \dots, 0.9\}$ and n . Here we fix the values of parameter $k_0 = 1$
1227 and $c_0 = 1$. Code to reproduce the plot is provided in [https://anonymous.4open.science/r/](https://anonymous.4open.science/r/gaussian_approximation_sgd-5C8F)
1228 [gaussian_approximation_sgd-5C8F](https://anonymous.4open.science/r/gaussian_approximation_sgd-5C8F).

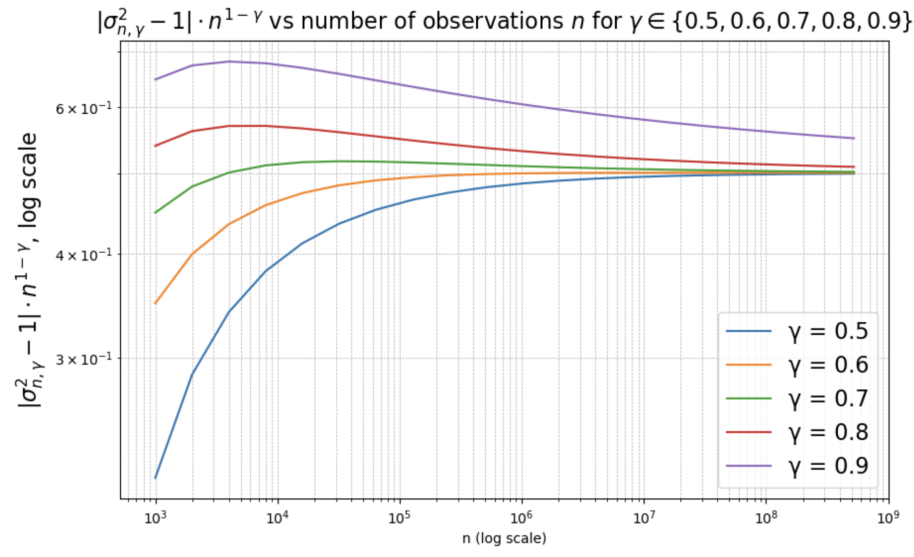


Figure 1: Numerical verification of the lower bound given in Proposition 1