NONASYMPTOTIC ANALYSIS OF STOCHASTIC GRADIENT DESCENT WITH THE RICHARDSON–ROMBERG EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the problem of solving strongly convex and smooth minimization problems using stochastic gradient descent (SGD) algorithm with a constant step size. Previous works suggested to combine the Polyak-Ruppert averaging procedure with the Richardson-Romberg extrapolation technique to reduce the asymptotic bias of SGD at the expense of a mild increase of the variance. We significantly extend previous results by providing an expansion of the mean-squared error of the resulting estimator with respect to the number of iterations n. More precisely, we show that the mean-squared error can be decomposed into the sum of two terms: a leading one of order $\mathcal{O}(n^{-1/2})$ with explicit dependence on a minimax-optimal asymptotic covariance matrix, and a second-order term of order $\mathcal{O}(n^{-3/4})$ where the power $3/4$ is best known. We also extend this result to the p-th moment bound keeping optimal scaling of the remainders with respect to n. Our analysis relies on the properties of the SGD iterates viewed as a time-homogeneous Markov chain. In particular, we establish that this chain is geometrically ergodic with respect to a suitably defined weighted Wasserstein semimetric.

1 INTRODUCTION

Stochastic gradient methods are a fundamental approach for solving a wide range of optimization problems, with a broad range of applications including generative modeling (Goodfellow et al., 2014; 2016), empirical risk minimization (Van der Vaart, 2000), and reinforcement learning (Sutton & Barto, 2018; Schulman et al., 2015; Mnih et al., 2015). These methods are devoted to solving the stochastic minimization problem

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

where ξ is a random variable on $(\mathsf{Z},\mathcal{Z})$ and we can access the gradient ∇f of the function f only through (unbiased) noisy estimates ∇F. Throughout this paper, we consider strongly convex minimization problems admitting a unique solution $\theta^\star$. Arguably the simplest and one of the most widely used approaches to solve (1) is the stochastic gradient descent (SGD), defining the sequence of updates

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

where $\{\gamma_k\}_{k\in\mathbb{N}}$ are step sizes, either diminishing or constant, and $\{\xi_k\}_{k\in\mathbb{N}}$ is an i.i.d. sequence with distribution $\mathbb{P}_\xi$. The algorithm (2) can be viewed as a special instance of the Robbins-Monro procedure (Robbins & Monro, 1951). While the SGD algorithm remains one of the core algorithms in statistical inference, its performance can be enhanced by means of additional techniques that use e.g., momentum (Qian, 1999), averaging (Polyak & Juditsky, 1992), or variance reduction (Defazio et al., 2014; Nguyen et al., 2017). In particular, the celebrated Polyak-Ruppert algorithm proceeds with a trajectory-wise averaging of the estimates

$$\bar\theta_{n_0,n} = \tfrac{1}{n}\sum_{k=n_0+1}^{n+n_0}\theta_k \tag{3}$$

for some $n_0>0$. It is known (Polyak & Juditsky, 1992; Fort, 2015), that under appropriate assumptions on f and γ_k, the sequence of estimates $\{\bar\theta_{n_0,n}\}_{n\in\mathbb{N}}$ is asymptotically normal, that is,

$$\sqrt{n}(\bar\theta_{n_0,n}-\theta^\star) \xrightarrow{d} \mathrm{N}(0,\Sigma_\infty)\,, \quad n\to\infty \tag{4}$$

where $\xrightarrow{d}$ denotes the convergence in distribution and $\mathrm{N}(0, \Sigma_\infty)$ denotes the zero-mean Gaussian distribution with covariance matrix Σ_∞, which is asymptotically optimal from the Rao-Cramer lower bound, see Fort (2015) for a discussion. On the other hand, quantitative counterparts of (4) rely on the mean-squared error bounds of the form

$$\mathsf{E}^{1/2}[\|\bar{\theta}_{n_0,n} - \theta^\star\|^2] \leq \frac{\sqrt{\operatorname{Tr}\Sigma_\infty}}{n^{1/2}} + \frac{C(f,d)}{n^{1/2+\delta}} + \mathcal{R}(\|\theta_0 - \theta^\star\|, n)\,. \tag{5}$$

Here $\mathcal{R}(\|\theta_0 - \theta^\star\|, n)$ is a remainder term which reflects the dependence upon initial condition, $C(f,d)$ is some instance-dependent constant and $\delta > 0$. There are many studies establishing (5) for Polyak-Ruppert averaged SGD under various model assumptions, including Bach & Moulines (2013), Gadat & Panloup (2023). In particular, Li et al. (2022) derived the bound (5) with the rate $\delta = 1/4$ and proved that the second order term is unavoidable under a natural setup. However, their results apply a modified algorithm with control variates and multiple restarts. In our work, we show that the same non-asymptotic upper bound is achieved by a simple modification of the estimate $\bar{\theta}_{n_0,n}$ based on Richardson-Romberg extrapolation. The main contributions of the current paper are as follows:

- We show that a version of SGD algorithm with constant step size, Polyak-Ruppert averaging, and Richardson-Romberg extrapolation lead to the root-MSE bound (5) with $\delta = 1/4$ when applied to strongly convex minimization problems. We obtain this result by leveraging the analysis of iterates generated by the constant step-size SGD as a Markov chain. This process turns out to be geometrically ergodic with respect to a carefully designed Wasserstein semi-metric (see detailed discussion in Section 3). It is important to note that this result is obtained for a fixed step size γ of order $1/\sqrt{n}$ with n being a total number of iterations. This result requires that the number of samples, n, is known a priori to optimize the step size γ.
- We generalize the above result to the p-th moment error bounds. Under a similar step size $\gamma \asymp 1/\sqrt{n}$, we obtain the error bound of the order

$$\mathsf{E}^{1/p}[\|\bar{\theta}_n^{(RR)} - \theta^\star\|^p] \leq \frac{Cp^{1/2}\sqrt{\operatorname{Tr}\Sigma_\infty}}{n^{1/2}} + \frac{C(f,d,p)}{n^{3/4}} + \mathcal{R}(\|\theta_0 - \theta^\star\|, n, p)\,, \tag{6}$$

  where C is a universal constant, and $\bar{\theta}_n^{(RR)}$ is a counterpart of the quantity $\bar{\theta}_{n_0,n}$ when using Richardson-Romberg extrapolation, see related definitions at Section 4. Our proof is based on a novel version of the Rosenthal inequality, which might be of independent interest.

The rest of the paper is organized as follows. First, we provide a literature review on the non-asymptotic analysis of the first order optimization methods, with an emphasis on the constant step-size algorithms and Richardson-Romberg procedure in Section 2. Then, we provide analysis of the constant step size SGD viewed as a Markov chain together with the properties of the Polyak-Ruppert averaged estimator (3) in Section 3. In Section 4, we discuss the Richardson-Romberg extrapolation applied to the Polyak-Ruppert averaged SGD and derive the respective 2-nd and p-th moment error bounds.

Notations and definitions. For $\theta_1, \ldots, \theta_k$ being the iterates of stochastic first-order method, we denote $\mathcal{F}_k = \sigma(\theta_0, \theta_1, \ldots, \theta_k)$ and E_k be an alias for $\mathsf{E}\left[\cdot|\mathcal{F}_k\right]$. We call a function $c: \mathsf{Z} \times \mathsf{Z} \to \mathbb{R}_+$ a *distance-like* function, if it is symmetric, lower semi-continuous and $c(x,y) = 0$ if and only if $x = y$, and there exists $q \in \mathbb{N}$ such that $(d(x,y) \wedge 1)^q \leq c(x,y)$. For two probability measures ξ and ξ' we denote by $\mathscr{C}(\xi, \xi')$ the set of couplings of two probability measures, that is, for any $\mathcal{C} \in \mathscr{C}(\xi,\xi')$ and any $A \in \mathcal{Z}$ it holds $\mathcal{C}(\mathsf{Z} \times A) = \xi'(A)$ and $\mathcal{C}(A \times \mathsf{Z}) = \xi(A)$. We define the Wasserstein semi-metrics associated to the distance-like function $c(\cdot,\cdot)$, as

$$\mathbf{W}_c(\xi,\xi') = \inf_{\mathcal{C} \in \mathscr{C}(\xi,\xi')} \int_{\mathsf{Z}\times\mathsf{Z}} c(z,z')\mathcal{C}(\mathrm{d}z, \mathrm{d}z')\,. \tag{7}$$

Note that $\mathbf{W}_c(\xi,\xi')$ is not necessarily a distance, as it may fail to satisfy the triangle inequality. In the particular case of $\mathsf{Z} = \mathbb{R}^d$, and $c_p(x,y) = \|x-y\|^p$, $x, y \in \mathbb{R}^d$, $p \geq 1$, we denote the corresponding Wasserstein metrics by $\mathbf{W}_p(\xi,\xi')$. Let $\mathrm{Q}(z, A)$ be a Markov kernel on $(\mathsf{Z}, \mathcal{Z})$. We say that K is a Markov coupling of Q if for all $(z,z') \in \mathsf{Z}^2$ and $A \in \mathcal{Z}$, $\mathrm{K}((z,z'), A \times \mathsf{Z}) = \mathrm{Q}(z,A)$ and $\mathrm{K}((z,z'), \mathsf{Z} \times A) = \mathrm{Q}(z',A)$. If K is a kernel coupling of Q, then for all $n \in \mathbb{N}$, K^n is a kernel coupling of Q^n and for any $\mathcal{C} \in \mathscr{C}(\xi,\xi')$, $\mathcal{C}\mathrm{K}^n$ is a coupling of $(\xi\mathrm{Q}^n, \xi'\mathrm{Q}^n)$ and it holds

$$\mathbf{W}_c(\xi\mathrm{Q}^n, \xi'\mathrm{Q}^n) \leq \int_{\mathsf{Z}\times\mathsf{Z}} \mathrm{K}^n c(z,z')\mathcal{C}(\mathrm{d}z\mathrm{d}z')\,,$$

see (Douc et al., 2018, Corollary 20.1.4). For any probability measure $\mathcal{C}$ on $(\mathsf{Z}^2, \mathcal{Z}^{\otimes 2})$, we denote by $\mathbb{P}^{\mathrm{K}}_{\zeta}$ and $\mathsf{E}^{\mathrm{K}}_{\zeta}$ the probability and the expectation on the canonical space $((\mathsf{Z}^2)^{\mathbb{N}}, (\mathcal{Z}^{\otimes 2})^{\otimes \mathbb{N}})$ such that the canonical process $\{(Z_n, Z'_n), n \in \mathbb{N}\}$ is a Markov chain with initial probability $\mathcal{C}$ and Markov kernel K. We write $\mathsf{E}^{\mathrm{K}}_{z,z'}$ instead $\mathsf{E}^{\mathrm{K}}_{\delta_{z,z'}}$. For all $x, y \in \mathbb{R}^d$ denote by $x \otimes y$ the tensor product of x and y and by $x^{\otimes k}$ the k-th tensor power of x. In addition, for a function $f : \mathbb{R}^d \to \mathbb{R}$ we denote by $\nabla^k f(\theta)$ the k-th differential of f, that is $\nabla^k f(\theta)_{i_1,\cdots,i_k} = \frac{\partial^k f}{\prod_{j=1}^k \partial x_{i_j}}$. For any tensor $M \in (\mathbb{R}^d)^{\otimes(k-1)}$, we define $\nabla^k f(\theta) M \in \mathbb{R}^d$ by the relation $(\nabla^k f(\theta) M)_l = \sum_{i_1,\ldots,i_{k-1}} M_{i_1,\cdots,i_{k-1}} \nabla^k f(\theta)_{i_1,\cdots,i_{k-1},l}$, where $l \in \{1, \ldots, d\}$. For two sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ we write $a_n \lesssim b_n$, if there is an absolute constant c_0, such that $a_n \leq c_0 b_n$. Throughout this paper we use c_0 for an absolute constant, which values may vary from line to line.

2 Literature review

Gadat & Panloup (2023) derive (5) with $\delta = 1/8$ for a certain class of functions f, including strongly convex functions, improving Moulines & Bach (2011) which obtain this result for $\delta = 1/12$. Li et al. (2022) suggested the Root-SGD algorithm combining the ideas of the Polyak-Ruppert averaged SGD with control variates and established (5) with $\delta = 1/4$. The recent series of papers (Huo et al., 2023; Zhang & Xie, 2024; Zhang et al., 2024) investigate stochastic approximation algorithms with both i.i.d. and Markovian data and constant step sizes. The authors consider both linear SA problems and Q-learning, quantify bias, and propose precise characterization of the bias together with a Richardson-Romberg extrapolation procedure. However, these results only consider 2-nd moment of the error and provide MSE bounds of order $\mathcal{O}(1/n) + \mathcal{O}(\gamma)$ with no explicit expression for the leading term.

Richardson-Romberg extrapolation. Richardson-Romberg extrapolation is a technique used to improve the accuracy of numerical approximations (Hildebrand, 1987), such as those from numerical differentiation or integration. It involves using approximations with different step sizes and then extrapolating to reduce the error, typically by removing the leading term in the error expansion. The one-step Richardson-Romberg was introduced to reduce the discretization error induced by an Euler scheme to simulate stochastic differential equation in Talay & Tubaro (1990), and later generalized for non-smooth functions in Bally & Talay (1996). This technique was extended using multistep discretizations in Pagès (2007). Finally, Richardson-Romberg extrapolation have been applied to Stochastic Gradient Descent (SGD) methods in Durmus et al. (2016), Merad & Gaïffas (2023) and Huo et al. (2024b), to improve convergence and reduce error in optimization problems, particularly when dealing with noisy or high-variance gradient estimates.

3 Finite-time analysis of the SGD dynamics for strongly convex minimization problems

3.1 Geometric ergodicity of SGD iterates

We consider the following assumption on the function f in the minimization problem (1).

A1. *The function f is μ-strongly convex on $\mathbb{R}^d$, that is, it is continuously differentiable and there exists a constant $\mu > 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that*

$$\frac{\mu}{2}\|\theta - \theta'\|^2 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \,. \tag{8}$$

A2. *The function f is 4 times continuously differentiable and L_2-smooth on $\mathbb{R}^d$, i.e., it is continuously differentiable and there is a constant $\mathrm{L}_2 \geq 0$, such that for any $\theta, \theta' \in \mathbb{R}^d$,*

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq \mathrm{L}_2 \|\theta - \theta'\| \,. \tag{9}$$

Moreover, f has uniformly bounded 3-rd and 4-th derivatives, there exist $\mathrm{L}_3, \mathrm{L}_4 \geq 0$ such that

$$\|\nabla^i f(\theta)\| \leq \mathrm{L}_i \;\; \textit{for } i \in \{3, 4\} \,. \tag{10}$$

We aim to solve the problem (1) using SGD with a constant step size, starting from initial distribution ν. That is, for $k \geq 0$ and a step size $\gamma \geq 0$, we consider the following recurrent scheme

$$\theta^{(\gamma)}_{k+1} = \theta^{(\gamma)}_k - \gamma \nabla F(\theta^{(\gamma)}_k, \xi_{k+1}) \,, \quad \theta^{(\gamma)}_0 = \theta_0 \sim \nu \,, \tag{11}$$

where $\{\xi_k\}_{k\in\mathbb{N}}$ is a sequence satisfying the following condition.

A3 (p). *$\{\xi_k\}_{k\in\mathbb{N}}$ is a sequence of independent and identically distributed (i.i.d.) random variables with distribution $\mathbb{P}_\xi$, such that ξ_i and θ_0 are independent and for any $\theta \in \mathbb{R}^d$ it holds that*

$$\mathsf{E}_{\xi\sim\mathbb{P}_\xi}[\nabla F(\theta,\xi)] = \nabla f(\theta)\,.$$

Moreover, there exists τ_p, such that $\mathsf{E}^{1/p}[\|\nabla F(\theta^\star,\xi)\|^p] \le \tau_p$, and for any $q = 2,\dots,p$ it holds with some $\mathrm{L}_1 > 0$ that for any $\theta_1,\theta_2 \in \mathbb{R}^d$,

$$\mathrm{L}_1^{q-1}\,\|\theta_1-\theta_2\|^{q-2}\langle\nabla f(\theta_1)-\nabla f(\theta_2),\theta_1-\theta_2\rangle \ge \mathsf{E}_{\xi\sim\mathbb{P}_\xi}[\|\nabla F(\theta_1,\xi)-\nabla F(\theta_2,\xi)\|^q]\,. \tag{12}$$

Assumption **A**3(p) generalizes the well-known L_1-co-coercivity assumption, see Dieuleveut et al. (2020). A sufficient condition which allows for **A**3(p) is to assume that $F(\theta,\xi)$ is $\mathbb{P}_\xi$-a.s. convex with respect to $\theta \in \mathbb{R}^d$. For ease of notation, we set

$$\mathrm{L} = \max(\mathrm{L}_1,\mathrm{L}_2,\mathrm{L}_3,\mathrm{L}_4)\,, \tag{13}$$

and trace only L in our subsequent bounds. In this paper we focus on the convergence to $\theta^\star$ of the Polyak-Ruppert averaging estimator defined for any $n \ge 0$,

$$\bar\theta_n^{(\gamma)} = \tfrac{1}{n}\textstyle\sum_{k=n+1}^{2n}\theta_k^{(\gamma)}\,. \tag{14}$$

Many previous studies instead consider $\bar\vartheta_n^{(\gamma)} = \frac{1}{n-n_0}\sum_{k=n_0+1}^{n}\theta_k^{(\gamma)}$ rather than $\bar\theta_n^{(\gamma)}$, where $n \ge n_0+1$ and n_0 denotes a burn-in period. However, when the sample size n is sufficiently large, the choice of the optimal burn-in size n_0 affects the leading terms in the MSE bound of $\bar\theta_n^{(\gamma)}-\theta^\star$ only by a constant factor. Therefore, we focus on (14), or equivalently, use $2n$ observations and set $n_0 = n$.

Properties of $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ viewed as a Markov chain. Under assumptions **A**1, **A**2 and **A**3(2), the sequence $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ defined by the relation (11) is a time-homogeneous Markov chain with the Markov kernel

$$\mathrm{Q}_\gamma(\theta,\mathsf{A}) = \int_{\mathbb{R}^d}\mathbb{1}_{\mathsf{A}}(\theta-\gamma\nabla F(\theta,z))\mathsf{P}_\xi(\mathrm{d}z)\,,\quad \theta\in\mathbb{R}^d\,,\ \mathsf{A}\in\mathsf{B}(\mathbb{R}^d)\,, \tag{15}$$

where $\mathsf{B}(\mathbb{R}^d)$ denoted the Borel σ-field of $\mathbb{R}^d$. In Dieuleveut et al. (2020) it has been established that, under the stated assumptions, Q_γ admits a unique invariant distribution π_γ, if γ is small enough. Previous studies, such as Dieuleveut et al. (2020) or Merad & Gaïffas (2023), studied the convergence of the distributions of $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ to π_γ in the 2-Wasserstein distance $\mathbf{W}_2$, associated with the Euclidean distance in $\mathbb{R}^d$. However, our main results would require to switch to the non-standard distance-like function, which is defined under **A**1 and **A**3(2) as follows:

$$c(\theta,\theta') = \|\theta-\theta'\|\left(\|\theta-\theta^*\|+\|\theta'-\theta^*\|+\tfrac{2\sqrt{2}\tau_2\sqrt{\gamma}}{\sqrt{\mu}}\right)\,,\qquad \theta,\theta^\star\in\mathbb{R}^d\,. \tag{16}$$

Here the constants τ_2 and μ are given in **A**3(2) and **A**1, respectively. Note that this distance-like function is specifically designed to analyze $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ under **A**1 and **A**3(2). In particular, it depends on the step size γ and $\theta^\star$. Our first main result establishes *geometric ergodicity* of the Markov kernel Q_γ with respect to the distance-like function c from (16).

Proposition 1. *Assume A1, A2, and A3(2). Then for any $\gamma \in (0;1/(2\,\mathrm{L})]$, the Markov kernel Q_γ defined in (15) admits a unique invariant distribution π_γ. Moreover, Q_γ is geometrically ergodic with respect to the cost function c, that is, for any initial distribution ν on $\mathbb{R}^d$ and $k\in\mathbb{N}$,*

$$\mathbf{W}_c(\nu\mathrm{Q}_\gamma^k,\pi_\gamma) \le 4(1/2)^{k/m(\gamma)}\mathbf{W}_c(\nu,\pi_\gamma)\,, \tag{17}$$

where $m(\gamma) = \lceil 2\log 4/(\gamma\mu)\rceil$.

Discussion. The proof of Proposition 1 is provided in Appendix A.1. Properties of the invariant distribution π_γ were previously studied in literature, see e.g. Dieuleveut et al. (2020). It particular, it is known (Dieuleveut et al., 2020, Lemma 13), that the 2-nd moment of $\theta_\infty^{(\gamma)}$, where $\theta_\infty^{(\gamma)}$ is distributed according to the stationary distribution π_γ, scales linearly with γ:

$$\textstyle\int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\pi_\gamma(\mathrm{d}\theta) \lesssim \frac{\gamma\tau_2}{\mu}\,. \tag{18}$$

This property yields, using Lyapunov's inequality, that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \theta'\| \pi_\gamma(\mathrm{d}\theta)\pi_\gamma(\mathrm{d}\theta') \lesssim \sqrt{\tfrac{\gamma \tau_2}{\mu}}\,.$$

At the same time, expectation of the cost function $c(\theta, \theta')$ scale linearly with the step size γ:

$$\int_{\mathbb{R}^d} c(\theta, \theta')\pi(\mathrm{d}\theta)\pi(\mathrm{d}\theta') \lesssim \tfrac{\gamma \tau_2}{\mu}\,. \tag{19}$$

The property (19) is crucial to obtain tighter (with respect to the step size γ) error bounds for the Richardson-Romberg estimator, as well as in the Rosenthal inequality for additive functional of $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ derived in Proposition 8. Precisely, the additional $\sqrt{\gamma}$ factor obtained in (19) would allow us to obtain sharper bounds on the remainder terms in Theorem 6.

Next, we analyze the error $\theta_\infty^{(\gamma)} - \theta^\star$ where $\theta_\infty^{(\gamma)}$ is distributed according to the stationary distribution π_γ. To this end, we consider the following condition.

C1 (p). *There exist constants $\mathsf{D}_{\mathrm{last},p}, \mathsf{C}_{\mathrm{step},p} \geq 2$ depending only on p, such that for any step size $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},p})]$, and any initial distribution ν it holds that*

$$\mathsf{E}_\nu^{2/p}\big[\|\theta_k^{(\gamma)} - \theta^\star\|^p\big] \leq (1-\gamma\mu)^k \mathsf{E}_\nu^{2/p}\big[\|\theta_0 - \theta^\star\|^p\big] + \mathsf{D}_{\mathrm{last},p}\gamma\tau_p^2/\mu\,. \tag{20}$$

Moreover, for the stationary distribution π_γ it holds that

$$\mathsf{E}_{\pi_\gamma}^{2/p}\big[\|\theta_\infty^{(\gamma)} - \theta^\star\|^p\big] \leq \mathsf{D}_{\mathrm{last},p}\gamma\tau_p^2/\mu\,. \tag{21}$$

Note that, since $\mathrm{L} \geq \mu$, it holds that $\gamma\mu \leq 1/2$. It is important to recognize that **C**1 is not independent from the previous assumptions **A**1 - **A**3(p). In particular, Dieuleveut et al. (2020, Lemma 13) implies that, under **A**1, **A**3(p) with $p \geq 2$, and **A**2, the bound (21) holds for $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},p})]$ with some constants $\mathsf{D}_{\mathrm{last},p}$ and $\mathsf{C}_{\mathrm{step},p}$, which depends only upon p. Unfortunately, it is complicated to obtain precise dependence of $\mathsf{C}_{\mathrm{step},p}$ and $\mathsf{D}_{\mathrm{last},p}$ upon p, as well as to obtain the bound (21) with tight numerical constants. The results available in the literature (Gadat & Panloup, 2023; Li et al., 2022; Merad & Gaïffas, 2023) either are obtained under alternative set of assumptions, or are not explicit with respect to their dependence upon p. That is why we prefer to state **C**1(p) as a separate assumption. In the subsequent bounds we use **C**1(p) together with **A**1, **A**3(p) with $p \geq 2$, and **A**2, tracking the dependence of our bounds upon $\mathsf{C}_{\mathrm{step},p}$ and $\mathsf{D}_{\mathrm{last},p}$. We leave the problem of deriving **C**1(p) with sharp constants $\mathsf{D}_{\mathrm{last},p}, \mathsf{C}_{\mathrm{step},p}$ as an interesting direction for the future research.

Under the assumption **C**1, we control the fluctuations of $\{\theta_k^{(\gamma)}\}$ around the solution $\theta^\star$ of (1). However, unless the function f is quadratic, it is known that $\int_{\mathbb{R}^d} \theta\pi_\gamma(\mathrm{d}\theta) \neq \theta^\star$. In the following proposition, we quantify this bias under milder assumptions compared to the ones from Dieuleveut et al. (2020, Theorem 4). Namely, the following result holds:

Proposition 2. *Assume **A**1, **A**2, **A**3(6), and **C**1(6). Then there exist such $\Delta_1 \in \mathbb{R}^d, \Delta_2 \in \mathbb{R}^{d\times d}$, not depending upon γ, that for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, it holds*

$$\bar{\theta}_\gamma := \int_{\mathbb{R}^d} \theta\pi_\gamma(\mathrm{d}\theta) = \theta^\star + \gamma\Delta_1 + B_1\gamma^{3/2}\,, \tag{22}$$

$$\bar{\Sigma}_\gamma := \int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma\Delta_2 + B_2\gamma^{3/2}\,. \tag{23}$$

Here $B_1 \in \mathbb{R}^d$ and $B_2 \in \mathbb{R}^{d\times d}$ satisfy $\|B_i\| \leq \mathsf{C}_1$, $i = 1, 2$, where C_1 defined in (51) is a constant independent of γ. Moreover, for any initial distribution ν on $\mathbb{R}^d$, it holds that

$$\mathsf{E}_\nu[\bar{\theta}_n^{(\gamma)}] = \theta^\star + \gamma\Delta_1 + B_1\gamma^{3/2} + \mathcal{R}_1(\theta_0 - \theta^\star, \gamma, n)\,, \tag{24}$$

where

$$\|\mathcal{R}_1(\theta_0 - \theta^\star, \gamma, n)\| \lesssim \tfrac{\mathrm{e}^{-\gamma\mu(n+1)/2}}{n\gamma\mu}\left(\mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^2\big] + \tfrac{\sqrt{\gamma}\tau_2}{\sqrt{\mu}}\right)\,. \tag{25}$$

The proof is postponed to Appendix A. Results of this type were already obtained in the literature for stochastic approximation algorithms, see e.g. Huo et al. (2024a) and Allmeier & Gast (2024). As already highlighted, the additive term Δ_1 vanishes in the case of minimizing the quadratic function f, see Bach & Moulines (2013).

3.2 Analysis of the Polyak-Ruppert averaged estimator $\bar{\theta}_n^{(\gamma)}$.

In this section, we analyze the finite-sample properties of the estimator $\bar{\theta}_n^{(\gamma)}$ from (14). The analysis is based on techniques previously used in Moulines & Bach (2011), as well as in the analysis of the Polyak-Ruppert averaged LSA (Linear Stochastic Approximation) algorithms, see Mou et al. (2020); Durmus et al. (2024). Next, we define the k-th step noise level at the point $\theta \in \mathbb{R}^d$ by:

$$\varepsilon_k(\theta) = \nabla F(\theta, \xi_k) - \nabla f(\theta)\,, \tag{26}$$

Note that $\varepsilon_{k+1}(\theta_k^{(\gamma)})$ is a martingale-difference sequence w.r.t. the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. With these notations, (11) writes

$$\theta_{k+1}^{(\gamma)} - \theta^\star = \theta_k^{(\gamma)} - \theta^\star - \gamma\big(\nabla f(\theta_k^{(\gamma)}) + \varepsilon_{k+1}(\theta_k^{(\gamma)})\big)\,. \tag{27}$$

Setting

$$\eta(\theta) = \nabla f(\theta) - \mathrm{H}^\star(\theta - \theta^\star)\,, \tag{28}$$

with

$$\mathrm{H}^\star = \nabla^2 f(\theta^\star) \in \mathbb{R}^{d\times d}\,. \tag{29}$$

we get from (27) with additional rearranging the terms, that

$$\mathrm{H}^\star(\theta_k^{(\gamma)} - \theta^\star) = \frac{\theta_k^{(\gamma)} - \theta_{k+1}^{(\gamma)}}{\gamma} - \varepsilon_{k+1}(\theta_k^{(\gamma)}) - \eta(\theta_k^{(\gamma)})\,. \tag{30}$$

Taking average of (30) with $k = n+1$ to $2n$, we arrive at the final representation:

$$\mathrm{H}^\star(\bar{\theta}_n^{(\gamma)} - \theta^\star) = \frac{\theta_{n+1}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{\theta_{2n}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{1}{n}\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta_k^{\gamma}) - \frac{1}{n}\sum_{k=n+1}^{2n} \eta(\theta_k^{\gamma})\,. \tag{31}$$

Representation (31) is the key ingredient in the proof of the next result where the variance of noise $\varepsilon_k(\theta^\star)$ measured at the optimal point $\theta^\star$ naturally appears, that is,

$$\Sigma_\varepsilon^\star = \mathsf{E}_{\xi\sim\mathbb{P}_\xi}[\nabla F(\theta^\star, \xi)^{\otimes 2}]\,. \tag{32}$$

Note that $\Sigma_\varepsilon^\star$ does not depend on the step size γ of (11) and is related to the ”optimal” covariance matrix of the Polyak-Ruppert averaged iterates $\bar{\theta}_n^{(\gamma)}$, see Fort (2015). In our first main result below, we establish non-asymptotic properties of the averaged Polyak-Ruppert estimator (14).

Theorem 3. *Assume **A1**, **A2**, **A3**(6), and **C1**(6). Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, $n \in \mathbb{N}$, and initial distribution ν on $\mathbb{R}^d$, the sequence of Polyak-Ruppert estimates (14) satisfies*

$$\mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(\gamma)} - \theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{\sqrt{n}} + \frac{\mathsf{C}_2}{\gamma^{1/2} n} + \mathsf{C}_3\gamma + \frac{\mathsf{C}_4\gamma^{1/2}}{n^{1/2}} + \mathcal{R}_2(n, \gamma, \|\theta_0 - \theta^\star\|)\,, \tag{33}$$

where the constants C_2 to C_4 are defined in Appendix B (see equation (61)), and

$$\mathcal{R}_2(n, \gamma, \|\theta_0 - \theta^\star\|) = \frac{c_0(1-\gamma\mu)^{(n+1)/2}\,\mathrm{L}}{\gamma\mu n}\mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^2] + \frac{c_0\,\mathrm{L}(1-\gamma\mu)^{n+1}}{2n\gamma\mu}\mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^4]\,,$$

where c_0 is an absolute (numerical) constant.

The version of Theorem 3 with explicit constants together with the proof is provided in Appendix B, see Theorem 13. Note that the result of Theorem 3 is valid for arbitrary $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$. At the same time, this bound can be optimized over step size of the form $\gamma = n^{-\beta}$, $\beta \in (0,1)$.

Corollary 4. *Under the assumptions of Theorem 3, by setting $\gamma = n^{-2/3}$ with $n \ge (\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})^{3/2}$ it holds that*

$$\mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(\gamma)} - \theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\big(n^{-2/3}\big) + \mathcal{R}_2(n, 1/n^{2/3}, \|\theta_0 - \theta^\star\|)\,. \tag{34}$$

Corollary 4 implies that, if n is known in advance and $\gamma \asymp n^{-2/3}$, then $\bar{\theta}_n^{(\gamma)}$ satisfies (5) with $\delta = 1/6$. A closer inspection of the sum (31) reveals that $\mathsf{E}_{\pi_\gamma}[\eta(\theta_k^\gamma)] \asymp \gamma$ meaning that we can not hope to provide a better bound for the term $\frac{1}{n}\sum_{k=n+1}^{2n}\eta(\theta_k^\gamma)$ compared to the one coming from the Minkowski's inequality. Thus, this is the *bias* of the stationary distribution, which prevents us from gaining the optimal second-order term w.r.t. the sample size n from Corollary 4.

Note that in case of deterministic problems $\varepsilon_k(\theta) = 0$ for any k and θ, and **C**1(6) is satisfied for any $p \geq 2$ with $\mathsf{D}_{\mathrm{last},p} = 0$. In such a setting, $\Sigma_\varepsilon^\star = 0$, and the remainder terms are proportional to $\mathsf{D}_{\mathrm{last},p}$ with $p = 2, 4$, or 6, and therefore also vanishes. Therefore, Theorem 3 provides exponential convergence bounds, which are embedded in the remainder term. The decay rate of the second-order (w.r.t. n) term in (34) is well studied in the literature. In particular, Moulines & Bach (2011) obtained a second order term of order $n^{-5/8}$ for the SGD algorithm with Polyak-Ruppert averaging. A similar rate under more general assumptions was reported in (Gadat & Panloup, 2023, Theorem 2) for p-th moment bounds. However, all these results are known to be suboptimal for first-order methods. The recent work by Li et al. (2022) shows that the best known second-order error term in the bounds (34) is of order $\mathcal{O}(n^{-3/4})$ and can be achieved by the Root-SGD algorithm. In the next section we can mirror this bound using the constant step-size SGD algorithm combined with the Richardson-Romberg extrapolation technique (Dieuleveut et al., 2020).

4 Richardson-Romberg extrapolation

Our analysis presented in Theorem 3 was based on the summation by parts formula (31) and Taylor expansion of the gradient $\nabla f(\theta)$ in the vicinity of $\theta^\star$, yielding the remainder quantity $\eta(\theta)$. It is important to notice that

$$\int_{\mathbb{R}^d} \eta(\theta)\pi_\gamma(\mathrm{d}\theta) \neq 0\,, \tag{35}$$

which prevents us from using more aggressive (larger) step sizes γ in the optimized bound (34). In this section we show that Richardson-Romberg extrapolation technique is sufficient to significantly reduce the bias associated with $\eta(\theta)$ and improve the second-order term in the MSE bound (34). Instead of considering a single SGD trajectory $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$, and then relying on the tail-averaged estimator $\bar{\theta}_n^{(\gamma)}$, we construct two parallel chains based on the same sequence $\{\xi_k\}_{k\in\mathbb{N}}$:

$$\begin{aligned}
\theta_{k+1}^{(\gamma)} &= \theta_k^{(\gamma)} - \gamma\nabla F(\theta_k^{(\gamma)}, \xi_{k+1})\,, \quad \bar{\theta}_n^{(\gamma)} = \tfrac{1}{n}\textstyle\sum_{k=n+1}^{2n}\theta_k^{(\gamma)}\,, \\
\theta_{k+1}^{(2\gamma)} &= \theta_k^{(2\gamma)} - 2\gamma\nabla F(\theta_k^{(2\gamma)}, \xi_{k+1})\,, \quad \bar{\theta}_n^{(2\gamma)} = \tfrac{1}{n}\textstyle\sum_{k=n+1}^{2n}\theta_k^{(2\gamma)}\,.
\end{aligned} \tag{36}$$

Based on $\bar{\theta}_n^{(\gamma)}$ and $\bar{\theta}_n^{(2\gamma)}$ defined above, we construct a Richardson-Romberg estimator as

$$\bar{\theta}_n^{(RR)} := 2\bar{\theta}_n^{(\gamma)} - \bar{\theta}_n^{(2\gamma)}\,. \tag{37}$$

Note that it is possible to use different sources of randomness $\{\xi_k\}_{k\in\mathbb{N}}$ and $\{\xi_k'\}_{k\in\mathbb{N}}$ when constructing the sequences $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ and $\{\theta_k^{(2\gamma)}\}_{k\in\mathbb{N}}$, respectively. At the same time, it is possible to show the benefits of using the same sequence of random variables $\{\xi_k\}_{k\in\mathbb{N}}$ in (36). Indeed, consider the decomposition (31) and further expand the term $\eta(\theta)$ defined in (28) as

$$\eta(\theta) = \psi(\theta) + G(\theta)\,,$$

where we have defined, for $\theta \in \mathbb{R}^d$, the following vector-valued functions:

$$\begin{aligned}
\psi(\theta) &= (1/2)\nabla^3 f(\theta^\star)(\theta - \theta^\star)^{\otimes 2}\,, \\
G(\theta) &= (1/6)\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\,\mathrm{d}t\right)(\theta - \theta^\star)^{\otimes 3}\,.
\end{aligned} \tag{38}$$

We further rewrite the decomposition (31) as

$$\begin{aligned}
\mathrm{H}^\star(\bar{\theta}_n^{(\gamma)} - \theta^\star) &= \frac{\theta_{n+1}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{\theta_{2n}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star) \\
&\quad - \frac{1}{n}\sum_{k=n+1}^{2n}\{\varepsilon_{k+1}(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta^\star)\} - \frac{1}{n}\sum_{k=n+1}^{2n}\psi(\theta_k^{(\gamma)}) - \frac{1}{n}\sum_{k=n+1}^{2n}G(\theta_k^{(\gamma)})\,.
\end{aligned} \tag{39}$$

Note that in the decomposition (39), the linear statistics $W = n^{-1}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$ does not depend upon γ. Moreover, when setting the step size $\gamma \simeq n^{-\beta}$ with an appropriate $\beta \in (0,1)$, we can show that the moments of all other terms except for W in the r.h.s. of (39) are small (see Theorem 9 for more details). Hence, using the same sequence $\{\xi_k\}_{k\in\mathbb{N}}$ of noise variables in (36) yields an estimator $\bar\theta_n^{(RR)}$, such that its leading component of the variance still equals W. Hence, using the Richardson-Romberg procedure will increase only the second-order (w.r.t. n) components of the variance. At the same time, using different random sequences $\{\xi_k\}_{k\in\mathbb{N}}$ and $\{\xi'_k\}_{k\in\mathbb{N}}$ for $\bar\theta_n^{(\gamma)}$ and $\bar\theta_n^{(2\gamma)}$ increase the leading component of the MSE by a constant factor. Hence, it is preferable to use synchronous noise construction as introduced in (36).

Proposition 2 implies the following improved bound on the bias of $\bar\theta_n^{(RR)}$:

Proposition 5. *Assume **A**1, **A**2, **A**3(6), and **C**1(6). Then, for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, and any initial distribution ν on $\mathbb{R}^d$, it holds that*

$$\mathsf{E}_\nu[\bar\theta_n^{(RR)}] = \theta^\star + B_3\gamma^{3/2} + \mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n)\,, \tag{40}$$

where $B_3 \in \mathbb{R}^d$ is a vector such that $\|B_3\| \le \mathsf{C}_1$, and

$$\|\mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n)\| \lesssim \tfrac{\mathrm{e}^{-\gamma\mu(n+1)/2}}{n\gamma\mu}\left(\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^2] + \tfrac{\sqrt{\gamma}\tau_2}{\sqrt{\mu}}\right)\,.$$

The proof of Proposition 5 is provided in Appendix A. This result is a simple consequence of Proposition 5, since the linear in γ component of the bias $\gamma\Delta_1$ from (24) cancels out when computing $\bar\theta_n^{(RR)}$. We are now ready to formulate the main result for the Richardson-Romberg estimate $\bar\theta_n^{(RR)}$.

Theorem 6. *Assume **A**1, **A**2, **A**3(6), and **C**1(6). Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, initial distribution ν and $n \in \mathbb{N}$, the Richardson-Romberg estimator $\bar\theta_n^{(RR)}$ defined in (37) satisfies*

$$\mathsf{E}_\nu^{1/2}[\|\mathrm{H}^\star(\bar\theta_n^{(RR)} - \theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},1}\gamma^{1/2}}{n^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},2}}{\gamma^{1/2}n} + \mathsf{C}_{\mathrm{RR},3}\gamma^{3/2} + \frac{\mathsf{C}_{\mathrm{RR},4}\gamma}{n^{1/2}} \\ + \mathcal{R}_4(n,\gamma,\|\theta_0-\theta^\star\|)\,,$$

where the constants $\mathsf{C}_{\mathrm{RR},1}$ to $\mathsf{C}_{\mathrm{RR},4}$ are defined in (69) and

$$\mathcal{R}_4(n,\gamma,\|\theta_0-\theta^\star\|) = \tfrac{c_0\,\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu} \\ \times\left(\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6] + \mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^4] + \mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^2] + \tfrac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu}\right)\,,$$

with c_0 being an absolute constant.

Proof of Theorem 6 is provided in Appendix C. Similarly to Theorem 3, we can optimize the above bound setting γ depending upon n.

Corollary 7. *Under the assumptions of Theorem 6, by setting $\gamma = n^{-1/2}$ with $n \ge (\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})^2$, it holds that*

$$\mathsf{E}_\nu^{1/2}[\|\mathrm{H}^\star(\bar\theta_n^{(RR)} - \theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\big(n^{-3/4}\big) + \mathcal{R}_4(n, 1/\sqrt{n}, \|\theta_0-\theta^\star\|)\,. \tag{41}$$

Discussion. Note that the result of Corollary 7 is a counterpart of (5) with $\delta = 1/4$. This decay rate of the second order term is the same as for the Root-SGD algorithm, and in general can not be improved, see Li et al. (2022). However, the assumptions of Theorem 6 are stronger compared to the ones imposed by Li et al. (2022). In particular, in **A**2 we require that f is 4 times continuously differentiable and uniformly bounded. At the same time, Li et al. (2022) impose Lipschitz continuity of the Hessian of f, which is essentially equivalent to bounded 3-rd derivative of f. Our proof of Theorem 6 essentially relies on the 4-th order Taylor expansion, and it is not clear, if this assumption can be relaxed. We leave further investigations of this question for future research.

Now we aim to generalize the previous result for the p-th moment bounds with $p \ge 2$. The key technical element of our proof for the p-th moment bound is the following statement, which can be viewed as a version of Rosenthal's inequality (Rosenthal, 1970; Pinelis, 1994).

Proposition 8. *Let $p \geq 2$ and assume* **A** *1,* **A** *2,* **A** *3(2p), and* **C** *1(2p). Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},2p})]$, it holds that*

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\| \textstyle\sum_{k=0}^{n-1}\{\psi(\theta_k^{(\gamma)}) - \pi_\gamma(\psi)\|^p\big] \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\, p\tau_{2p}^2\sqrt{n\gamma}}{\mu^{3/2}} + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\,\tau_{2p}}{\mu^2}\,, \tag{42}$$

where ψ is defined in (38).

Discussion. The proof of Proposition 8 is provided in Appendix D.1. It is important to acknowledge that there are numerous Rosenthal-type inequalities for dependent sequences in the literature. Proposition 8 can be viewed as an analogue to the classical Rosenthal inequality for strongly mixing sequences, see (Rio, 2017, Theorem 6.3). However, it should be emphasized that the Markov chain $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ is geometrically ergodic under the assumptions **A**1-**A**3(p) only in sense of the weighted Wasserstein semi-metric $\mathbf{W}_c(\xi,\xi')$ with respect to a cost function c defined in (16). As a result, the sequence $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ does not necessarily satisfy strong mixing conditions. At the same time, $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ satisfies the τ-mixing condition, see Merlevède et al. (2011). However, the considered function $\psi(\theta)$ is quadratic, which makes the respective result of (Merlevède et al., 2011, Theorem 1) inapplicable. Similar Rosenthal-type inequalities have been explored in (Durmus et al., 2023), but in Proposition 8 we obtain the bound with tighter dependence of the right-hand side upon γ. Below we provide the p-th moment bound together with corollary for the step size γ optimized w.r.t. n.

Theorem 9. *Let $p \geq 2$ and assume* **A** *1,* **A** *2,* **A** *3(3p), and* **C** *1(3p). Then for any step size $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},3p})]$, initial distribution ν, and $n \in \mathbb{N}$, the estimator $\bar{\theta}_n^{(RR)}$ defined in (37) satisfies*

$$\begin{aligned}\mathsf{E}_{\nu}^{1/p}[\| \mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^p] \leq \frac{c_1\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\,p^{1/2}}{n^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},5}}{n\gamma^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},6}\gamma^{1/2}}{n^{1/2}} + \mathsf{C}_{\mathrm{RR},7}\gamma^{3/2} \\ + \frac{c_2 p\tau_p}{n^{1-1/p}} + \frac{\mathsf{C}_{\mathrm{RR},8}}{n} + \mathcal{R}_5(n,\gamma,\|\theta_0-\theta^\star\|)\,,\end{aligned} \tag{43}$$

where c_1 and c_2 the absolute constants form the Pinelis version of Rosenthal inequality (Pinelis, 1994, Theorem 4.1), problem-specific constants $\mathsf{C}_{\mathrm{RR},5}$ to $\mathsf{C}_{\mathrm{RR},8}$ are defined in (96), and

$$\mathcal{R}_5(n,\gamma,\|\theta_0-\theta^\star\|) = (1-\gamma\mu)^{(n+1)/2}C_{f,p}\big(\mathsf{E}_\nu^{1/p}\big[\|\theta_0-\theta^\star\|^p\big] + \mathsf{E}_\nu^{1/p}\big[\|\theta_0-\theta^\star\|^{2p}\big] + \mathsf{E}_\nu^{1/p}\big[\|\theta_0-\theta^\star\|^{3p}\big]\big)\,.$$

Here constant $C_{f,p}$ can be traced from (97).

Corollary 10. *Under the assumptions of Theorem 9, by setting $\gamma = n^{-1/2}$ with $n \geq (\mathrm{L}\,\mathsf{C}_{\mathrm{step},3p})^2$, it holds that*

$$\mathsf{E}_{\nu}^{1/p}[\| \mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\,p^{1/2}}{n^{1/2}} + \mathcal{O}\big(n^{-3/4}\big) + \mathcal{R}_5(n, 1/\sqrt{n}, \|\theta_0-\theta^\star\|)\,. \tag{44}$$

Discussion. Proof of Theorem 9 is provided in Appendix D. Note that the result above is a direct generalization of Theorem 6, which reveals the same optimal scaling of the step size γ with respect to n. To the best of our knowledge, this is the first analysis of a first-order method, which provides a bound for the second-order term of order $\mathcal{O}\big(n^{-3/4}\big)$ while keeping the precise leading term related to the minimax-optimal covariance matrix $\mathrm{H}^\star$. Such results were previously known only for the setting of linear stochastic approximation (LSA), which corresponds to the case of quadratic function f in the initial minimization problem (1), see Durmus et al. (2024). In such a case, no bias occurs: $\int_{\mathbb{R}^d}\theta\pi_\gamma(\mathrm{d}\theta) = \theta^\star$, and Polyak-Ruppert averaging with the step size $\gamma \simeq n^{-1/2}$ allows for the same scaling of the remainder terms in n, as in (44). This result can be found, for example, in Durmus et al. (2024, Theorem 1). Hence, the main result of Theorem 9 is as follows: Richardson-Romberg extrapolation applied to strongly convex minimization problems allows to restore the p-th moment error moment bounds from the LSA setting.

5 Numerical results

In this section we illustrate numerically the scale of the second-order terms in Corollary 7, that is, in (41). We recall the error representation (39), and move the term $\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$ to the

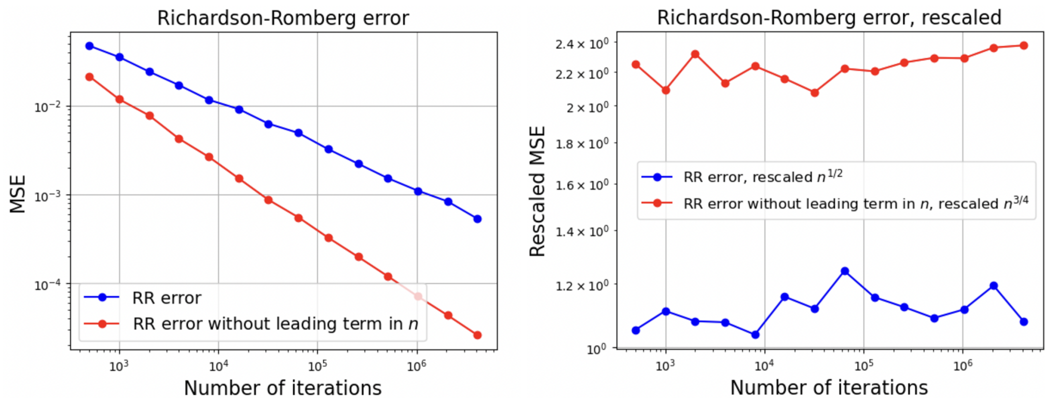

Figure 1: Left picture: Richardson-Romberg experimental error with and without the main term $\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$. Right picture: same errors after rescaling by $n^{1/2}$ and $n^{3/4}$, respectively.

right-hand side:

$$\mathrm{H}^\star(\bar{\theta}_n^{(\gamma)}-\theta^\star)+\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)=\frac{\theta_{n+1}^{(\gamma)}-\theta^\star}{\gamma n}-\frac{\theta_{2n}^{(\gamma)}-\theta^\star}{\gamma n}$$
$$-\frac{1}{n}\sum_{k=n+1}^{2n}\{\varepsilon_{k+1}(\theta_k^{(\gamma)})-\varepsilon_{k+1}(\theta^\star)\}-\frac{1}{n}\sum_{k=n+1}^{2n}\psi(\theta_k^{(\gamma)})-\frac{1}{n}\sum_{k=n+1}^{2n}G(\theta_k^{(\gamma)})\,. \quad (45)$$

Under **A**3(6), the statistics $\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$ is a sum of independent random variables, and

$$n^{-2}\mathsf{E}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)\|^2]=\frac{\operatorname{Tr}\Sigma_\varepsilon^\star}{n}\,.$$

Hence, in order to trace the rate of the second-order terms in (41), it is enough to find the decay rate of the right-hand side in (45). We show that, for a particular minimization problem, setting $\gamma=n^{-1/2}$, we achieve the desired scaling of order $n^{-3/4}$. We consider the minimization problem

$$f(\theta)=\theta^2+\cos\theta\to\min_{\theta\in\mathbb{R}}\,,$$

with the stochastic gradient oracles $\nabla F(\theta,\xi)$ given by $\nabla F(\theta,\xi)=2\theta-\sin\theta+\xi\,,\quad \xi\sim\mathcal{N}(0,1)$. This example clearly satisfies the assumptions **A**1, **A**2, **A**3(p) with any $p\geq 2$. We selected different sample sizes $n=250\times 2^k$, where $k=0,\ldots,14$, and run the SGD procedure (2) based on the constant step sizes γ and 2γ, selecting $\gamma=1/\sqrt{n}$. Then we construct the associated estimates $\{\bar{\theta}_n^{(\gamma)}\}$ and $\{\bar{\theta}_n^{(2\gamma)}\}$. Then for each n we compute the Richardson-Romberg estimates $\bar{\theta}_n^{(RR)}$ from (36) alongside with its versions without the leading term in n, i.e. $\bar{\theta}_n^{(RR)}+\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$. We provide first the plot for $\|\bar{\theta}_n^{(RR)}-\theta^\star\|$ and $\|\bar{\theta}_n^{(RR)}+\frac{1}{n}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)-\theta^\star\|$ in Figure 1. On the same figure we also provide the plots for rescaled errors

$$n^{1/2}\|\bar{\theta}_n^{(RR)}-\theta^\star\| \text{ and } n^{3/4}\|\bar{\theta}_n^{(RR)}+n^{-1}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)-\theta^\star\|\,.$$

The corresponding plot indicates that the proper scaling of the r.h.s. is $n^{-3/4}$, as predicted by Corollary 7. We provide code to reproduce the experiments at https://anonymous.4open.science/r/richardson_romberg_example-3DD4/.

6 Conclusion

In this paper, we have demonstrated that Polyak-Ruppert averaged SGD iterates with a constant step size achieve optimal root-MSE and p-th moment error bounds. More precisely, we have shown that these bounds admit both the sharp, optimal leading term, which aligns with the optimal covariance matrix Σ_∞, and a second-order term of order $\mathcal{O}(n^{-3/4})$, which is best known among the first order methods. Directions for future research include, firstly, generalizing the proposed algorithm to the setting of dependent noise sequences $\{\xi_k\}_{k\in\mathbb{N}}$ in the stochastic gradients (1). Another natural question is to study the properties of $\bar{\theta}_n^{(RR)}$ under relaxed assumptions on f. In particular, it would be interesting to remove additional smoothness assumptions on f (bounded 3-rd and 4-th derivatives), and to relax the strong convexity condition **A**1. One more research direction is to study a relation between the parameter δ in (5) and rates in the corresponding Berry-Esseen type results, following the technique of Shao & Zhang (2022).

References

Sebastian Allmeier and Nicolas Gast. Computing the bias of constant-step stochastic approximation with markovian noise. *arXiv preprint arXiv:2405.14285*, 2024.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/7fe1f8abaad094e0b5cb1b01d712f708-Paper.pdf.

V. Bally and D. Talay. The law of the euler scheme for stochastic differential equations. *Probability Theory and Related Fields*, 104(1):43–60, 1996. doi: 10.1007/BF01303802. URL https://doi.org/10.1007/BF01303802.

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.

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. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018. ISBN 978-3-319-97703-4.

Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient richardson-romberg markov chain monte carlo. *Advances in neural information processing systems*, 29, 2016.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Marina Sheshukova. Rosenthal-type inequalities for linear statistics of Markov chains. *arXiv preprint arXiv:2303.05838*, 2023.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 2024.

Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2022.11.012. URL https://www.sciencedirect.com/science/article/pii/S0304414922002447.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

Francis Begnaud Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 81–82, 2023.

Dongyan Huo, Yixuan Zhang, Yudong Chen, and Qiaomin Xie. The collusion of memory and nonlinearity in stochastic approximation with constant stepsize. *arXiv preprint arXiv:2405.16732*, 2024a.

Dongyan Lucy Huo, Yudong Chen, and Qiaomin Xie. Effectiveness of constant stepsize in markovian lsa and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20447–20455, 2024b.

Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 909–981. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/li22a.html.

Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant stepsize sgd through markov chains. *arXiv preprint arXiv:2306.11497*, 2023.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4): 435–474, 2011.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

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

A. Osekowski. *Sharp Martingale and Semimartingale Inequalities*. Monografie Matematyczne 72. Birkhäuser Basel, 1 edition, 2012. ISBN 3034803699,9783034803694.

Gilles Pagès. Multi-step Richardson-Romberg Extrapolation: Remarks on Variance Control and Complexity. *Monte Carlo Methods and Applications*, 13(1):37–70, 2007. doi: doi:10.1515/MCMA.2007.003. URL https://doi.org/10.1515/MCMA.2007.003.

Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994. doi: 10.1214/aop/1176988477. URL https://doi.org/10.1214/aop/1176988477.

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

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12 (1):145–151, 1999.

Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80. 2017. ISBN 978-3-662-54322-1. doi: 10.1007/978-3-662-54323-8.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Haskell P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970. ISSN 0021-2172. doi: 10.1007/BF02771562. URL https://doi.org/10.1007/BF02771562.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Qi-Man Shao and Zhuo-Song Zhang. Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3): 1548–1576, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990. doi: 10.1080/07362999008809220. URL https://doi.org/10.1080/07362999008809220.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Yixuan Zhang and Qiaomin Xie. Constant stepsize q-learning: Distributional convergence, bias and extrapolation. *arXiv preprint arXiv:2401.13884*, 2024.

Yixuan Zhang, Dongyan Huo, Yudong Chen, and Qiaomin Xie. Prelimit coupling and steady-state convergence of constant-stepsize nonsmooth contractive sa. *arXiv preprint arXiv:2404.06023*, 2024.

A Proof of Proposition 2 and Proposition 5

Throughout this appendix we use c_0 for an absolute constant, which values may vary from line to line.

A.1 Proof of Proposition 1

Consider the synchronous coupling construction defined by the recursions

$$\begin{aligned} \theta_{k+1}^{(\gamma)} &= \theta_k^{(\gamma)} - \gamma \nabla F(\theta_k^{(\gamma)}, \xi_{k+1}), \quad \theta_0^{(\gamma)} = \theta \in \mathbb{R}^d, \\ \tilde{\theta}_{k+1}^{(\gamma)} &= \tilde{\theta}_k^{(\gamma)} - \gamma \nabla F(\tilde{\theta}_k^{(\gamma)}, \xi_{k+1}), \quad \tilde{\theta}_0^{(\gamma)} = \tilde{\theta} \in \mathbb{R}^d. \end{aligned} \tag{46}$$

The pair $(\theta_k^{(\gamma)}, \tilde{\theta}_k^{(\gamma)})_{k\in\mathbb{N}}$ defines a Markov chain with the Markov kernel $\mathrm{K}_\gamma(\cdot,\cdot)$, which is a coupling kernel of $(\mathrm{Q}_\gamma, \mathrm{Q}_\gamma)$. From now on we omit an upper index (γ) and write simply $(\theta_k, \tilde{\theta}_k)_{k\in\mathbb{N}}$. Applying now **A**3(2), for $\gamma \leq 2/\mathrm{L}$, we get that

$$\begin{aligned} \mathsf{E}_\nu[\|\theta_{k+1} - \tilde{\theta}_{k+1}\|^2 | \mathcal{F}_k] &= \mathsf{E}[\|\theta_k - \tilde{\theta}_k - \gamma(\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}))\|^2 | \mathcal{F}_k] \\ &= \|\theta_k - \tilde{\theta}_k\|^2 + \gamma^2 \mathsf{E}[\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^2 | \mathcal{F}_k] \\ &\qquad - 2\gamma \langle \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k \rangle \\ &\leq (1 - \gamma\mu)\|\theta_k - \tilde{\theta}_k\|^2, \end{aligned} \tag{47}$$

where in the last inequality we additionally used $(1 - 2\gamma\mu(1 - \gamma\mathrm{L}/2)) \leq 1 - \gamma\mu$. Similarly, for a cost function c defined in (16), we get using Hölder's and Minkowski's inequalities, that for any $r \in \mathbb{N}$

$$\begin{aligned} \mathsf{E}[c(\theta_{k+r}, \tilde{\theta}_{k+r}) | \mathcal{F}_k] \leq \mathsf{E}^{1/2}[\|\theta_{k+r} - \tilde{\theta}_{k+r}\|^2 | \mathcal{F}_k] \big(&\mathsf{E}^{1/2}[\|\theta_{k+r} - \theta^\star\|^2 | \mathcal{F}_k] \\ &+ \mathsf{E}^{1/2}[\|\tilde{\theta}_{k+r} - \theta^\star\|^2 | \mathcal{F}_k] + \frac{2^{3/2}\gamma^{1/2}\tau_2}{\mu^{1/2}} \big). \end{aligned}$$

Combining the above inequalities and using (64), we obtain

$$\begin{aligned} \mathsf{E}[c(\theta_{k+r}, \theta'_{k+r}) | \mathcal{F}_k] &\leq (1-\gamma\mu)^{r/2} \|\theta_k - \tilde{\theta}_k\| \big((1-\gamma\mu)^{r/2} (\|\theta_k - \theta^\star\| + \|\tilde{\theta}_k - \theta^\star\|) + \frac{2^{5/2}\gamma^{1/2}\tau_2}{\mu^{1/2}} \big) \\ &\leq 2(1-\gamma\mu)^{r/2} c(\theta_k, \theta'_k). \end{aligned}$$

Note that $2(1-\gamma\mu)^{r/2} \leq 2$ for any $r \leq m(\gamma) - 1$ and $2(1-\gamma\mu)^{m(\gamma)/2} \leq 1/2$. Hence, applying the theorem Douc et al. (2018, Theorem 20.3.4), we obtain that the Markov kernel Q_γ admits a unique invariant distribution π_γ. Moreover,

$$\mathbf{W}_c(\nu \mathrm{Q}_\gamma^k, \pi_\gamma) \leq 2(1/2)^{\lfloor k/m(\gamma) \rfloor} \mathbf{W}_c(\nu, \pi_\gamma). \tag{48}$$

It remains to note that $(1/2)^{\lfloor k/m(\gamma) \rfloor} \leq 2(1/2)^{k/m(\gamma)}$.

A.2 Proof of Proposition 2

We begin with proving (22) and (23). First we introduce some additional notations. Under assumptions **A**1 – **A**3(2), we define a matrix-valued function $\mathcal{C}(\theta) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as

$$\mathcal{C}(\theta) = \mathsf{E}[\varepsilon_1(\theta)^{\otimes 2}]. \tag{49}$$

The result below is essentially based on an appropriate modification of the bounds presented in Dieuleveut et al. (2020, Lemma 18). A careful inspection of the respective proof reveals that we do not need specific assumptions for $\mathcal{C}(\theta)$ defined in (49), instead we use Lemma 12. For completeness, we present the respective result below:

Lemma 11. *Assume **A**1, **A**2, **A**3(6), and **C**1(6). Then, for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, it holds*

$$\bar{\theta}_\gamma - \theta^\star = -(\gamma/2)\{\mathrm{H}^\star\}^{-1}\{\nabla^3 f(\theta^\star)\}\mathbf{T}\mathcal{C}(\theta^\star) + B_1\gamma^{3/2}, \tag{50}$$

where $\bar{\theta}_\gamma$ is defined in (22), $\mathcal{C}(\theta)$ is defined in (49), and $B_1 \in \mathbb{R}^d$ satisfies $\|B_1\| \leq \mathsf{C}_1$, where

$$\mathsf{C}_1 = \left(\frac{\mathrm{L}^2 \,\mathsf{D}_{\mathrm{last},2}}{\sqrt{\mu}} + \frac{\mathrm{L}\sqrt{\mathsf{D}_{\mathrm{last},2}}}{\sqrt{\mu}} \right) \frac{\tau_2^2}{\mu} + \frac{\mathrm{L}}{\mu} \,. \tag{51}$$

Moreover,

$$\bar{\Sigma}_\gamma = \gamma \mathbf{T} \mathcal{C}(\theta^\star) + B_2 \gamma^{3/2} \,, \tag{52}$$

where the operator $\mathbf{T} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d \times d}$ is defined by the relation

$$\operatorname{vec}(\mathbf{T}A) = (\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star)^{-1} \operatorname{vec}(A)$$

for any matrix $A \in \mathbb{R}^{d\times d}$, and $B_2 \in \mathbb{R}^{d\times d}$ is a matrix, such that $\|B_2\| \leq \mathsf{C}_1$.

Proof. Let $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ be a recurrence defined in (11) with initial distribution $\theta_0 \sim \pi_\gamma$. Moreover, we assume that θ_0 is independent of the noise variables $(\xi_k)_{k\geq 1}$. First, applying a third-order Taylor expansion of $\nabla f(\theta)$ around $\theta^\star$, for any $\theta \in \mathbb{R}^d$, we obtain

$$\nabla f(\theta) = \mathrm{H}^\star(\theta - \theta^\star) + (1/2)\{\nabla^3 f(\theta^\star)\}(\theta - \theta^\star)^{\otimes 2} + G(\theta) \,, \tag{53}$$

where $G(\theta)$ is defined in (38) and writes as

$$G(\theta) = \frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\, dt \right)(\theta - \theta^\star)^{\otimes 3} \,.$$

Thus, using **A**2,

$$\|G(\theta)\| \lesssim \mathrm{L}_4 \|\theta - \theta^\star\|^3 \,.$$

Thus, integrating (53) with respect to π_γ, we get from (53) that

$$\mathrm{H}^\star(\bar{\theta}_\gamma - \theta^\star) + (1/2)\{\nabla^3 f(\theta^\star)\}\left[\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \right] = -\int_{\mathbb{R}^d} G(\theta)\pi_\gamma(\mathrm{d}\theta) \,. \tag{54}$$

Now we need to provide an explicit expression for the covariance matrix

$$\bar{\Sigma}_\gamma = \int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \,. \tag{55}$$

Using the recurrence (11), we obtain that

$$\theta_1 - \theta^\star = (\mathrm{I} - \gamma\,\mathrm{H}^\star)(\theta_0 - \theta^\star) - \gamma\varepsilon_1(\theta_0) - \gamma\eta(\theta_0) \,,$$

where the function $\eta(\cdot)$ is defined in (28). Hence, taking second moment w.r.t. π_γ from both sides, we get that

$$\begin{aligned} \bar{\Sigma}_\gamma = (\mathrm{I} - \gamma\,\mathrm{H}^\star)\bar{\Sigma}_\gamma(\mathrm{I} - \gamma\,\mathrm{H}^\star) + \gamma^2 \int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) + \gamma^2 \int_{\mathbb{R}^d} \{\eta(\theta)\}^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \\ - \gamma \int_{\mathbb{R}^d} \left[(\mathrm{I} - \gamma\,\mathrm{H}^\star)(\theta - \theta^\star)\{\eta(\theta)\}^\top + \eta(\theta)(\theta - \theta^\star)^\top(\mathrm{I} - \gamma\,\mathrm{H}^\star) \right] \pi_\gamma(\mathrm{d}\theta) \,. \end{aligned} \tag{56}$$

In the above equation $\mathcal{C}(\theta)$ is defined in (49), and we additionally used that $\mathsf{E}\left[\varepsilon_1(\theta_0)|\mathcal{F}_0\right] = 0$. Moreover, (49) together with **C**1(6) implies that

$$\int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) = \mathcal{C}(\theta^\star) + B\gamma^{1/2} \,,$$

where $B \in \mathbb{R}^{d\times d}$ satisfies $\|B\| \leq \mathsf{C}_2$. Thus, from (56) together with **C**1(6) we obtain that $\bar{\Sigma}_\gamma$ is a solution to the matrix equation

$$\mathrm{H}^\star\,\bar{\Sigma}_\gamma + \bar{\Sigma}_\gamma\,\mathrm{H}^\star - \gamma\,\mathrm{H}^\star\,\bar{\Sigma}_\gamma\,\mathrm{H}^\star = \gamma\mathcal{C}(\theta^\star) + B\gamma^{3/2} \,,$$

which can be written using the vectorization operation as

$$\operatorname{vec}\left(\bar{\Sigma}_\gamma\right) = \gamma(\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star - \gamma\,\mathrm{H}^\star \otimes \mathrm{H}^\star)^{-1} \operatorname{vec}(\mathcal{C}(\theta^\star)) + \gamma^{3/2}(\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star - \gamma\,\mathrm{H}^\star \otimes \mathrm{H}^\star)^{-1} \operatorname{vec}(B) \,.$$

Now we check that the latter operator $\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star - \gamma \mathrm{H}^\star \otimes \mathrm{H}^\star$ is indeed invertible for $\gamma \in (0, 2/\mathrm{L})$. Moreover, assumption **A**1 guarantees that the symmetric matrix $\mathrm{H}^\star$ is non-degenerate and positive-definite. Let $u_1, \ldots, u_d \in \mathbb{R}^d$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq \mu > 0$ be its eigenvectors and eigenvalues, respectively. Then we notice that

$$\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star - \gamma \mathrm{H}^\star \otimes \mathrm{H}^\star = \mathrm{H}^\star \otimes (\mathrm{I} - (\gamma/2)\mathrm{H}^\star) + (\mathrm{I} - (\gamma/2)\mathrm{H}^\star) \otimes \mathrm{H}^\star \,.$$

Hence, the latter operator is also diagonalizable in the orthogonal basis $u_i \otimes u_j$ in $\mathbb{R}^{d^2}$ with the respective eigenvalues being equal to $\lambda_i(1 - (\gamma/2)\lambda_j) + \lambda_j(1 - (\gamma/2)\lambda_i)$. Set now

$$\begin{aligned} S &= \mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star \in \mathbb{R}^{d^2 \times d^2} \\ R &= \mathrm{H}^\star \otimes \mathrm{H}^\star \in \mathbb{R}^{d^2 \times d^2} \,. \end{aligned} \tag{57}$$

Then it is easy to observe that

$$(S - \gamma R)^{-1} = S^{-1} + S^{-1} \sum_{k=1}^{\infty} \gamma^k (RS^{-1})^k \,,$$

provided that $\gamma \|RS^{-1}\| \leq 1$. Sice R and S are diagonalizable in the same orthogonal basis $\{u_i \otimes u_j\}_{1 \leq i,j \leq d}$ with the eigenvalues $\lambda_i \lambda_j$ and $\lambda_i + \lambda_j$, respectively, the condition $\gamma \|RS^{-1}\| \leq 1$ holds provided that $\gamma \leq 2/\mathrm{L}$. Hence, for $\gamma \leq 1/\mathrm{L}$, it holds that

$$(\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star - \gamma \mathrm{H}^\star \otimes \mathrm{H}^\star)^{-1} = (\mathrm{H}^\star \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{H}^\star)^{-1} + D \,,$$

where $D \in \mathbb{R}^{d^2 \times d^2}$ satisfies

$$\|D\| \lesssim \gamma \|S\|^{-1} \|RS^{-1}\| \lesssim \frac{\gamma \mathrm{L}}{\mu} \,.$$

Combining the above bounds in (54), we arrive at the expansion formula (50). $\square$

We now state an auxiliary lemma about the function $\mathcal{C}(\theta)$ from (49).

Lemma 12. *Assume **A**1, **A**2, **A**3(2), and **C**1(2). Then, for any $\gamma \in (0, 1/(\mathrm{L}\, \mathsf{C}_{\mathrm{step},2})]$, it holds*

$$\| \int_{\mathbb{R}^d} \mathcal{C}(\theta) \pi_\gamma(\mathrm{d}\theta) - \mathcal{C}(\theta^\star) \| \leq \mathsf{C}_2 \gamma^{1/2} \,,$$

where the constant C_2 is given by

$$\mathsf{C}_2 = \left(\frac{\mathrm{L}^2 \, \mathsf{D}_{\mathrm{last},2}}{\sqrt{\mu}} + \frac{\mathrm{L} \sqrt{\mathsf{D}_{\mathrm{last},2}}}{\sqrt{\mu}} \right) \tau_2^2 \,. \tag{58}$$

Proof. Recall that

$$\varepsilon_1(\theta) = \nabla F(\theta, \xi_1) - \nabla f(\theta) \,.$$

Hence, using the definition of $\mathcal{C}(\theta)$ in (49), we get

$$\begin{aligned} \mathcal{C}(\theta) - \mathcal{C}(\theta^\star) = \mathsf{E}_{\xi_1 \sim \mathbb{P}_\xi}[(\varepsilon_1(\theta) - \varepsilon_1(\theta^\star))(\varepsilon_1(\theta) - \varepsilon_1(\theta^\star))^T] + \mathsf{E}_{\xi_1 \sim \mathbb{P}_\xi}[\varepsilon_1(\theta^\star)(\varepsilon_1(\theta) - \varepsilon_1(\theta^\star))^T] \\ + \mathsf{E}_{\xi_1 \sim \mathbb{P}_\xi}[(\varepsilon_1(\theta) - \varepsilon_1(\theta^\star))\varepsilon_1(\theta^\star)^T]. \end{aligned}$$

Using **A**3(2), we obtain

$$\mathsf{E}_\xi[\|\varepsilon_1(\theta) - \varepsilon_1(\theta^\star)\|^2] \lesssim \mathrm{L} \langle \nabla f(\theta) - \nabla f(\theta^\star), \theta - \theta^\star \rangle - \|\nabla f(\theta) - \nabla f(\theta^\star)\|^2 \lesssim \mathrm{L}^2 \|\theta - \theta^\star\|^2.$$

Hence, combining the previous inequalities and using Hölder's inequality, we obtain for any $\theta \in \mathbb{R}^d$, that

$$\|\mathcal{C}(\theta) - \mathcal{C}(\theta^\star)\| \lesssim \mathrm{L}^2 \|\theta - \theta^\star\|^2 + \tau_2 \mathrm{L} \|\theta - \theta^\star\|.$$

Applying now **C**1(2), we obtain

$$\| \int_{\mathbb{R}^d} \mathcal{C}(\theta) \pi_\gamma(\mathrm{d}\theta) - \mathcal{C}(\theta^\star) \| \leq \int_{\mathbb{R}^d} \|\mathcal{C}(\theta) - \mathcal{C}(\theta^\star)\| \pi_\gamma(\mathrm{d}\theta) \lesssim \mathrm{L}^2 \frac{\mathsf{D}_{\mathrm{last},2} \gamma \tau_2^2}{\mu} + \tau_2 \mathrm{L} \sqrt{\frac{\mathsf{D}_{\mathrm{last},2} \gamma \tau_2^2}{\mu}}.$$

We conclude the proof by noting that $\gamma \mu \leq 1$. $\square$

Now we prove (24). We use synchronous coupling construction defined by the pair of recursions:

$$\theta_{k+1} = \theta_k - \gamma \nabla F(\theta_k, \xi_{k+1}), \quad \theta_0 \sim \nu$$
$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma \nabla F(\tilde{\theta}_k, \xi_{k+1}), \quad \tilde{\theta}_0 \sim \pi_\gamma \,.$$

Recall that the corresponding coupling kernel is denoted as $\mathrm{K}_\gamma(\cdot,\cdot)$. Then we obtain

$$\begin{aligned}\mathsf{E}_\nu[\bar{\theta}_n] - \theta^\star &= n^{-1} \sum_{k=n+1}^{2n} \mathsf{E}^{\mathrm{K}_\gamma}_{\nu,\pi_\gamma}[\theta_k - \tilde{\theta}_k] + n^{-1} \sum_{k=n+1}^{2n} \mathsf{E}_{\pi_\gamma}[\tilde{\theta}_k - \theta^\star] \\ &= n^{-1} \sum_{k=n+1}^{2n} \mathsf{E}^{\mathrm{K}_\gamma}_{\nu,\pi_\gamma}[\theta_k - \tilde{\theta}_k] + (\bar{\theta}_\gamma - \theta^\star)\,.\end{aligned}$$

Using (47) and **C**1(2), we obtain

$$\begin{aligned}\|\mathsf{E}^{K}_{\nu,\pi_\gamma}[\{\theta_k - \tilde{\theta}_k\}]\| &\le (1-\gamma\mu)^{k/2} \{\mathsf{E}^{\mathrm{K}_\gamma}_{\nu,\pi_\gamma}\|\theta_0 - \tilde{\theta}_0\|^2\}^{1/2} \\ &\le (1-\gamma\mu)^{k/2} (\mathsf{E}^{1/2}_\nu[\|\theta_0 - \theta^\star\|^2] + \frac{\sqrt{2\gamma}\tau_2}{\sqrt{\mu}})\,.\end{aligned}$$

Summing the above bounds for k from $n+1$ to $2n$, we obtain (24).

A.3 Proof of Proposition 5

Note that

$$\mathsf{E}_\nu[\bar{\theta}^{(RR)}_n - \theta^\star] = 2\mathsf{E}_\nu[\bar{\theta}^\gamma_n - \theta^\star] - \mathsf{E}_\nu[\bar{\theta}^{2\gamma}_n - \theta^\star].$$

Applying (24), we obtain

$$\|\mathsf{E}_\nu[\bar{\theta}^{(RR)}_n - \theta^\star]\| \lesssim \mathsf{C}_1 \gamma^{3/2} + \mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n), \tag{59}$$

where

$$\|\mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n)\| \lesssim \frac{(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu} (\mathsf{E}^{1/2}_\nu[\|\theta_0 - \theta^\star\|^2] + \frac{\sqrt{\gamma}\tau_2}{\sqrt{\mu}})\,. \tag{60}$$

B PROOF OF THEOREM 3

Theorem 13 (Version of Theorem 3 with explicit constants)**.** *Assume* **A**1*,* **A**2*,* **A**3(6)*, and* **C**1(6)*. Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, $n \in \mathbb{N}$, and initial distribution ν on $\mathbb{R}^d$, the sequence of Polyak-Ruppert estimates (14) satisfies*

$$\mathsf{E}_\nu^{1/2}[\| \mathrm{H}^\star(\bar{\theta}_n^{(\gamma)} - \theta^\star)\|^2] \leq \frac{\sqrt{\operatorname{Tr} \Sigma_\varepsilon^\star}}{\sqrt{n}} + \frac{\mathsf{C}_2}{\gamma^{1/2} n} + \mathsf{C}_3 \gamma + \frac{\mathsf{C}_4 \gamma^{1/2}}{n^{1/2}} + \mathcal{R}_2(n, \gamma, \|\theta_0 - \theta^\star\|) \,,$$

where we have set

$$\mathsf{C}_2 = c_0 \mathsf{D}_{\mathrm{last},2}^{1/2} \tau_2 \,, \quad \mathsf{C}_3 = c_0 \frac{\mathrm{L}\, \mathsf{D}_{\mathrm{last},4} \tau_4^2}{2\mu} \,, \quad \mathsf{C}_4 = c_0\, \mathrm{L}\, \mathsf{D}_{\mathrm{last},2}^{1/2} \tau_2 \,. \tag{61}$$

and the remainder term $\mathcal{R}_2(n, \gamma, \|\theta_0 - \theta^\star\|)$ is given by

$$\begin{aligned} \mathcal{R}_2(n, \gamma, \|\theta_0 - \theta^\star\|) = \frac{c_0\, \mathrm{L} (1-\gamma\mu)^{(n+1)/2}}{\gamma \mu n} \mathsf{E}_\nu^{1/2} \big[\|\theta_0 - \theta^\star\|^2\big] \\ + \frac{\mathrm{L}\, c_0 (1-\gamma\mu)^{n+1}}{2 n \gamma \mu} \mathsf{E}_\nu^{1/2} \big[\|\theta_0 - \theta^\star\|^4\big] \,. \end{aligned} \tag{62}$$

Proof. Throughout the proof we omit upper index (γ) both for the elements of the sequence $\{\theta_k^{(\gamma)}\}_{k \in \mathbb{N}}$ and Polyak-Ruppert averaged estimates $\bar{\theta}_n^{(\gamma)}$. Instead, we write simply θ_k and $\bar{\theta}_n$, respectively. Summing the recurrence (31), we obtain that

$$\mathrm{H}^\star(\bar{\theta}_n - \theta^\star) = \frac{\theta_{n+1} - \theta^\star}{\gamma n} - \frac{\theta_{2n} - \theta^\star}{\gamma n} - \frac{1}{n} \sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta_k) - \frac{1}{n} \sum_{k=n+1}^{2n} \eta(\theta_k) \,. \tag{63}$$

Applying the 3-rd order Taylor expansion with integral remainder, we get that

$$\nabla f(\theta_k) = \mathrm{H}^\star(\theta_k - \theta^\star) + \frac{1}{2} \left(\int_0^1 \nabla^3 f(t\theta^\star + (1-t)\theta_k)\, dt \right) (\theta_k - \theta^\star)^{\otimes 2} \,,$$

where $\nabla^3 f(\cdot) \in \mathbb{R}^{d \times d \times d}$. Using **A**2, we thus obtain that

$$\|\eta(\theta_k)\| \leq \frac{1}{2} \mathrm{L}_3 \, \|\theta_k - \theta^\star\|^2 \,.$$

Applying Minkowski's inequality to the decomposition (33) and to the last term therein, we get

$$\begin{aligned} \mathsf{E}_\nu^{1/2}[\| \mathrm{H}^\star(\bar{\theta}_n - \theta^\star)\|^2] \leq \frac{\mathsf{E}_\nu^{1/2}[\|\theta_{n+1} - \theta^\star\|^2]}{\gamma n} + \frac{\mathsf{E}_\nu^{1/2}[\|\theta_{2n} - \theta^\star\|^2]}{\gamma n} + \frac{1}{n} \mathsf{E}_\nu^{1/2}\Big[\| \sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta_k)\|^2\Big] \\ + \frac{\mathrm{L}_3}{2n} \sum_{k=n+1}^{2n} \mathsf{E}_\nu^{1/2}\big[\|\theta_k - \theta^\star\|^4\big] \,. \end{aligned}$$

Applying **C**1(2), we obtain that for $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},2})]$ it holds that

$$\mathsf{E}_\nu \|\theta_k - \theta^\star\|^2 \lesssim (1-\gamma\mu)^k \mathsf{E}_\nu\big[\|\theta_0 - \theta^\star\|^2\big] + \frac{\mathsf{D}_{\mathrm{last},2} \gamma \tau_2^2}{\mu} \,. \tag{64}$$

Moreover, from $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},4})]$ it holds that

$$\mathsf{E}_\nu^{1/2} \|\theta_k - \theta^\star\|^4 \lesssim (1-\gamma\mu)^k \mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^4\big] + \frac{\mathsf{D}_{\mathrm{last},4} \gamma \tau_4^2}{\mu} \,. \tag{65}$$

Combining Lemma 14 with previous inequalities, we obtain

$$\begin{aligned} \mathsf{E}_\nu^{1/2}[\| \mathrm{H}^\star(\bar{\theta}_n - \theta^\star)\|^2] \lesssim \frac{\sqrt{\operatorname{Tr} \Sigma_\varepsilon^\star}}{\sqrt{n}} + \frac{\mathsf{D}_{\mathrm{last},2}^{1/2} \tau_2}{\gamma^{1/2} n} + \frac{\mathrm{L}\, \mathsf{D}_{\mathrm{last},4} \gamma \tau_4^2}{2\mu} + \frac{\mathrm{L}\, \mathsf{D}_{\mathrm{last},2}^{1/2} \gamma^{1/2} \tau_2}{\mu^{1/2} n^{1/2}} \\ + \frac{(1-\gamma\mu)^{(n+1)/2}}{\gamma n} \left(\frac{\mathrm{L}}{\mu} + 1 \right) \mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^2\big] + \frac{\mathrm{L} (1-\gamma\mu)^{n+1}}{n \gamma \mu} \mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^4\big] \,, \end{aligned}$$

and the result follows. □

Lemma 14. *Assume* **A**1, **A**3(2), **A**2, *and* **C**1(2). *Then for any* $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},2})]$ *and any* $n \in \mathbb{N}$, *it holds*

$$\mathsf{E}_\nu^{1/2}[\| \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^2] \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2}^{1/2}\sqrt{\gamma n}\tau_2}{\mu^{1/2}} + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{\gamma\mu}\mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^2]\,. \tag{66}$$

Moreover, under **A**1, **A**3(p), **A**2, *and* **C**1(p), *for any* $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},p})]$ *and* $n \in \mathbb{N}$ *it holds that*

$$\begin{aligned}\mathsf{E}_\nu^{1/p}[\| \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^p] &\lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},p}^{1/2}\sqrt{\gamma n}p\tau_p}{\mu^{1/2}} \\ &\quad + \frac{\mathrm{L}\,p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2}}\mathsf{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p]\,.\end{aligned} \tag{67}$$

Proof. Since $\{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}$ is a martingale-difference sequence with respect to $\mathcal{F}_k$, we have

$$\mathsf{E}_\nu\big[\| \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^2\big] = \sum_{k=n+1}^{2n} \mathsf{E}_\nu\big[\|\{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^2\big]\,.$$

where $\varepsilon_{k+1}(\theta^\star) = \nabla F(\theta^\star, \xi_{k+1})$ uses the same noise variable ξ_{k+1} as $F(\theta_k, \xi_{k+1})$. Note that

$$\begin{aligned}&\mathsf{E}_\nu[\|\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^2] = \mathsf{E}_\nu[\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\theta^\star, \xi_{k+1})\|^2 \\ &\quad - 2\mathsf{E}_\nu\big[\langle \nabla F(\theta_k, \xi_{k+1}) - \nabla F(\theta^\star, \xi_{k+1}), \nabla f(\theta_k) - \nabla f(\theta^\star)\rangle\big] + \|\nabla f(\theta_k) - \nabla f(\theta^\star)\|^2]\,.\end{aligned}$$

Using **A**2, **A**3(2), and taking conditional expectation with respect to $\mathcal{F}_k$, we obtain

$$\begin{aligned}\mathsf{E}_\nu[\|\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^2] &\leq \mathsf{E}_\nu[\mathrm{L}\langle \nabla f(\theta_k) - \nabla f(\theta^\star), \theta_k - \theta^\star\rangle - \|\nabla f(\theta_k) - \nabla f(\theta^\star)\|^2] \\ &\leq \mathrm{L}^2\,\mathsf{E}_\nu[\|\theta_k - \theta^\star\|^2].\end{aligned}$$

Thus, we obtain that

$$\mathsf{E}_\nu[\| \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^2] \leq \mathrm{L}^2 \sum_{k=n+1}^{2n} \mathsf{E}_\nu[\|\theta_k - \theta^\star\|^2]\,,$$

and the statement (66) follows from the assumption **C** 1(2). In order to prove (67), we apply Burkholder's inequality Osekowski (2012, Theorem 8.6) and obtain

$$\begin{aligned}\mathsf{E}_\nu^{1/p}[\| \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^p] &\leq p\mathsf{E}_\nu^{1/p}\big[\big(\sum_{k=n+1}^{2n} \|\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^2\big)^{p/2}\big] \\ &\leq p\big(\sum_{k=n+1}^{2n} \mathsf{E}_\nu^{2/p}\big[\|\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^p\big]\big)^{1/2} \\ &\lesssim p\,\mathrm{L}\big(\sum_{k=n+1}^{2n} \mathsf{E}_\nu^{2/p}\big[\|\theta_k - \theta^\star\|^p\big]\big)^{1/2} \\ &\overset{(a)}{\lesssim} \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},p}^{1/2}\sqrt{\gamma n}p\tau_p}{\mu^{1/2}} + \frac{\mathrm{L}\,p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2}}\mathsf{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p]\,,\end{aligned}$$

where in (a) we have additionally used **C**1(p). □

C PROOF OF THEOREM 6

Within this section we often use the definition of the function $\psi : \mathbb{R}^d \to \mathbb{R}^d$ from (38):

$$\psi(\theta) = (1/2)\nabla^3 f(\theta^*)(\theta - \theta^\star)^{\otimes 2} \tag{68}$$

Theorem 15 (Version of Theorem 6 with explicit constants)**.** *Assume* ***A**1*, ***A**2*, ***A**3*(6), *and* ***C**1*(6). *Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, initial distribution ν and $n \in \mathbb{N}$, the Richardson-Romberg estimator $\bar{\theta}_n^{(RR)}$ defined in (37) satisfies*

$$\mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^2] \leq \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},1}\gamma^{1/2}}{n^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},2}}{\gamma^{1/2}n} + \mathsf{C}_{\mathrm{RR},3}\gamma^{3/2} + \frac{\mathsf{C}_{\mathrm{RR},4}\gamma}{n^{1/2}} + \mathcal{R}_4(n, \gamma, \|\theta_0 - \theta^\star\|)\,,$$

where we have set

$$\begin{aligned} \mathsf{C}_{\mathrm{RR},1} &= \frac{c_0 \mathsf{D}_{\mathrm{last},4}\,\mathrm{L}\,\tau_4^2}{\mu^{3/2}} + \frac{c_0\,\mathrm{L}\,\mathsf{D}_{\mathrm{last},2}^{1/2}\tau_2}{\mu^{1/2}}\,, \quad \mathsf{C}_{\mathrm{RR},2} = \frac{c_0 \mathsf{D}_{\mathrm{last},2}^{1/2}\tau_2}{\mu^{1/2}} \\ \mathsf{C}_{\mathrm{RR},3} &= c_0\left(\frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},6}^{3/2}\tau_6^3}{\mu^{3/2}} + \mathsf{C}_1\right)\,, \quad \mathsf{C}_{\mathrm{RR},4} = \frac{c_0 \mathsf{D}_{\mathrm{last},4}\,\mathrm{L}\,\tau_4^2}{\mu}\,, \end{aligned} \tag{69}$$

C_1 *is defined in* (51)*, and the remainder term $\mathcal{R}_4(n, \gamma, \|\theta_0 - \theta^\star\|)$ is given by*

$$\mathcal{R}_4(n, \gamma, \|\theta_0 - \theta^\star\|) = \frac{c_0\,\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu} \times \left(\mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^6] + \mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^4] + \mathsf{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^2] + \frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu}\right)\,. \tag{70}$$

Proof. Using the recursion (31), we obtain that

$$\begin{aligned} \mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star) &= \frac{2(\theta_{n+1}^{(\gamma)} - \theta^\star)}{\gamma n} - \frac{2(\theta_{2n}^{(\gamma)} - \theta^\star)}{\gamma n} - \frac{\theta_{n+1}^{(2\gamma)} - \theta^\star}{2\gamma n} + \frac{\theta_{2n}^{(2\gamma)} - \theta^\star}{2\gamma n} \\ &\quad - \frac{1}{n}\sum_{k=n+1}^{2n}[2\varepsilon_{k+1}(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta_k^{(2\gamma)})] - \frac{1}{n}\sum_{k=n+1}^{2n}[2\eta(\theta_k^{(\gamma)}) - \eta(\theta_k^{(2\gamma)})]\,. \end{aligned} \tag{71}$$

Therefore, applying Minkowski's inequality to the decomposition (71), we obtain for any initial distribution ν that

$$\begin{aligned} \mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^2] &\lesssim \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)\|^2]}_{T_1} + \underbrace{\frac{1}{\gamma n}\mathsf{E}_\nu^{1/2}[\|\theta_{n+1}^{(\gamma)} - \theta^\star\|^2] + \frac{1}{\gamma n}\mathsf{E}_\nu^{1/2}[\|\theta_{2n}^{(\gamma)} - \theta^\star\|^2]}_{T_2} \\ &\quad + \underbrace{\frac{1}{\gamma n}\mathsf{E}_\nu^{1/2}[\|\theta_{n+1}^{(2\gamma)} - \theta^\star\|^2] + \frac{1}{\gamma n}\mathsf{E}_\nu^{1/2}[\|\theta_{2n}^{(2\gamma)} - \theta^\star\|^2]}_{T_3} \\ &\quad + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta^\star)\|^2]}_{T_4} \\ &\quad + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta_k^{(2\gamma)}) - \varepsilon_{k+1}(\theta^\star)\|^2]}_{T_5} + \underbrace{\|2\pi_\gamma(\psi) - \pi_{2\gamma}(\psi)\|}_{T_6} \\ &\quad + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\eta(\theta_k^{(\gamma)}) - \pi_\gamma(\psi)\|^2] + \frac{1}{n}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\eta(\theta_k^{(2\gamma)}) - \pi_{2\gamma}(\psi)\|^2]}_{T_7}\,. \end{aligned}$$

Now we upper bound the terms in the right-hand side of the above bound separately. First, we note that

$$T_1 = \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{\sqrt{n}}\,.$$

Using **C**1(2), we get

$$T_2 + T_3 \lesssim \frac{(1-\gamma\mu)^{n+1/2}}{\gamma n}\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^2] + \frac{\mathsf{D}_{\text{last},2}^{1/2}\tau_2}{\mu^{1/2}\gamma^{1/2}n}\,.$$

Applying Lemma 14, we get

$$T_4 + T_5 \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\text{last},2}^{1/2}\gamma^{1/2}\tau_2}{\mu^{1/2}n^{1/2}} + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{\mu\gamma n}\mathsf{E}_\nu^{1/2}\left[\|\theta_0-\theta^\star\|^2\right].$$

Now we proceed with the term T_6. Applying the recurrence (11), we obtain that

$$\theta_1^{(\gamma)} - \theta^\star = (\mathrm{I} - \gamma\,\mathrm{H}^\star)(\theta_0^{(\gamma)} - \theta^\star) - \gamma\varepsilon_1(\theta_0^{(\gamma)}) - \gamma\eta(\theta_0^{(\gamma)})\,. \tag{72}$$

Thus, taking expectation w.r.t. π_γ in both sides above, we get

$$\mathrm{H}^\star(\bar{\theta}_\gamma - \theta^\star) = \mathsf{E}_{\pi_\gamma}\left[\eta(\theta_0^{(\gamma)})\right] = \pi_\gamma(\psi) + \pi_\gamma(G)\,,$$

where $G(\theta)$ is defined in (38) and writes as

$$G(\theta) = \frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\,dt\right)(\theta-\theta^\star)^{\otimes 3}\,.$$

Hence, applying **A**2 together with Proposition 2, we obtain that

$$T_6 = \|2\pi_\gamma(\psi) - \pi_{2\gamma}(\psi)\| \lesssim \mathsf{C}_1\gamma^{3/2}\,. \tag{73}$$

Finally, using Lemma 19, Lemma 18, and Lemma 16, we obtain that

$$\begin{aligned} T_7 \lesssim\; & \frac{\mathsf{D}_{\text{last},4}\,\mathrm{L}\,\gamma\tau_4^2}{\mu n^{1/2}} + \frac{\mathsf{D}_{\text{last},4}\,\mathrm{L}\,\gamma^{1/2}\tau_4^2}{\mu^{3/2}n^{1/2}} + \frac{\mathrm{L}\,\mathsf{D}_{\text{last},6}^{3/2}\gamma^{3/2}\tau_6^3}{\mu^{3/2}} \\ & + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}\left(\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6] + \mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^4] + \frac{\mathsf{D}_{\text{last},4}\gamma\tau_4^2}{\mu}\right). \end{aligned}$$

Combining the bounds above completes the proof. □

Below we provide some auxiliary technical lemmas.

Lemma 16. *Assume **A**1, **A**2, **A**3(4), and **C**1(4). Then for any $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\text{step},4})]$ and any $n \in \mathbb{N}$ it holds*

$$n^{-1}\mathsf{E}_{\pi_\gamma}^{1/2}[\|\sum_{k=n+1}^{2n}\{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \lesssim \frac{\mathsf{D}_{\text{last},4}\,\mathrm{L}_3\,\gamma\tau_4^2}{\mu n^{1/2}} + \frac{\mathsf{D}_{\text{last},4}\,\mathrm{L}_3\,\gamma^{1/2}\tau_4^2}{\mu^{3/2}n^{1/2}}\,. \tag{74}$$

Proof. Using the fact that π_γ is a stationary distribution, we obtain that

$$\begin{aligned} \mathsf{E}_{\pi_\gamma}\left[\|\sum_{k=n+1}^{2n}\{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2\right] &= n\mathsf{E}_{\pi_\gamma}[\|\psi(\theta_0) - \pi_\gamma(\psi)\|^2] \\ &\quad + \sum_{k=1}^{n-1}(n-k)\mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\psi(\theta_k) - \pi_\gamma(\psi))] \end{aligned}$$

Using the Markov property, Cauchy–Schwartz inequality, Proposition 1, and Lemma 20, we obtain

$$\begin{aligned} &\mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\psi(\theta_k) - \pi_\gamma(\psi))] && (75) \\ &\quad = \mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\mathrm{Q}_\gamma^k\psi(\theta_0) - \pi_\gamma(\psi))] && (76) \\ &\quad \overset{(a)}{\lesssim} (1/2)^{k/m(\gamma)}\,\mathrm{L}_3\,\mathsf{E}_{\pi_\gamma}\left[\|\psi(\theta_0) - \pi_\gamma(\psi)\|\int c(\theta_0,\vartheta)d\pi_\gamma(\vartheta)\right], && (77) \end{aligned}$$

where in (a) we additionally used the fact that

$$\mathbf{W}_c(\delta_{\theta_0}, \pi_\gamma) = \int c(\theta_0, \vartheta) d\pi_\gamma(\vartheta)\,.$$

Using **C**1(4), we get

$$\mathsf{E}_{\pi_\gamma}[\|\psi(\theta_0) - \pi_\gamma\|^2] \leq \mathsf{E}_{\pi_\gamma}[\|\psi(\theta_0)\|^2] \leq \mathrm{L}_3^2\, \mathsf{E}_{\pi_\gamma}[\|\theta_0 - \theta^\star\|^4] \leq \frac{\mathrm{L}_3^2\, \mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^4}{\mu^2}\,, \tag{78}$$

and, using **C**1(2) and **C**1(4), we get

$$\int\int c^2(\theta_0, \vartheta) d\pi_\gamma(\vartheta) d\pi_\gamma(\theta_0) \tag{79}$$

$$\leq \int\int \|\theta_0 - \vartheta\|^2 \left(\|\theta_0 - \theta^\star\| + \|\vartheta - \theta^\star\| + \frac{2^{3/2}\gamma^{1/2}\tau_2}{\mu^{1/2}}\right)^2 d\pi_\gamma(\vartheta) d\pi_\gamma(\theta_0) \tag{80}$$

$$\lesssim \int\int (\|\theta_0 - \theta^\star\|^4 + \|\vartheta - \theta^\star\|^4) + \frac{\gamma\tau_2^2}{\mu}(\|\theta_0 - \theta^\star\|^2 + \|\vartheta - \theta^\star\|^2) d\pi_\gamma(\vartheta) d\pi_\gamma(\theta_0) \tag{81}$$

$$\lesssim \frac{\mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^2}{\mu^2} + \frac{\mathsf{D}_{\mathrm{last},2}\gamma^2\tau_2^4}{\mu^2} \lesssim \frac{\mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^4}{\mu^2}\,. \tag{82}$$

Using (78), (79), and Cauchy–Schwartz inequality for (75), we obtain

$$\mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\psi(\theta_k) - \pi_\gamma(\psi))] \lesssim (1/2)^{k/m(\gamma)} \frac{\mathrm{L}_3\, \mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^4}{\mu^2}\,.$$

Combining the inequalities above and using that $m(\gamma) = \lceil 2\frac{\log 4}{\gamma\mu}\rceil \leq \frac{2\log 4 + 1}{\gamma\mu}$, we get

$$n^{-1}\mathsf{E}_{\pi_\gamma}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \leq \left(\frac{\mathsf{D}_{\mathrm{last},4}\, \mathrm{L}_3^2\, \gamma^2\tau_4^4}{\mu^2 n} + \frac{\mathsf{D}_{\mathrm{last},4} m(\gamma)\, \mathrm{L}_3^2\, \gamma^2\tau_4^4}{\mu^2 n}\right)^{1/2}$$
$$\lesssim \frac{\mathsf{D}_{\mathrm{last},4}\, \mathrm{L}_3\, \gamma\tau_4^2}{\mu n^{1/2}} + \frac{\mathsf{D}_{\mathrm{last},4}\, \mathrm{L}_3\, \gamma^{1/2}\tau_4^2}{\mu^{3/2} n^{1/2}}\,.$$

□

Lemma 17. *Assume **A**1, **A**2, **A**3(4). Then for any $\gamma \in (0; \frac{2}{11L}]$, and any $k \in \mathbb{N}$ it holds that*

$$\mathsf{E}[\|\theta_{k+1} - \tilde{\theta}_{k+1}\|^4 | \mathcal{F}_k] \leq (1 - \gamma\mu)^2 \|\theta_k - \tilde{\theta}_k\|^4. \tag{83}$$

Proof. Recall that the sequences $\{\theta_k\}_{k\in\mathbb{N}}$ and $\{\tilde{\theta}_k\}_{k\in\mathbb{N}}$ are defined by the recurrences

$$\theta_{k+1} = \theta_k - \gamma\nabla F(\theta_k, \xi_{k+1})\,, \quad \theta_0 = \theta \in \mathbb{R}^d\,, \tag{84}$$

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma\nabla F(\tilde{\theta}_k, \xi_{k+1})\,, \quad \tilde{\theta}_0 = \tilde{\theta} \in \mathbb{R}^d\,. \tag{85}$$

Expanding the brackets, we obtain that

$$\begin{aligned}
\|\theta_{k+1} - \tilde{\theta}_{k+1}\|^4 &= \|\theta_k - \tilde{\theta}_k\|^4 + \gamma^4\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^4 \\
&+ 4\gamma^2\langle\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle^2 \\
&+ 2\gamma^2\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^2\|\theta_k - \tilde{\theta}_k\|^2 \\
&- 4\gamma\langle\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2 \\
&- 4\gamma^3\langle\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^2
\end{aligned}$$

Using **A**3(4) and Cauchy–Schwartz inequality, we get

$$\begin{aligned}
\mathsf{E}[\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^4 | \mathcal{F}_k] &\leq \mathrm{L}^3\langle\nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2, \\
\mathsf{E}[\langle\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle^2 | \mathcal{F}_k] &\leq \mathrm{L}\langle\nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2, \\
\mathsf{E}[\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1})\|^2\|\theta_k - \theta'_k\|^2 | \mathcal{F}_k] &\leq \mathrm{L}\langle\nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2 \\
\mathsf{E}[\langle\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\tilde{\theta}_k, \xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2 | \mathcal{F}_k] &= \langle\nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle\|\theta_k - \tilde{\theta}_k\|^2\,.
\end{aligned}$$

Similarly,

$$\mathsf{E}[\langle \nabla F(\theta_k,\xi_{k+1}) - \nabla F(\tilde{\theta}_k,\xi_{k+1}), \theta_k - \tilde{\theta}_k\rangle \|\nabla F(\theta_k,\xi_{k+1}) - \nabla F(\tilde{\theta}_k,\xi_{k+1})\|^2 |\mathcal{F}_k] \\ \leq \mathrm{L}^2 \langle \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle \|\theta_k - \tilde{\theta}_k\|^2$$

Combining all inequalities above, we obtain

$$\mathsf{E}[\|\theta_{k+1} - \theta'_{k+1}\|^4 |\mathcal{F}_k] \leq \|\theta_k - \tilde{\theta}_k\|^4 \\ - (4\gamma - \gamma^4 \mathrm{L}^3 - 4\gamma^2 \mathrm{L} - 2\gamma^2 \mathrm{L} - 4\gamma^3 \mathrm{L}^2)\langle \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k), \theta_k - \tilde{\theta}_k\rangle \|\theta_k - \tilde{\theta}_k\|^2$$

Using **A**1 and since $1 - \gamma^3 \mathrm{L}^3 /4 - 3\gamma \mathrm{L} /2 - \gamma^2 \mathrm{L}^2 \geq 1 - 11\gamma \mathrm{L} /4$, we get

$$\mathsf{E}[\|\theta_{k+1} - \tilde{\theta}_{k+1}\|^4 |\mathcal{F}_k] \leq (1 - 4\gamma\mu(1 - 11\gamma \mathrm{L} /4))\|\theta_k - \tilde{\theta}_k\|^4 \\ \leq (1 - 2\gamma\mu(1 - 11\gamma \mathrm{L} /4))^2 \|\theta_k - \tilde{\theta}_k\|^4 .$$

Since $1 - 11\gamma \mathrm{L} /4 \geq 1/2$ for $\gamma \leq 2/(11 \mathrm{L})$, we complete the proof. □

Lemma 18. *Assume* ***A****1,* ***A****2,* ***A****3(4), and* ***C****1(4). Then for any $\gamma \in (0; 1/(\mathrm{L}\, \mathsf{C}_{\mathrm{step},4})]$, any $n \in \mathbb{N}$ and initial distribution ν it holds*

$$n^{-1}\mathsf{E}_{\nu}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \lesssim n^{-1}\mathsf{E}_{\pi_\gamma}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \\ + \frac{\mathrm{L}_3(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}\left(\mathsf{E}_{\nu}^{1/2}[\|\theta_0 - \theta^\star\|^4] + \frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu}\right) .$$

Proof. Using the synchronous coupling construction defined in (46) and the corresponding coupling kernel K_γ, we obtain that

$$\begin{aligned} \mathsf{E}_{\nu}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] &= (\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2])^{1/2} \\ &\leq \mathsf{E}_{\pi_\gamma}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\tilde{\theta}_k) - \pi_\gamma(\psi)\}\|^2] + (\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \psi(\tilde{\theta}_k)\}\|^2])^{1/2} \end{aligned} \tag{86}$$

Applying Minkowski's inequality to the last term and using Lemma 20, we get

$$(\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \psi(\tilde{\theta}_k)\}\|^2])^{1/2} \leq \sum_{k=n+1}^{2n} (\mathsf{E}_{\nu,\pi_\gamma}^{K}[\|\{\psi(\theta_k) - \psi(\tilde{\theta}_k)\}\|^2])^{1/2} \\ \leq \frac{\mathrm{L}_3}{2}\sum_{k=n+1}^{2n} (\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[c^2(\theta_k,\tilde{\theta}_k)])^{1/2} .$$

Using Hölder's and Minkowski's inequality and applying Lemma 17 , (64) and (65), we obtain

$$\begin{aligned} &(\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[c^2(\theta_k,\tilde{\theta}_k)])^{1/2} \\ &\quad\leq (\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\|\theta_k - \tilde{\theta}_k\|^4])^{1/4}(\mathsf{E}_{\pi_\gamma}^{1/4}[\|\tilde{\theta}_k - \theta^\star\|^4] + \mathsf{E}_{\eta}^{1/4}[\|\theta_k - \theta^\star\|^4 + \frac{\gamma^{1/2}\tau_2}{\mu^{1/2}}]) \\ &\quad\leq (1-\gamma\mu)^{k/2}(\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\|\theta_0 - \tilde{\theta}_0\|^4])^{1/4}(\mathsf{E}_{\eta}^{1/4}[\|\theta_0 - \theta^\star\|^4] + \frac{\mathsf{D}_{\mathrm{last},4}^{1/2}\gamma^{1/2}\tau_4}{\mu^{1/2}} + \frac{\gamma^{1/2}\tau_2}{\mu^{1/2}}) \\ &\quad\lesssim (1-\gamma\mu)^{k/2}\left(\frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu} + \mathsf{E}_{\nu}^{1/2}\|\theta_0 - \theta^\star\|^4\right) \end{aligned}$$

Combining all inequalities above, we get

$$(\mathsf{E}_{\nu,\pi_\gamma}^{\mathrm{K}_\gamma}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \psi(\theta'_k)\}\|^2])^{1/2} \lesssim \frac{\mathrm{L}_3(1-\gamma\mu)^{(n+1)/2}}{\gamma\mu}\left(\mathsf{E}_{\nu}^{1/2}[\|\theta_0 - \theta^\star\|^4] + \frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu}\right) .$$

Substituting the last inequality into (86) we complete the proof. □

Lemma 19. *Assume* ***A****1, **A**2, **A**3(6), and **C**1(6). Then for any $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, $n \in \mathbb{N}$, and initial distribution ν, it holds that*

$$
\begin{aligned}
n^{-1}\mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2\Big] &\le n^{-1}\mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2\Big] \\
&\quad + \frac{\mathrm{L}_4(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6] + \frac{\mathrm{L}_4\,\mathsf{D}_{\mathrm{last},6}^{3/2}\gamma^{3/2}\tau_6^3}{3\mu^{3/2}}\,.
\end{aligned} \tag{87}
$$

Proof. Applying the 4-rd order Taylor expansion with integral remainder, we get that

$$
\eta(\theta) = \psi(\theta) + \frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\,dt\right)(\theta-\theta^\star)^{\otimes 3}\,, \tag{88}
$$

and using **A**2, we obtain

$$
\|\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\,dt\right)(\theta-\theta^\star)^{\otimes 3}\| \le \mathrm{L}_4\,\|\theta-\theta^\star\|^3\,. \tag{89}
$$

Therefore, combining (88), **A**2, and applying Minkowski's inequality, we get

$$
\begin{aligned}
\mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2\Big] &\le \mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2\Big] \\
&\quad + \frac{\mathrm{L}_4}{6}\sum_{k=n+1}^{2n}\mathsf{E}_\nu^{1/2}[\|\theta_k-\theta^\star\|^6]
\end{aligned} \tag{90}
$$

Applying **C**1(6) for the last term of (90), we get

$$
\begin{aligned}
\mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2\Big] &\lesssim \mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2\Big] + \frac{\mathrm{L}_4\,n\mathsf{D}_{\mathrm{last},6}^{3/2}\gamma^{3/2}\tau_6^3}{\mu^{3/2}} \\
&\quad + \frac{\mathrm{L}_4(1-\gamma\mu)^{3(n+1)/2}}{1-(1-\gamma\mu)^{3/2}}\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6]\,.
\end{aligned} \tag{91}
$$

In remains to notice that $(1-\gamma\mu)^{3/2} \le (1-\gamma\mu)$, and the statement follows. □

We conclude this section with a technical statement on the properties of the function ψ from (68).

Lemma 20. *Let $\psi(\cdot)$ be a function defined in (68). Then for any $\theta, \theta' \in \mathbb{R}^d$, it holds that*

$$
\|\psi(\theta)-\psi(\theta')\| \le \frac{1}{2}\,\mathrm{L}_3\,c(\theta,\theta').
$$

Proof. For simplicity, let us denote $T = \nabla^3 f(\theta^*)$. Hence,

$$
\|\psi(\theta)-\psi(\theta')\| \le \frac{1}{2}\|T(\theta-\theta^\star)^{\otimes 2} - T(\theta'-\theta^\star)^{\otimes 2}\|. \tag{92}
$$

Note that

$$
\|T\| = \sup_{x\neq 0, y\neq 0, z\neq 0}\frac{\sum_{i,j,k} T_{ijk}x_iy_jz_k}{\|x\|\,\|y\|\,\|z\|} \ge \sup_{x\neq 0, y\neq 0}\sup_{z\neq 0}\frac{\sum_k z_k \sum_{i,j} T_{ijk}x_iy_j}{\|z\|\,\|y\|\,\|x\|} = \sup_{x\neq 0, y\neq 0}\frac{\|t(x,y)\|}{\|y\|\,\|x\|}, \tag{93}
$$

where $t(x,y)_k = \sum_{i,j} T_{ijk}x_iy_j$. Therefore, for any $x, y \in \mathbb{R}^d$, it holds that

$$
\|t(x,y)\| \le \|x\|\,\|y\|\,\|T\| \tag{94}
$$

We denote $v = Tx^{\otimes 2} - Ty^{\otimes 2}$. Then

$$v_k = \sum_{i,j} T_{ijk}(x_i x_j - y_i y_j) = \sum_{i,j} T_{ijk}((x_i - y_i)x_j + (x_i - y_i)y_j) = \\ \sum_{i,j} T_{ijk}(x_i - y_i)x_j + \sum_{i,j} T_{ijk}(x_i - y_i)y_j, \tag{95}$$

where the first inequality is true since $T_{ijk} = T_{jik}$ by definition of T. Combining (94) and (C) and using triangle inequality, we obtain

$$\|v\| \leq \|T\|\|x - y\|(\|x\| + \|y\|) \leq \|T\|\|x - y\|(\|x\| + \|y\| + \frac{2\sqrt{2}\tau_2\sqrt{\gamma}}{\sqrt{\mu}}).$$

We complete the proof setting $x = \theta - \theta^\star, y = \theta' - \theta^\star$ $\square$

D PROOF OF THEOREM 9

Theorem 21 (Version of Theorem 9 with explicit constants)**.** *Let $p \geq 2$ and assume **A**1, **A**2, **A**3(3p), and **C**1(3p). Then for any $\gamma \in (0; 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},3p})]$, initial distribution ν, and $n \in \mathbb{N}$, the estimator $\bar{\theta}_n^{(RR)}$ defined in (37) satisfies*

$$\mathsf{E}_\nu^{1/p}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^p] \leq \frac{c_1\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\,p^{1/2}}{n^{1/2}} + \frac{c_2 p\tau_p}{n^{1-1/p}} + \frac{\mathsf{C}_{\mathrm{RR},5}}{n\gamma^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},6}\gamma^{1/2}}{n^{1/2}} + \mathsf{C}_{\mathrm{RR},7}\gamma^{3/2} + \frac{\mathsf{C}_{\mathrm{RR},8}}{n} + \mathcal{R}_5(n,\gamma,\|\theta_0-\theta^\star\|)\,,$$

where we have set

$$\begin{aligned} \mathsf{C}_{\mathrm{RR},5} &= \frac{c_0\mathsf{D}_{\mathrm{last},p}^{1/2}\tau_p}{\mu^{1/2}}\,, \quad \mathsf{C}_{\mathrm{RR},6} = \frac{c_0\,\mathrm{L}\,\mathsf{D}_{\mathrm{last},p}^{1/2}p\tau_p}{\mu^{1/2}} + \frac{c_0\,\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}p\tau_{2p}^2}{\mu^{3/2}}\,, \\ \mathsf{C}_{\mathrm{RR},7} &= c_0\left(\mathsf{C}_1 + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},3p}^{3/2}\tau_{3p}^3}{\mu^{3/2}}\right)\,, \quad \mathsf{C}_{\mathrm{RR},8} = \frac{c_0\,\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\tau_{2p}}{\mu^2}\,, \end{aligned} \tag{96}$$

C_1 is defined in (51), and the remainder term $\mathcal{R}_5(n,\gamma,\|\theta_0-\theta^\star\|)$ is given by

$$\begin{aligned} \mathcal{R}_5(n,\gamma,\|\theta_0-\theta^\star\|) &= \frac{c_0(1-\gamma\mu)^{(n+1)/2}}{\gamma n}\mathsf{E}_\nu^{1/p}[\|\theta_0-\theta^\star\|^p] + \frac{c_0\,\mathrm{L}\,p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2}n}\mathsf{E}_\nu^{1/p}[\|\theta_0-\theta^\star\|^p] \\ &+ \frac{c_0\,\mathrm{L}(1-\gamma\mu)^{(n+1)/2}p^2}{\gamma\mu^2}\mathsf{E}_\nu^{1/p}[\|\theta_0-\theta^\star\|^{2p}] + \frac{c_0\,\mathrm{L}(1-\gamma\mu)^{(3/2)n}}{\gamma\mu}\mathsf{E}_\nu^{1/p}[\|\theta_0-\theta^\star\|^{3p}] \end{aligned} \tag{97}$$

Proof. Using the decomposition (71), we obtain that for any $p \geq 2$, it holds that

$$\begin{aligned} \mathsf{E}_\nu^{1/p}[\|\,\mathrm{H}^\star(\bar{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim\; & \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)\|^p]}_{T_1} + \underbrace{\frac{1}{\gamma n}\mathsf{E}_\nu^{1/p}[\|\theta_{n+1}^{(\gamma)}-\theta^\star\|^p] + \frac{1}{\gamma n}\mathsf{E}_\nu^{1/p}[\|\theta_{2n}^{(\gamma)}-\theta^\star\|^p]}_{T_2} \\ & + \underbrace{\frac{1}{\gamma n}\mathsf{E}_\nu^{1/p}[\|\theta_{n+1}^{(2\gamma)}-\theta^\star\|^p] + \frac{1}{\gamma n}\mathsf{E}_\nu^{1/p}[\|\theta_{2n}^{(2\gamma)}-\theta^\star\|^p]}_{T_3} \\ & + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta^\star)\|^p]}_{T_4} \\ & + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta_k^{(2\gamma)}) - \varepsilon_{k+1}(\theta^\star)\|^p]}_{T_5} + \underbrace{\|2\pi_\gamma(\psi) - \pi_{2\gamma}(\psi)\|}_{T_6} \\ & + \underbrace{\frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\psi(\theta_k^{(\gamma)}) - \pi_\gamma(\psi)\|^p] + \frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\psi(\theta_k^{(2\gamma)}) - \pi_{2\gamma}(\psi)\|^p]}_{T_7} \\ & + \underbrace{\frac{1}{n}\sum_{k=n+1}^{2n}\mathsf{E}_\nu^{1/p}[\|G(\theta_k^{(\gamma)})\|^p] + \frac{1}{n}\sum_{k=n+1}^{2n}\mathsf{E}_\nu^{1/p}[\|G(\theta_k^{(2\gamma)})\|^p]}_{T_8}\,. \end{aligned}$$

Now we upper bounds the terms above separately. Applying first the Pinelis version of Rosenthal inequality (Pinelis, 1994) together with **A**3(p), we obtain that

$$T_1 \leq \frac{c_1\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\,p^{1/2}}{n^{1/2}} + \frac{c_2 p\tau_p}{n^{1-1/p}}\,.$$

Applying **C**1(p) (which is implied by **C**1($3p$)), we obtain that

$$T_2 + T_3 \lesssim \frac{\mathsf{D}_{\mathrm{last},p}^{1/2}\tau_p}{\mu^{1/2} n \gamma^{1/2}} + \frac{(1-\gamma\mu)^{(n+1)/2}}{\gamma n} \mathsf{E}_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^p\big]\,.$$

Applying Lemma 14 (see the bound (67)), we get that

$$T_4 + T_5 \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},p}^{1/2}\gamma^{1/2} p \tau_p}{\mu^{1/2} n^{1/2}} + \frac{\mathrm{L}\,p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2} n} \mathsf{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p]\,.$$

Using the bounds (72) and (73), we obtain

$$T_6 \lesssim \mathsf{C}_1 \gamma^{3/2}\,.$$

Applying Proposition 8, we get

$$\frac{1}{n}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n} \psi(\theta_k^{(\gamma)}) - \pi_\gamma(\psi)\|^p] \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p} p \tau_{2p}^2 \gamma^{1/2}}{\mu^{3/2} n^{1/2}} + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\tau_{2p}}{\mu^2 n}\,.$$

Using this bound and adopting the result of (Durmus et al., 2023, Theorem 4), we obtain that

$$T_7 \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p} p \tau_{2p}^2 \gamma^{1/2}}{\mu^{3/2} n^{1/2}} + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\tau_{2p}}{\mu^2 n} + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2} p^2}{\gamma\mu^2} \mathsf{E}_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^{2p}\big]\,.$$

Finally, applying the definition of $G(\theta)$ in (38) together with **C**1($3p$), we obtain that

$$\begin{aligned} T_8 &\lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},3p}^{3/2}\gamma^{3/2}\tau_{3p}^3}{\mu^{3/2}} + \frac{\mathrm{L}}{n}\sum_{k=n+1}^{2n}(1-\gamma\mu)^{(3/2)k}\mathsf{E}_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^{3p}\big] \\ &\lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},3p}^{3/2}\gamma^{3/2}\tau_{3p}^3}{\mu^{3/2}} + \frac{\mathrm{L}(1-\gamma\mu)^{(3/2)n}}{\gamma\mu}\mathsf{E}_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^{3p}\big]\,. \end{aligned}$$

To complete the proof it remains to combine the bounds for T_1 to T_8. □

D.1 Proof of Proposition 8

In the proof below we use the notation

$$\bar{\psi}(\theta) = \psi(\theta) - \pi_\gamma(\psi)\,.$$

We proceed with the blocking technique. Indeed, let us set the parameter

$$m = m(\gamma) = \left\lceil \frac{2\log 4}{\gamma\mu} \right\rceil\,. \tag{98}$$

Our choice of parameter $m(\gamma)$ is due to Proposition 1. For notation conciseness we write it simply as m, dropping its dependence upon γ. Using Minkowski's inequality, we obtain that

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=0}^{n-1}\bar{\psi}(\theta_k)\|^p\big] \le \mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=0}^{\lfloor n/m\rfloor m-1}\bar{\psi}(\theta_k)\|^p\big] + m\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\bar{\psi}(\theta_0)\|^p\big]\,. \tag{99}$$

Now we consider the Poisson equation, associated with Q_γ^m and function $\bar{\psi}$, that is,

$$g_m(\theta) - \mathrm{Q}_\gamma^m g_m(\theta) = \bar{\psi}(\theta)\,. \tag{100}$$

The function

$$g_m(\theta) = \sum_{k=0}^{\infty}\mathrm{Q}_\gamma^{km}\bar{\psi}(\theta) \tag{101}$$

is well-defined under the assumptions **A**1, **A**2, **A**3($2p$), and **C**1($2p$). Moreover, g_m is a solution of the Poisson equation (100). Define $q := \lfloor n/m \rfloor$, then we have

$$\sum_{k=0}^{qm-1}\bar{\psi}(\theta_k) = \sum_{r=0}^{m-1} B_{m,r}\,, \quad \text{with} \quad B_{m,r} = \sum_{k=0}^{q-1}\big\{g_m(\theta_{km+r}) - \mathrm{Q}_\gamma^m g_m(\theta_{km+r})\big\}\,. \tag{102}$$

Using Minkowski's inequality, we get from (99), that

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=0}^{n-1}\bar{\psi}(\theta_k)\|^p\big] \le m\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=1}^{q}\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^p\big] + 2m\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\psi(\theta_0\|^p\big] \tag{103}$$

Now we upper bound both terms of (103) separately. Under assumption **A**2, and applying **C**1($2p$), we get

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\psi(\theta_0\|^p\big] \le \frac{\mathrm{L}}{2}\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\theta_0-\theta^\star\|^{2p}\big] \le \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\gamma\tau_{2p}^2}{2\mu}\,. \tag{104}$$

To proceed with the first term, we apply Burkholder's inequality (Osekowski, 2012, Theorem 8.6), and obtain that

$$\begin{aligned}\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=1}^{q}\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^p\big]&\\ &\le p\mathsf{E}_{\pi_\gamma}^{1/p}\big[\big(\sum_{k=1}^{q}\|\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^2\big)^{p/2}\big]\,. \end{aligned}\tag{105}$$

Applying now Minkowski's inequality again, we get

$$\begin{aligned}\mathsf{E}_{\pi_\gamma}^{2/p}\big[\big(\sum_{k=1}^{q}\|\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^2\big)^{p/2}\big] &\le q\mathsf{E}_{\pi_\gamma}^{2/p}\big[\|\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^p\big]\\ &\lesssim q\left(\mathsf{E}_{\pi_\gamma}^{2/p}[\|g_m(\theta_0)\|^p] + \mathsf{E}_{\pi_\gamma}^{2/p}[\|\mathrm{Q}_\gamma^m g_m(\theta_0)\|^p]\right)\\ &\lesssim q\mathsf{E}_{\pi_\gamma}^{2/p}[\|g_m(\theta_0)\|^p]\,.\end{aligned}$$

It remains to upper bound the moment $\mathsf{E}_{\pi_\gamma}^{2/p}[\|g_m(\theta_0)\|^p]$. In order to do this, we first note that due to the duality theorem (Douc et al., 2018, Theorem 20.1.2.), we get that for any $k \in \mathbb{N}$,

$$\|\mathrm{Q}^{mk}\psi(\theta) - \pi_\gamma(\psi)\| \le \frac{1}{2}\,\mathrm{L}_3\,\mathbf{W}_c(\delta_\theta\mathrm{Q}_\gamma^{km},\pi_\gamma) \le 2\,\mathrm{L}_3(1/2)^k\mathbf{W}_c(\delta_\theta,\pi_\gamma)\,,$$

where the last inequality is due to Proposition 1. Hence, applying the definition of $g_m(\theta)$ in (101), we obtain that

$$\mathsf{E}_{\pi_\gamma}^{1/p}[\|g_m(\theta_0)\|^p] \le \sum_{k=0}^{\infty}\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\mathrm{Q}_\gamma^{km}\bar{\psi}(\theta)\|^p\big] \le 2\,\mathrm{L}_3\sum_{k=0}^{\infty}(1/2)^k\mathsf{E}_{\pi_\gamma}^{1/p}\big[\{\mathbf{W}_c(\delta_\theta,\pi_\gamma)\}^p\big]\,.$$

To control the latter term, we simply apply the definition of $\mathbf{W}_c(\delta_\theta,\pi_\gamma)$ and a cost function $c(\theta,\theta')$ together with **C**1($2p$), we get

$$\begin{aligned}\mathsf{E}_{\pi_\gamma}^{1/p}\big[\{\mathbf{W}_c(\delta_\theta,\pi_\gamma)\}^p\big] &\lesssim \left(\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\theta-\theta'\|^p\left(\|\theta-\theta^\star\| + \|\theta'-\theta^\star\| + \frac{\tau_2\sqrt{\gamma}}{\sqrt{\mu}}\right)^p\pi_\gamma(\mathrm{d}\theta)\pi_\gamma(\mathrm{d}\theta')\right)^{1/p}\\ &\le \big(\int\|\theta-\theta'\|^{2p}\pi_\gamma(\mathrm{d}\theta)\pi_\gamma(\mathrm{d}\theta')\big)^{1/2p}\big(\int\left(\|\theta-\theta^\star\| + \|\theta'-\theta^\star\| + \frac{\tau_2\sqrt{\gamma}}{\sqrt{\mu}}\right)^{2p}\pi_\gamma(\mathrm{d}\theta)\pi_\gamma(\mathrm{d}\theta')\big)^{1/2p}\\ &\lesssim \frac{\mathsf{D}_{\mathrm{last},2p}\tau_{2p}^2\gamma}{\mu}\,.\end{aligned}$$

Combining now the bounds above in (105), we get that

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=1}^{q}\{g_m(\theta_{km}) - \mathrm{Q}_\gamma^m g_m(\theta_{(k-1)m})\}\|^p\big] \lesssim \frac{\mathsf{D}_{\mathrm{last},2p}p\,\mathrm{L}_3\,\tau_{2p}^2\gamma\sqrt{q}}{\mu}\,, \tag{106}$$

and, hence, substituting into (99), we get

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\|\sum_{k=0}^{n-1}\bar{\psi}(\theta_k)\|^p\big] \lesssim \frac{\mathsf{D}_{\mathrm{last},2p}p\,\mathrm{L}_3\,\tau_{2p}^2\gamma\sqrt{q}m}{\mu} + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\tau_{2p}^2\gamma m}{2\mu}\,. \tag{107}$$

Now the statement follows from the definition of $m = m(\gamma)$ in (98) and $q = \lfloor n/m\rfloor \le n/m$.