Nonasymptotic Bounds and Enhanced Convergence of Stochastic Gradient Descent via Richardson–Romberg Extrapolation

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 8, 2024

Abstract

Stochastic gradient descent (SGD) remains one of the most widely used methods for solving large-scale optimization problems in machine learning. In this paper, we present a novel analysis of SGD combined with the Richardson-Romberg extrapolation technique. This modification allows for sharper convergence rates in strongly convex optimization settings, achieving optimal root-mean-square error bounds with reduced bias. We further extend the results to the p-th moment error bounds, providing generalizations to multiple scenarios. By viewing the iterates of SGD as a Markov chain, we leverage geometric ergodicity properties to analyze the error terms and derive explicit moment bounds. Our approach shows that the Richardson-Romberg extrapolation offers a significant improvement in the accuracy of SGD, particularly for constant step-size implementations.

Keywords: Stochastic Gradient Descent (SGD), Richardson-Romberg Extrapolation, Strong Convexity, Nonasymptotic Bounds, Moment Analysis.

1 Introduction

Stochastic gradient descent (SGD) is a fundamental optimization algorithm extensively used in machine learning, reinforcement learning, and statistical estimation. Its simplicity and computational efficiency make it suitable for large-scale problems. Despite its popularity, achieving optimal convergence rates in practical settings remains challenging, particularly when noisy gradients and finite-sample complexities are involved. Numerous techniques have been introduced to enhance the performance of SGD, including momentum methods, variance reduction, and adaptive learning rates. However, there is still significant interest in improving the theoretical convergence rates of SGD without overly complex modifications.

In this work, we focus on combining SGD with the Richardson-Romberg extrapolation technique, which has traditionally been applied in numerical analysis to accelerate the convergence of approximations. We investigate how this technique can be used in the context of strongly convex optimization problems to improve the convergence of SGD, particularly in terms of nonasymptotic error bounds.

Our approach offers two major benefits: first, it reduces the bias in the error estimates of the SGD iterates by leveraging multiple gradient computations with different step sizes; second, it achieves optimal moment bounds, including second- and p-th moment error bounds, under a constant step-size setting. This modification can be viewed as a lightweight enhancement to standard SGD, making it practical for real-world applications where precision and fast convergence are crucial.

Backgrounds. Stochastic gradient methods are a fundamental approach for solving a wide range of optimization problems, with a broad range of applications including generative modeling [GPAM$^+$14, GBC16], empirical risk minimization [VdV00], and reinforcement learning [SB18, SLA$^+$15, MKS$^+$15]. 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 [RM51]. 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 [Qia99], averaging [PJ92], or variance reduction [DBLJ14, NLST17]. In particular, the celebrated Polyak-Ruppert algorithm proceeds with a trajectory-wise averaging of the estimates

$$\overline{\theta}_{n_0,n} = \frac{1}{n}\sum_{k=n_0+1}^{n+n_0}\theta_k \tag{3}$$

for some $n_0 > 0$. It is known [PJ92, For15], that under appropriate assumptions on f and γ_k, the sequence of estimates $\{\overline{\theta}_{n_0,n}\}_{n\in\mathbb{N}}$ is asymptotically normal, that is,

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

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 [For15] for a discussion. On the other hand, quantitative counterparts of (4) rely on the mean-squared error bounds of the form

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

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 [BM13], [GP23]. In particular, [LMWJ22] derived the bound (5) with the optimal rate $\delta = 1/2$ and proved the corresponding lower bound. 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 $\overline{\theta}_{n_0,n}$ based on Richardson-Romberg extrapolation. The main results 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 an 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 2). 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.

- 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

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

where C is a universal constant, and $\overline{\theta}_n^{(RR)}$ is a counterpart of the quantity $\overline{\theta}_{n_0,n}$ when using Richardson-Romberg extrapolation, see related definitions at Section 3. Our proof is based on a novel version of the Rosenthal inequality, which might be of independent interest.

1.1 Literature review

Richardson-Romberg extrapolation. Richardson-Romberg extrapolation is a technique used to improve the accuracy of numerical approximations [Hil87], 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 [TT90], and later generalized for non-smooth functions in [BT96]. This technique was extended using multistep discretizations in [Pag07]. Finally, Richardson-Romberg extrapolation have been applied to Stochastic Gradient Descent (SGD) methods in [DSM$^+$16], [MG23] and [HCX24], to improve convergence and reduce error in optimization problems, particularly when dealing with noisy or high-variance gradient estimates.

SGD with Polyak-Ruppert averaging. [GP23] derive (5) with $\delta = 1/4$ for a certain class of functions f, including strongly convex functions, improving [MB11] which obtain this result for $\delta = 1/6$. [LMWJ22] suggested the Root-SGD algorithm combining the ideas of the Polyak-Ruppert averaged SGD with control variates and established (5) with $\delta = 1/2$. The recent series of papers [HCX23, ZX24, ZHCX24] 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.

Contributions. Our main contributions are as follows:

- We introduce a novel modification to the standard SGD algorithm by incorporating the Richardson-Romberg extrapolation, and show how this leads to improved convergence rates.
- We derive nonasymptotic upper bounds on the root-mean-square error (RMSE) and p-th moment error in the strongly convex setting.
- Our analysis provides new insights into the ergodicity of the SGD iterates, leveraging their Markov chain properties to sharpen the error bounds.

Organization. The rest of the paper is organized as follows. In Section 1.1, we provide a literature review on non-asymptotic analysis of first-order optimization methods, with a focus on constant step-size algorithms and the Richardson-Romberg procedure. In Section 2, we analyze constant step-size SGD as a Markov chain and examine the properties of the Polyak-Ruppert averaged estimator (3). Section 3 discusses the Richardson-Romberg extrapolation applied to Polyak-Ruppert averaged SGD and derives the RMSE and p-th moment error bounds. Finally, Section 4 concludes with extensions, practical implications, and directions for future research.

Notations and definitions. For $\theta_1, \dots, \theta_k$ being the iterates of stochastic first-order method, we denote $\mathcal{F}_k = \sigma(\theta_0, \theta_1, \dots, \theta_k)$ and E_k be an alias for $\mathsf{E}\big[\cdot|\mathcal{F}_k\big]$. 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 $(d(x, y) \wedge 1)^q \le c(x, y)$. 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 $c_p(x, y) = \|x - y\|^p$, $x, y \in \mathbb{R}^d$, $p \ge 1$, 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

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

see [DMPS18, 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 $(\nabla^k f(\theta) M)_l = \sum_{i_1, \dots, i_{k-1}} M_{i_1, \cdots, i_{k-1}} \nabla^k f(\theta)_{i_1, \cdots, i_{k-1}, l}$, where $l \in \{1, \dots, d\}$.

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 $a_n \le c_0 b_n$.

2 Finite-time analysis of the SGD dynamics for strongly convex minimization problems: Geometric ergodicity of SGD iterates

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

Assumption 1. *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*

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

Assumption 2. *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

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

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

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma \nabla F(\theta_k^{(\gamma)}, \xi_{k+1}) \quad \theta_0^{(\gamma)} = \theta_0 \sim \nu \tag{11}$$

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

Assumption 3 (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 $\mathsf{E}^{1/p}[\|\nabla F(\theta^\star, \xi)\|^p] \leq \tau_p$, and for any $q = 2, \ldots, p$ it holds with some $\mathrm{L}_1 > 0$ that for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

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

Assumption Assumption 3(p) generalizes the well-known L_1-co-coercivity assumption, see [DDB20]. A sufficient condition which allows for Assumption 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 $n \geq 0$,

$$\overline{\theta}_n^{(\gamma)} = \frac{1}{n} \sum_{k=n+1}^{2n} \theta_k^{(\gamma)} \tag{14}$$

Many previous studies instead consider $\overline{\vartheta}_n^{(\gamma)} = \frac{1}{n-n_0} \sum_{k=n_0+1}^{n} \theta_k^{(\gamma)}$ rather than $\overline{\theta}_n^{(\gamma)}$, where $n \geq n_0 + 1$ 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 $\overline{\theta}_n^{(\gamma)} - \theta^\star$ 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 Assumption 1, Assumption 2 and Assumption 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 [DDB20] 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 [DDB20] or [MG23], 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 Assumption 1 and Assumption 3(2) as follows:

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

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

Proposition 1. *Assume Assumption 1, Assumption 2, and Assumption 3(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, 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. [DDB20]. It particular, it is known [DDB20, 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:

$$\int_{\mathbb{R}^d}\|\theta-\theta^\star\|^2\pi_\gamma(\mathrm{d}\theta) \le \frac{\mathsf{D}_{\mathrm{last},2}\gamma\tau_2}{\mu} \tag{18}$$

This property yields, using Lyapunov's inequality, that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\theta-\theta'\|\pi_\gamma(\mathrm{d}\theta)\pi_\gamma(\mathrm{d}\theta') \lesssim \sqrt{\frac{\mathsf{D}_{\mathrm{last},2}\gamma\tau_2}{\mu}}$$

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

$$\int_{\mathbb{R}^d} c(\theta,\theta')\pi(\mathrm{d}\theta)\pi(\mathrm{d}\theta') \lesssim \frac{\mathsf{D}_{\mathrm{last},2}\gamma\tau_2}{\mu} \tag{19}$$

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 4.

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.

C 1 (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 Assumption 1 - Assumption 3(p). In particular, [DDB20, Lemma 13] implies that, under Assumption 1, Assumption 3(p) with $p \geq 2$, and Assumption 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 [GP23, LMWJ22, MG23] 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 Assumption 1, Assumption 3(p) with $p \geq 2$, and Assumption 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 [DDB20, Theorem 4]. Namely, the following result holds:

Proposition 2. *Assume Assumption 1, Assumption 2, Assumption 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*

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

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

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* (45) *is a constant independent of* γ*. Moreover, for any initial distribution* ν *on* $\mathbb{R}^d$*, it holds that*

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

where

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

The proof is postponed to Appendix A. Results of this type were already obtained in the literature for stochastic approximation algorithms, see e.g. [HZCX24] and [AG24]. As already highlighted, the additive term Δ_1 vanishes in the case of minimizing the quadratic function f, see [BM13].

3 Richardson-Romberg extrapolation

Our analysis will be based on the summation by parts formula

$$\mathrm{H}^\star(\overline{\theta}_n^{(\gamma)} - \theta^\star) = \frac{\theta_{n+1}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{\theta_{2n}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{1}{n}\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta_k^\gamma) - \frac{1}{n}\sum_{k=n+1}^{2n} \eta(\theta_k^\gamma) \tag{26}$$

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

$$\int_{\mathbb{R}^d} \eta(\theta)\pi_\gamma(\mathrm{d}\theta) \neq 0 \tag{27}$$

which prevents us from using more aggressive (larger) step sizes γ in the optimized bound presented in [MB11]:

$$\mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\overline{\theta}_n^{(\gamma)} - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\big(n^{-2/3}\big) \tag{28}$$

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 (28). Instead of considering a single SGD trajectory $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$, and then relying on the tail-averaged estimator $\overline{\theta}_n^{(\gamma)}$, we construct two parallel chains based on the same sequence of noise variables $\{\xi_k\}_{k\in\mathbb{N}}$:

(29)

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

$$\overline{\theta}_n^{(RR)} := 2\overline{\theta}_n^{(\gamma)} - \overline{\theta}_n^{(2\gamma)} \tag{30}$$

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 (29). Indeed, consider the decomposition (26) and further expand the term $\eta(\theta)$ defined in

$$\eta(\theta) = \nabla f(\theta) - \mathrm{H}^\star(\theta - \theta^\star) \qquad \mathrm{H}^\star = \nabla^2 f(\theta^\star) \in \mathbb{R}^{d\times d} \tag{31}$$

so

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

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

$$\begin{aligned} \psi(\theta) &= (1/2)\nabla^3 f(\theta^*)(\theta - \theta^\star)^{\otimes 2} \\ G(\theta) &= (1/6)\left(\textstyle\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\,\mathrm{d}t\right)(\theta - \theta^\star)^{\otimes 3} \end{aligned} \tag{32}$$

We further rewrite the decomposition (26) as

$$\mathrm{H}^\star(\overline{\theta}_n^{(\gamma)} - \theta^\star) = \frac{\theta_{n+1}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{\theta_{2n}^{(\gamma)} - \theta^\star}{\gamma n} - \frac{1}{n}\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta^\star)$$

$$- \frac{1}{n} \sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta^\star)\} - \frac{1}{n} \sum_{k=n+1}^{2n} \psi(\theta_k^{(\gamma)}) - \frac{1}{n} \sum_{k=n+1}^{2n} G(\theta_k^{(\gamma)}) \quad (33)$$

Note that in the decomposition (33), 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 (33) are small (see Theorem 2 for more details). Hence, using the same sequence $\{\xi_k\}_{k\in\mathbb{N}}$ of noise variables in (29) yields an estimator $\overline{\theta}_n^{(RR)}$, 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 $\overline{\theta}_n^{(\gamma)}$ and $\overline{\theta}_n^{(2\gamma)}$ increase the leading component of the MSE by a constant factor. Hence, it is preferable to use synchronous noise construction as introduced in (29).

Proposition 2 implies the following improved bound on the bias of $\overline{\theta}_n^{(RR)}$:

Proposition 3. *Assume Assumption 1, Assumption 2, Assumption 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*

$$\mathsf{E}_\nu[\overline{\theta}_n^{(RR)}] = \theta^\star + B_3\gamma^{3/2} + \mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n) \quad (34)$$

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

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

The proof of Proposition 3 is provided in Appendix A. This result is a simple consequence of Proposition 3, since the linear in γ component of the bias $\gamma\Delta_1$ from (24) cancels out when computing $\overline{\theta}_n^{(RR)}$. We are now ready to formulate the main result for the Richardson-Romberg estimate $\overline{\theta}_n^{(RR)}$.

Theorem 1. *Assume Assumption 1, Assumption 2, Assumption 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 $\overline{\theta}_n^{(RR)}$ defined in (30) satisfies*

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

where the constants $\mathsf{C}_{\mathrm{RR},1}$ to $\mathsf{C}_{\mathrm{RR},4}$ are defined in (58) and

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

Proof of Theorem 1 is provided in Appendix B. We can optimize the above bound setting γ depending upon n.

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

$$\mathsf{E}_\nu^{1/2}[\|\,\mathrm{H}^\star(\overline{\theta}_n^{(RR)} - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\big(n^{-3/4}\big) + \mathcal{R}_4(n, 1/\sqrt{n}, \|\theta_0 - \theta^\star\|) \quad (35)$$

Discussion. Note that the result of Corollary 1 is a counterpart of (5) with $\delta = 1/4$. However, the assumptions of Theorem 1 are stronger compared to the ones imposed by [LMWJ22]. In particular, in Assumption 2 we require that f is 4 times continuously differentiable and uniformly bounded. At the same time, [LMWJ22] impose Lipschitz continuity of the Hessian of f, which is essentially equivalent to bounded 3-rd derivative of f. Our proof of Theorem 1 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 $p \geq 2$. 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 [Ros70, Pin94].

Proposition 4. *Let $p \geq 2$ and assume Assumption 1, Assumption 2, Assumption 3(2p), and **C** 1(2p). Then for any $\gamma \in (0, 1/(\mathrm{L}\, \mathsf{C}_{\mathrm{step},2p})]$, it holds that*

$$\mathsf{E}_{\pi_\gamma}^{1/p}\big[\| \textstyle\sum_{k=0}^{n-1}\{\psi(\theta_k^{(\gamma)}) - \pi_\gamma(\psi)\|^p\big] \lesssim \frac{\mathrm{L}\, \mathsf{D}_{\mathrm{last},2p} p \tau_{2p}^2 \sqrt{n\gamma}}{\mu^{3/2}} + \frac{\mathrm{L}\, \mathsf{D}_{\mathrm{last},2p} \tau_{2p}}{\mu^2} \tag{36}$$

where ψ is defined in (32).

Discussion. The proof of Proposition 4 is provided in Appendix C.1. It is important to acknowledge that there are numerous Rosenthal-type inequalities for dependent sequences in the literature. Proposition 4 can be viewed as an analogue to the classical Rosenthal inequality for strongly mixing sequences, see [Rio17, 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 Assumption 1-Assumption 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 [MPR11]. However, the considered function $\psi(\theta)$ is quadratic, which makes the respective result of [MPR11, Theorem 1] inapplicable. Similar Rosenthal-type inequalities have been explored in [DMN$^+$23], but in Proposition 4 we obtain the bound with tighter dependence of the right-hand side upon γ.

Below we provide the p-th moment bound together with its corollary for the step size γ optimized with respect to n.

Theorem 2. *Let $p \geq 2$ and assume Assumption 1, Assumption 2, Assumption 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 $\overline{\theta}_n^{(RR)}$ defined in* (30) *satisfies*

$$\begin{aligned}\mathsf{E}_\nu^{1/p}[\| \mathrm{H}^\star(\overline{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\, p^{1/2}}{n^{1/2}} + \frac{p\tau_p}{n^{1-1/p}} + \frac{\mathsf{C}_{\mathrm{RR},5}}{n\gamma^{1/2}} + \frac{\mathsf{C}_{\mathrm{RR},6}\gamma^{1/2}}{n^{1/2}} + \mathsf{C}_{\mathrm{RR},7}\gamma^{3/2} \\ + \frac{\mathsf{C}_{\mathrm{RR},8}}{n} + \mathcal{R}_5(n, \gamma, \|\theta_0 - \theta^\star\|)\end{aligned} \tag{37}$$

where constants $\mathsf{C}_{\mathrm{RR},5}$ to $\mathsf{C}_{\mathrm{RR},8}$ are defined in (86) *and*

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

where the constant $C_{f,p}$ can be traced from (87).

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

$$\mathsf{E}_\nu^{1/p}[\| \mathrm{H}^\star(\overline{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}\, p^{1/2}}{n^{1/2}} + \mathcal{O}\big(n^{-3/4}\big) + \mathcal{R}_5(n, 1/\sqrt{n}, \|\theta_0 - \theta^\star\|) \tag{38}$$

Discussion. Proof of Theorem 2 is provided in Appendix C. Note that the result above is a direct generalization of Theorem 1, 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 [DMNS24]. 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 (38). This result can be found, for example, in [DMNS24, Theorem 1]. Hence, the main result of Theorem 2 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.

4 Conclusion

In this paper, we tackled the problem of solving strongly convex and smooth minimization problems using stochastic gradient descent (SGD) with a constant step size, enhanced by Polyak-Ruppert averaging and the Richardson-Romberg extrapolation technique. We significantly extended previous results by providing a detailed nonasymptotic analysis of the mean-squared error (MSE) of the estimator. Specifically, we showed that the MSE can be decomposed into a leading term of order $\mathcal{O}(n^{-1/2})$, which depends on the minimax-optimal asymptotic covariance matrix, and a second-order term of order $\mathcal{O}(n^{-3/4})$, which is generally non-improvable. Furthermore, we extended these results to p-th moment bounds while maintaining optimal scaling of the remainders with respect to n.

Our analysis relied on viewing the SGD iterates as a time-homogeneous Markov chain, demonstrating that this chain is geometrically ergodic with respect to a carefully defined weighted Wasserstein semimetric. Directions for future research include generalizing our results to settings involving dependent noise sequences in the stochastic gradients, and studying the properties of the Richardson-Romberg extrapolated estimator under relaxed smoothness and convexity assumptions. Additionally, further exploration of the relationship between the parameter δ and rates in Berry-Esseen type results could provide deeper insights into the statistical properties of the algorithm.

References

[AG24] Sebastian Allmeier and Nicolas Gast. Computing the bias of constant-step stochastic approximation with markovian noise. *arXiv preprint arXiv:2405.14285*, 2024.

[BM13] 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, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[BT96] 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.

[DBLJ14] 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.

[DDB20] 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.

[DMN+23] 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.

[DMNS24] 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.

[DMPS18] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.

[DSM+16] 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.

[For15] Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.

[GBC16] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

[GP23] 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.

[GPAM+14] 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.

[HCX23] 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*, pages 81–82, 2023.

[HCX24] Dongyan Lucy Huo, Yudong Chen, and Qiaomin Xie. Effectiveness of constant stepsize in markovian lsa and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20447–20455, 2024.

[Hil87] Francis Begnaud Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

[HZCX24] 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*, 2024.

[LMWJ22] 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, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 909–981. PMLR, 02–05 Jul 2022.

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[MG23] Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant step-size sgd through markov chains. *arXiv preprint arXiv:2306.11497*, 2023.

[MKS+15] 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.

[MPR11] 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.

[NLST17] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

[Ose12] A. Osekowski. *Sharp Martingale and Semimartingale Inequalities*. Monografie Matematyczne 72. Birkhäuser Basel, 1 edition, 2012.

[Pag07] Gilles Pagès. Multi-step Richardson-Romberg Extrapolation: Remarks on Variance Control and Complexity. *Monte Carlo Methods and Applications*, 13(1):37–70, 2007.

[Pin94] Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994.

[PJ92] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[Rio17] Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80. 2017.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Ros70] 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.

[SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[SLA+15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[TT90] 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.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[ZHCX24] 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.

[ZX24] Yixuan Zhang and Qiaomin Xie. Constant stepsize q-learning: Distributional convergence, bias and extrapolation. *arXiv preprint arXiv:2401.13884*, 2024.

A Proof of Proposition 2 and Proposition 3

A.1 Proof of Proposition 1

Consider the synchronous coupling construction defined by the recursions

$$\begin{aligned} \theta_{k+1}^{(\gamma)} &= \theta_k^{(\gamma)} - \gamma \nabla F(\theta_k^{(\gamma)}, \xi_{k+1}) \quad \theta_0^{(\gamma)} = \theta \in \mathbb{R}^d \\ \widetilde{\theta}_{k+1}^{(\gamma)} &= \widetilde{\theta}_k^{(\gamma)} - \gamma \nabla F(\widetilde{\theta}_k^{(\gamma)}, \xi_{k+1}) \quad \widetilde{\theta}_0^{(\gamma)} = \widetilde{\theta} \in \mathbb{R}^d \end{aligned} \tag{39}$$

The pair $(\theta_k^{(\gamma)}, \widetilde{\theta}_k^{(\gamma)})_{k\in\mathbb{N}}$ 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 $(\theta_k, \widetilde{\theta}_k)_{k\in\mathbb{N}}$. Applying now Assumption 3(2), for $\gamma \le 2/\mathrm{L}$, we get that

$$\begin{aligned} \mathsf{E}_\nu[\|\theta_{k+1} - \widetilde{\theta}_{k+1}\|^2|\mathcal{F}_k] &= \mathsf{E}[\|\theta_k - \widetilde{\theta}_k - \gamma(\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\widetilde{\theta}_k, \xi_{k+1}))\|^2|\mathcal{F}_k] \\ &= \|\theta_k - \widetilde{\theta}_k\|^2 + \gamma^2 \mathsf{E}[\|\nabla F(\theta_k, \xi_{k+1}) - \nabla F(\widetilde{\theta}_k, \xi_{k+1})\|^2|\mathcal{F}_k] \\ &\qquad - 2\gamma\langle \nabla f(\theta_k) - \nabla f(\widetilde{\theta}_k), \theta_k - \widetilde{\theta}_k\rangle \\ &\le (1-\gamma\mu)\|\theta_k - \widetilde{\theta}_k\|^2, \end{aligned} \tag{40}$$

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

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

Combining the above inequalities and using standard SGD analysis

$$\mathsf{E}_\nu\|\theta_k - \theta^\star\|^2 \lesssim (1-\gamma\mu)^k \mathsf{E}_\nu\big[\|\theta_0 - \theta^\star\|^2\big] + \frac{\mathsf{D}_{\mathrm{last},2}\gamma\tau_2^2}{\mu} \tag{41}$$

we obtain

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

Note that $2(1-\gamma\mu)^{r/2} \le 2$ for any $r \le m(\gamma) - 1$ and $2(1-\gamma\mu)^{m(\gamma)/2} \le 1/2$. Hence, applying the theorem [DMPS18, Theorem 20.3.4], we obtain that the Markov kernel Q_γ admits a unique invariant distribution π_γ. Moreover,

$$\mathbf{W}_c(\nu\mathrm{Q}_\gamma^k, \pi_\gamma) \le 2(1/2)^{\lfloor k/m(\gamma)\rfloor}\mathbf{W}_c(\nu, \pi_\gamma) \tag{42}$$

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

A.2 Proof of Proposition 2

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

$$\mathcal{C}(\theta) = \mathsf{E}[\varepsilon_1(\theta)^{\otimes 2}] \tag{43}$$

The result below is essentially based on an appropriate modification of the bounds presented in [DDB20, Lemma 18]. A careful inspection of the respective proof reveals that we do not need specific assumptions for $\mathcal{C}(\theta)$ defined in (43), instead we use Lemma 2. For completeness, we present the respective result below:

Lemma 1. *Assume Assumption 1, Assumption 2, Assumption 3(6), and C1(6). Then, for any $\gamma \in (0, 1/(\mathrm{L}\,\mathsf{C}_{\mathrm{step},6})]$, it holds*

$$\overline{\theta}_\gamma - \theta^\star = -(\gamma/2)\{\mathrm{H}^\star\}^{-1}\{\nabla^3 f(\theta^\star)\}\mathbf{T}\mathcal{C}(\theta^\star) + B_1\gamma^{3/2} \tag{44}$$

where $\overline{\theta}_\gamma$ is defined in (22), $\mathcal{C}(\theta)$ is defined in (43), and $B_1 \in \mathbb{R}^d$ satisfies $\|B_1\| \le \mathsf{C}_1$, where

$$\mathsf{C}_1 = \left(\frac{\mathrm{L}^2\,\mathsf{D}_{\mathrm{last},2}}{\sqrt{\mu}} + \frac{\mathrm{L}\,\sqrt{\mathsf{D}_{\mathrm{last},2}}}{\sqrt{\mu}}\right)\frac{\tau_2^2}{\mu} + \frac{\mathrm{L}}{\mu} \tag{45}$$

Moreover,

$$\overline{\Sigma}_\gamma = \gamma\mathbf{T}\mathcal{C}(\theta^\star) + B_2\gamma^{3/2} \tag{46}$$

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

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

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

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 $(\xi_k)_{k\ge 1}$. First, applying a third-order Taylor expansion of $\nabla f(\theta)$ around $\theta^\star$, for any $\theta \in \mathbb{R}^d$, we obtain

$$\nabla f(\theta) = \mathrm{H}^\star(\theta - \theta^\star) + (1/2)\{\nabla^3 f(\theta^\star)\}(\theta - \theta^\star)^{\otimes 2} + G(\theta) \tag{47}$$

where $G(\theta)$ is defined in (32) 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 Assumption 2,

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

Thus, integrating (47) with respect to π_γ, we get from (47) that

$$\mathrm{H}^\star(\overline{\theta}_\gamma - \theta^\star) + (1/2)\{\nabla^3 f(\theta^\star)\}\left[\int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right] = -\int_{\mathbb{R}^d} G(\theta)\pi_\gamma(\mathrm{d}\theta) \tag{48}$$

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

$$\overline{\Sigma}_\gamma = \int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \tag{49}$$

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 (31). Hence, taking second moment w.r.t. π_γ from both sides, we get that

$$\begin{aligned}\overline{\Sigma}_\gamma = (\mathrm{I} - \gamma\,\mathrm{H}^\star)\overline{\Sigma}_\gamma(\mathrm{I} - \gamma\,\mathrm{H}^\star) + \gamma^2\int_{\mathbb{R}^d}\mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) + \gamma^2\int_{\mathbb{R}^d}\{\eta(\theta)\}^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \\ - \gamma\int_{\mathbb{R}^d}\left[(\mathrm{I} - \gamma\,\mathrm{H}^\star)(\theta - \theta^\star)\{\eta(\theta)\}^\top + \eta(\theta)(\theta - \theta^\star)^\top(\mathrm{I} - \gamma\,\mathrm{H}^\star)\right]\pi_\gamma(\mathrm{d}\theta) \quad (50)\end{aligned}$$

In the above equation $\mathcal{C}(\theta)$ is defined in (43), and we additionally used that $\mathsf{E}\big[\varepsilon_1(\theta_0)|\mathcal{F}_0\big] = 0$. Moreover, (43) together with $\mathbf{C}1(6)$ 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 $\|B\| \le \mathsf{C}_2$. Thus, from (50) together with $\mathbf{C}1(6)$ we obtain that $\overline{\Sigma}_\gamma$ is a solution to the matrix equation

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

which can be written using the vectorization operation as

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

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 Assumption 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 $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d \ge \mu > 0$ 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

$$\begin{aligned}S &= \mathrm{H}^\star\otimes\mathrm{I} + \mathrm{I}\otimes\mathrm{H}^\star \in \mathbb{R}^{d^2\times d^2} \\ R &= \mathrm{H}^\star\otimes\mathrm{H}^\star \in \mathbb{R}^{d^2\times d^2}\end{aligned} \quad (51)$$

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 $\gamma\|RS^{-1}\| \le 1$. Sice R and S are diagonalizable in the same orthogonal basis $\{u_i \otimes u_j\}_{1\le i,j\le d}$ with the eigenvalues $\lambda_i\lambda_j$ and $\lambda_i + \lambda_j$, respectively, the condition $\gamma\|RS^{-1}\| \le 1$ holds provided that $\gamma \le 2/\,\mathrm{L}$. Hence, for $\gamma \le 1/\,\mathrm{L}$, 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 (48), we arrive at the expansion formula (44). □

We now state an auxiliary lemma about the function $\mathcal{C}(\theta)$ from (43).

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

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

where the constant C_2 *is given by*

$$\mathsf{C}_2 = \left(\frac{\mathrm{L}^2\, \mathsf{D}_{\mathrm{last},2}}{\sqrt{\mu}} + \frac{\mathrm{L}\,\sqrt{\mathsf{D}_{\mathrm{last},2}}}{\sqrt{\mu}} \right) \tau_2^2 \tag{52}$$

Proof. Recall that

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

Hence, using the definition of $\mathcal{C}(\theta)$ in (43), 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 Assumption 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 $\mathbf{C}1(2)$, we obtain

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

We conclude the proof by noting that $\gamma\mu \le 1$. □

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

$$\begin{aligned}\theta_{k+1} &= \theta_k - \gamma\nabla F(\theta_k, \xi_{k+1}), \quad \theta_0 \sim \nu \\ \widetilde{\theta}_{k+1} &= \widetilde{\theta}_k - \gamma\nabla F(\widetilde{\theta}_k, \xi_{k+1}), \quad \widetilde{\theta}_0 \sim \pi_\gamma\end{aligned}$$

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

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

Using (40) and $\mathbf{C}1(2)$, we obtain

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

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

A.3 Proof of Proposition 3

Note that

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

Applying (24), we obtain

$$\|\mathsf{E}_\nu[\overline{\theta}_n^{(RR)} - \theta^\star]\| \lesssim \mathsf{C}_1\gamma^{3/2} + \mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n) \tag{53}$$

where

$$\|\mathcal{R}_3(\theta_0 - \theta^\star, \gamma, n)\| \lesssim \frac{(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}(\mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^2\big] + \frac{\sqrt{\gamma}\tau_2}{\sqrt{\mu}}) \tag{54}$$

Lemma 3. *Assume Assumption 1, Assumption 3(2), Assumption 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*

$$\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n}\{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^2] \lesssim \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2}^{1/2}\sqrt{\gamma n}\tau_2}{\mu^{1/2}} + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{\gamma\mu}\mathsf{E}_\nu^{1/2}\big[\|\theta_0 - \theta^\star\|^2\big] \tag{55}$$

*Moreover, under Assumption 1, Assumption 3(p), Assumption 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*

$$\begin{aligned}\mathsf{E}_\nu^{1/p}[\|\sum_{k=n+1}^{2n}\{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\}\|^p] \lesssim{}& \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},p}^{1/2}\sqrt{\gamma n}p\tau_p}{\mu^{1/2}} \\ &+ \frac{\mathrm{L}\,p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2}}\mathsf{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p]\end{aligned} \tag{56}$$

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

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

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

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

Thus, we obtain that

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

and the statement (55) follows from the assumption $\mathbf{C}\,1(2)$. In order to prove (56), we apply Burkholder's inequality [Ose12, Theorem 8.6] and obtain

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

where in (a) we have additionally used $\mathbf{C}1(p)$. □

B Proof of Theorem 1

Within this section we often use the definition of the function $\psi:\mathbb{R}^d\to\mathbb{R}^d$ from (32):

$$\psi(\theta)=(1/2)\nabla^3 f(\theta^*)(\theta-\theta^\star)^{\otimes 2} \tag{57}$$

Theorem 3 (Version of Theorem 1 with explicit constants)**.** *Assume Assumption 1, Assumption 2, Assumption 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 $\overline{\theta}_n^{(RR)}$ defined in (30) satisfies*

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

where we have set

$$\begin{aligned}
\mathsf{C}_{\mathrm{RR},1}&=\frac{\mathsf{D}_{\mathrm{last},4}\,\mathrm{L}\,\tau_4^2}{\mu^{3/2}}+\frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2}^{1/2}\tau_2}{\mu^{1/2}} \quad \mathsf{C}_{\mathrm{RR},2}=\frac{\mathsf{D}_{\mathrm{last},2}^{1/2}\tau_2}{\mu^{1/2}}\\
\mathsf{C}_{\mathrm{RR},3}&=\frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},6}^{3/2}\tau_6^3}{\mu^{3/2}}+\mathsf{C}_1 \quad \mathsf{C}_{\mathrm{RR},4}=\frac{\mathsf{D}_{\mathrm{last},4}\,\mathrm{L}\,\tau_4^2}{\mu}
\end{aligned} \tag{58}$$

C_1 is defined in (45), and the remainder term $\mathcal{R}_4(n,\gamma,\|\theta_0-\theta^\star\|)$ is given by

$$\begin{aligned}
\mathcal{R}_4(n,\gamma,\|\theta_0-\theta^\star\|)&=\frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}\\
&\times\left(\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6]+\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^4]+\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^2]+\frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu}\right)
\end{aligned} \tag{59}$$

Proof. Using the recursion (26), we obtain that

$$\mathrm{H}^\star(\overline{\theta}_n^{(RR)}-\theta^\star)=\frac{2(\theta_{n+1}^{(\gamma)}-\theta^\star)}{\gamma n}-\frac{2(\theta_{2n}^{(\gamma)}-\theta^\star)}{\gamma n}-\frac{\theta_{n+1}^{(2\gamma)}-\theta^\star}{2\gamma n}+\frac{\theta_{2n}^{(2\gamma)}-\theta^\star}{2\gamma n}$$

$$-\frac{1}{n}\sum_{k=n+1}^{2n}[2\varepsilon_{k+1}(\theta_k^{(\gamma)})-\varepsilon_{k+1}(\theta_k^{(2\gamma)})]-\frac{1}{n}\sum_{k=n+1}^{2n}[2\eta(\theta_k^{(\gamma)})-\eta(\theta_k^{(2\gamma)})] \tag{60}$$

Therefore, applying Minkowski's inequality to the decomposition (60), we obtain for any initial distribution ν that

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

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 $\mathbf{C}1(2)$, we get

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

Applying Lemma 3, we get

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

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

$$\theta_1^{(\gamma)}-\theta^\star=(\mathrm{I}-\gamma\,\mathrm{H}^\star)(\theta_0^{(\gamma)}-\theta^\star)-\gamma\varepsilon_1(\theta_0^{(\gamma)})-\gamma\eta(\theta_0^{(\gamma)}) \tag{61}$$

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

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

where $G(\theta)$ is defined in (32) 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 Assumption 2 together with Proposition 2, we obtain that

$$T_6 = \|2\pi_\gamma(\psi) - \pi_{2\gamma}(\psi)\| \lesssim \mathsf{C}_1 \gamma^{3/2} \tag{62}$$

Finally, using Lemma 7, Lemma 6, and Lemma 4, we obtain that

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

Combining the bounds above completes the proof. □

Below we provide some auxiliary technical lemmas.

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

$$n^{-1}\mathsf{E}_{\pi_\gamma}^{1/2}\big[\|\sum_{k=n+1}^{2n}\{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2\big] \lesssim \frac{\mathsf{D}_{\mathrm{last},4}\,\mathrm{L}_3\,\gamma\tau_4^2}{\mu n^{1/2}} + \frac{\mathsf{D}_{\mathrm{last},4}\,\mathrm{L}_3\,\gamma^{1/2}\tau_4^2}{\mu^{3/2} n^{1/2}} \tag{63}$$

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

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

Using the Markov property, Cauchy–Schwartz inequality, Proposition 1, and Lemma 8, we obtain

$$\begin{aligned} &\mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\psi(\theta_k) - \pi_\gamma(\psi))] && (64)\\ &\quad = \mathsf{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\mathrm{Q}_\gamma^k\psi(\theta_0) - \pi_\gamma(\psi))] && (65)\\ &\quad \overset{(a)}{\lesssim} (1/2)^{k/m(\gamma)}\,\mathrm{L}_3\,\mathsf{E}_{\pi_\gamma}\big[\|\psi(\theta_0) - \pi_\gamma(\psi)\|\int c(\theta_0,\vartheta)d\pi_\gamma(\vartheta)\big] && (66) \end{aligned}$$

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 $\mathbf{C}1(4)$, we get

$$\mathsf{E}_{\pi_\gamma}[\|\psi(\theta_0) - \pi_\gamma\|^2] \le \mathsf{E}_{\pi_\gamma}[\|\psi(\theta_0)\|^2] \le \mathrm{L}_3^2\,\mathsf{E}_{\pi_\gamma}[\|\theta_0 - \theta^\star\|^4] \le \frac{\mathrm{L}_3^2\,\mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^4}{\mu^2} \tag{67}$$

and, using $\mathbf{C}1(2)$ and $\mathbf{C}1(4)$, we get

$$\int\int c^2(\theta_0,\vartheta)d\pi_\gamma(\vartheta)d\pi_\gamma(\theta_0) \tag{68}$$

$$\leq \int\int \|\theta_0-\vartheta\|^2\left(\|\theta_0-\theta^\star\|+\|\vartheta-\theta^\star\|+\frac{2^{3/2}\gamma^{1/2}\tau_2}{\mu^{1/2}}\right)^2 d\pi_\gamma(\vartheta)d\pi_\gamma(\theta_0) \tag{69}$$

$$\lesssim \int\int(\|\theta_0-\theta^\star\|^4+\|\vartheta-\theta^\star\|^4)+\frac{\gamma\tau_2^2}{\mu}(\|\theta_0-\theta^\star\|^2+\|\vartheta-\theta^\star\|^2)d\pi_\gamma(\vartheta)d\pi_\gamma(\theta_0) \tag{70}$$

$$\lesssim \frac{\mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^2}{\mu^2}+\frac{\mathsf{D}_{\mathrm{last},2}\gamma^2\tau_2^4}{\mu^2}\lesssim\frac{\mathsf{D}_{\mathrm{last},4}\gamma^2\tau_4^4}{\mu^2} \tag{71}$$

Using (67), (68), and Cauchy–Schwartz inequality for (64), 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

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

□

Lemma 5. *Assume Assumption 1, Assumption 2, Assumption 3(4). Then for any $\gamma\in(0;\frac{2}{11L}]$, and any $k\in\mathbb{N}$ it holds that*

$$\mathsf{E}[\|\theta_{k+1}-\widetilde{\theta}_{k+1}\|^4|\mathcal{F}_k]\leq(1-\gamma\mu)^2\|\theta_k-\widetilde{\theta}_k\|^4 \tag{72}$$

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

$$\theta_{k+1}=\theta_k-\gamma\nabla F(\theta_k,\xi_{k+1})\quad\theta_0=\theta\in\mathbb{R}^d \tag{73}$$

$$\widetilde{\theta}_{k+1}=\widetilde{\theta}_k-\gamma\nabla F(\widetilde{\theta}_k,\xi_{k+1})\quad\widetilde{\theta}_0=\widetilde{\theta}\in\mathbb{R}^d \tag{74}$$

Expanding the brackets, we obtain that

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

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

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

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

Similarly,

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

Combining all inequalities above, we obtain

$$\begin{aligned}\mathsf{E}[\|\theta_{k+1} - \theta'_{k+1}\|^4 | \mathcal{F}_k] &\le \|\theta_k - \widetilde{\theta}_k\|^4 \\ &\quad - (4\gamma - \gamma^4 \mathrm{L}^3 - 4\gamma^2 \mathrm{L} - 2\gamma^2 \mathrm{L} - 4\gamma^3 \mathrm{L}^2) \langle \nabla f(\theta_k) - \nabla f(\widetilde{\theta}_k), \theta_k - \widetilde{\theta}_k\rangle \|\theta_k - \widetilde{\theta}_k\|^2\end{aligned}$$

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

$$\begin{aligned}\mathsf{E}[\|\theta_{k+1} - \widetilde{\theta}_{k+1}\|^4 | \mathcal{F}_k] &\le (1 - 4\gamma\mu(1 - 11\gamma \mathrm{L}/4)) \|\theta_k - \widetilde{\theta}_k\|^4 \\ &\le (1 - 2\gamma\mu(1 - 11\gamma \mathrm{L}/4))^2 \|\theta_k - \widetilde{\theta}_k\|^4\end{aligned}$$

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

Lemma 6. *Assume Assumption 1, Assumption 2, Assumption 3(4), and* ***C*** *1(4). Then for any* $\gamma \in (0; 1/(\mathrm{L}\, C_{\mathrm{step},4})]$*, any* $n \in \mathbb{N}$ *and initial distribution* ν *it holds*

$$\begin{aligned}n^{-1} E_\nu^{1/2}[\|\sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] &\lesssim n^{-1} E_{\pi_\gamma}^{1/2}[\|\sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \\ &\quad + \frac{\mathrm{L}_3 (1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu} \left(E_\nu^{1/2}[\|\theta_0 - \theta^\star\|^4] + \frac{D_{\mathrm{last},4}\gamma\tau_4^2}{\mu} \right)\end{aligned}$$

Proof. Using the synchronous coupling construction defined in (39) and the corresponding coupling kernel K_γ, we obtain that

$$\begin{aligned}\mathsf{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] &= (\mathsf{E}^{\mathrm{K}_\gamma}_{\nu,\pi_\gamma}[\|\sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2])^{1/2} \\ &\le \mathsf{E}_{\pi_\gamma}^{1/2}[\|\sum_{k=n+1}^{2n} \{\psi(\widetilde{\theta}_k) - \pi_\gamma(\psi)\}\|^2] + (\mathsf{E}^{\mathrm{K}_\gamma}_{\nu,\pi_\gamma}[\|\sum_{k=n+1}^{2n} \{\psi(\theta_k) - \psi(\widetilde{\theta}_k)\}\|^2])^{1/2}\end{aligned} \tag{75}$$

Applying Minkowski's inequality to the last term and using Lemma 8, we get

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

Using Hölder's and Minkowski's inequality and applying Lemma 5 , (41) and

$$\mathsf{E}_\nu^{1/2}\|\theta_k-\theta^\star\|^4 \lesssim (1-\gamma\mu)^k \mathsf{E}_\nu^{1/2}\big[\|\theta_0-\theta^\star\|^4\big] + \frac{\mathsf{D}_{\mathrm{last},4}\gamma\tau_4^2}{\mu} \tag{76}$$

we obtain

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

Combining all inequalities above, we get

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

Substituting the last inequality into (75) we complete the proof. □

Lemma 7. *Assume Assumption 1, Assumption 2, Assumption 3(6), and* $\boldsymbol{C}$ *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*

$$\begin{aligned}
n^{-1}\mathsf{E}^{1/2}_\nu[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2] &\le n^{-1}\mathsf{E}^{1/2}_\nu[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2] \\
&\quad + \frac{\mathrm{L}_4(1-\gamma\mu)^{(n+1)/2}}{n\gamma\mu}\mathsf{E}^{1/2}_\nu[\|\theta_0-\theta^\star\|^6] + \frac{\mathrm{L}_4\,\mathsf{D}^{3/2}_{\mathrm{last},6}\gamma^{3/2}\tau_6^3}{3\mu^{3/2}}
\end{aligned} \tag{77}$$

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

$$\eta(\theta) = \psi(\theta) + \frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star+(1-t)\theta)\,dt\right)(\theta-\theta^\star)^{\otimes 3} \tag{78}$$

and using Assumption 2, we obtain

$$\|\left(\int_0^1 \nabla^4 f(t\theta^\star+(1-t)\theta)\,dt\right)(\theta-\theta^\star)^{\otimes 3}\| \le \mathrm{L}_4\,\|\theta-\theta^\star\|^3 \tag{79}$$

Therefore, combining (78), Assumption 2, and applying Minkowski's inequality, we get

$$\begin{aligned}
\mathsf{E}^{1/2}_\nu[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2] &\le \mathsf{E}^{1/2}_\nu[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2] \\
&\quad + \frac{\mathrm{L}_4}{6}\sum_{k=n+1}^{2n}\mathsf{E}^{1/2}_\nu[\|\theta_k-\theta^\star\|^6]
\end{aligned} \tag{80}$$

Applying $\mathbf{C}1(6)$ for the last term of (80), we get

$$\begin{aligned}\mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\eta(\theta_k)-\pi_\gamma(\psi)\|^2\Big] \lesssim \mathsf{E}_\nu^{1/2}\Big[\sum_{k=n+1}^{2n}\|\psi(\theta_k)-\pi_\gamma(\psi)\|^2\Big] + \frac{\mathrm{L}_4\, n\mathsf{D}_{\mathrm{last},6}^{3/2}\gamma^{3/2}\tau_6^3}{\mu^{3/2}} \\ + \frac{\mathrm{L}_4(1-\gamma\mu)^{3(n+1)/2}}{1-(1-\gamma\mu)^{3/2}}\mathsf{E}_\nu^{1/2}[\|\theta_0-\theta^\star\|^6]\end{aligned} \tag{81}$$

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 (57).

Lemma 8. *Let $\psi(\cdot)$ be a function defined in* (57)*. 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,

$$\|\psi(\theta)-\psi(\theta')\| \le \frac{1}{2}\|T(\theta-\theta^\star)^{\otimes 2} - T(\theta'-\theta^\star)^{\otimes 2}\| \tag{82}$$

Note that

$$\|T\| = \sup_{x\neq 0, y\neq 0, z\neq 0}\frac{\sum\limits_{i,j,k}T_{ijk}x_iy_jz_k}{\|x\|\|y\|\|z\|} \ge \sup_{x\neq 0,y\neq 0}\sup_{z\neq 0}\frac{\sum\limits_k z_k\sum\limits_{i,j}T_{ijk}x_iy_j}{\|z\|\|y\|\|x\|} = \sup_{x\neq 0,y\neq 0}\frac{\|t(x,y)\|}{\|y\|\|x\|} \tag{83}$$

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

$$\|t(x,y)\| \le \|x\|\|y\|\|T\| \tag{84}$$

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

$$\begin{aligned}v_k = \sum_{i,j}T_{ijk}(x_ix_j - y_iy_j) = \sum_{i,j}T_{ijk}((x_i-y_i)x_j + (x_i-y_i)y_j) = \\ \sum_{i,j}T_{ijk}(x_i-y_i)x_j + \sum_{i,j}T_{ijk}(x_i-y_i)y_j,\end{aligned} \tag{85}$$

where the first inequality is true since $T_{ijk} = T_{jik}$ by definition of T. Combining (84) and (B) and using triangle inequality, we obtain

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

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

C Proof of Theorem 2

Theorem 4 (Version of Theorem 2 with explicit constants)**.** *Let $p \geq 2$ and assume Assumption 1, Assumption 2, Assumption 3(3p), and C1(3p). Then for any $\gamma \in (0; 1/(\mathrm{L}\, C_{\mathrm{step},3p})]$, initial distribution ν, and $n \in \mathbb{N}$, the estimator $\overline{\theta}_n^{(RR)}$ defined in (30) satisfies*

$$E_\nu^{1/p}[\| \mathrm{H}^\star(\overline{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\mathrm{Tr}\, \Sigma_\varepsilon^\star} p^{1/2}}{n^{1/2}} + \frac{p\tau_p}{n^{1-1/p}} + \frac{C_{\mathrm{RR},5}}{n\gamma^{1/2}} + \frac{C_{\mathrm{RR},6}\gamma^{1/2}}{n^{1/2}} + C_{\mathrm{RR},7}\gamma^{3/2} + \frac{C_{\mathrm{RR},8}}{n} + \mathcal{R}_5(n, \gamma, \|\theta_0 - \theta^\star\|)$$

where we have set

$$\begin{aligned} C_{\mathrm{RR},5} &= \frac{D_{\mathrm{last},p}^{1/2}\tau_p}{\mu^{1/2}} \quad C_{\mathrm{RR},6} = \frac{\mathrm{L}\, D_{\mathrm{last},p}^{1/2} p\tau_p}{\mu^{1/2}} + \frac{\mathrm{L}\, D_{\mathrm{last},2p} p\tau_{2p}^2}{\mu^{3/2}} \\ C_{\mathrm{RR},7} &= C_1 + \frac{\mathrm{L}\, D_{\mathrm{last},3p}^{3/2}\tau_{3p}^3}{\mu^{3/2}} \quad C_{\mathrm{RR},8} = \frac{\mathrm{L}\, D_{\mathrm{last},2p}\tau_{2p}}{\mu^2} \end{aligned} \tag{86}$$

C_1 is defined in (45), and the remainder term $\mathcal{R}_5(n, \gamma, \|\theta_0 - \theta^\star\|)$ is given by

$$\begin{aligned} \mathcal{R}_5(n, \gamma, \|\theta_0 - \theta^\star\|) &= \frac{(1-\gamma\mu)^{(n+1)/2}}{\gamma n} E_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^p\big] + \frac{\mathrm{L}\, p(1-\gamma\mu)^{(n+1)/2}}{\mu^{1/2}\gamma^{1/2} n} E_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p] \\ &\quad + \frac{\mathrm{L}(1-\gamma\mu)^{(n+1)/2} p^2}{\gamma\mu^2} E_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^{2p}\big] + \frac{\mathrm{L}(1-\gamma\mu)^{(3/2)n}}{\gamma\mu} E_\nu^{1/p}\big[\|\theta_0 - \theta^\star\|^{3p}\big] \end{aligned} \tag{87}$$

Proof. Using the decomposition (60), we obtain that for any $p \geq 2$, it holds that

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

$$\underbrace{+\frac{1}{n}\sum_{k=n+1}^{2n}\mathsf{E}_\nu^{1/p}\big[\|G(\theta_k^{(\gamma)})\|^p\big]+\frac{1}{n}\sum_{k=n+1}^{2n}\mathsf{E}_\nu^{1/p}\big[\|G(\theta_k^{(2\gamma)})\|^p\big]}_{T_8}$$

Now we upper bounds the terms above separately. Applying first the Pinelis version of Rosenthal inequality [Pin94] together with Assumption $3(p)$, we obtain that

$$T_1 \lesssim \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}p^{1/2}}{n^{1/2}}+\frac{p\tau_p}{n^{1-1/p}}$$

Applying $\mathbf{C}1(p)$ (which is implied by $\mathbf{C}1(3p)$), 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 3 (see the bound (56)), 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 (61) and (62), we obtain

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

Applying Proposition 4, 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 [DMN+23, 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 (32) together with $\mathbf{C}1(3p)$, 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. □

C.1 Proof of Proposition 4

In the proof below we use the notation

$$\overline{\psi}(\theta)=\psi(\theta)-\pi_\gamma(\psi)$$

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

$$m = m(\gamma) = \left\lceil \frac{2\log 4}{\gamma\mu} \right\rceil \tag{88}$$

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

$$\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=0}^{n-1}\overline{\psi}(\theta_k)\|^p\big] \le \mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=0}^{\lfloor n/m\rfloor m-1}\overline{\psi}(\theta_k)\|^p\big] + m\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\overline{\psi}(\theta_0)\|^p\big] \tag{89}$$

Now we consider the Poisson equation, associated with Q^m_γ and function $\overline{\psi}$, that is,

$$g_m(\theta) - \mathrm{Q}^m_\gamma g_m(\theta) = \overline{\psi}(\theta) \tag{90}$$

The function

$$g_m(\theta) = \sum_{k=0}^{\infty}\mathrm{Q}^{km}_\gamma\overline{\psi}(\theta) \tag{91}$$

is well-defined under the assumptions Assumption 1, Assumption 2, Assumption 3($2p$), and $\mathbf{C}1(2p)$. Moreover, g_m is a solution of the Poisson equation (90). Define $q := \lfloor n/m\rfloor$, then we have

$$\sum_{k=0}^{qm-1}\overline{\psi}(\theta_k) = \sum_{r=0}^{m-1}B_{m,r} \quad \text{with} \quad B_{m,r} = \sum_{k=0}^{q-1}\big\{g_m(\theta_{km+r}) - \mathrm{Q}^m_\gamma g_m(\theta_{km+r})\big\} \tag{92}$$

Using Minkowski's inequality, we get from (89), that

$$\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=0}^{n-1}\overline{\psi}(\theta_k)\|^p\big] \le m\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=1}^{q}\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^p\big] + 2m\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\psi(\theta_0\|^p\big] \tag{93}$$

Now we upper bound both terms of (93) separately. Under assumption Assumption 2, and applying $\mathbf{C}1(2p)$, we get

$$\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\psi(\theta_0\|^p\big] \le \frac{\mathrm{L}}{2}\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\theta_0 - \theta^\star\|^{2p}\big] \le \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\gamma\tau^2_{2p}}{2\mu} \tag{94}$$

To proceed with the first term, we apply Burkholder's inequality [Ose12, Theorem 8.6], and obtain that

$$\begin{aligned}\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=1}^{q}\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^p\big] \\ \le p\mathsf{E}^{1/p}_{\pi_\gamma}\big[\big(\sum_{k=1}^{q}\|\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^2\big)^{p/2}\big]\end{aligned} \tag{95}$$

Applying now Minkowski's inequality again, we get

$$\mathsf{E}^{2/p}_{\pi_\gamma}\big[\big(\sum_{k=1}^{q}\|\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^2\big)^{p/2}\big] \le q\mathsf{E}^{2/p}_{\pi_\gamma}\big[\|\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^p\big]$$

$$\lesssim q\left(\mathsf{E}^{2/p}_{\pi_\gamma}[\|g_m(\theta_0)\|^p] + \mathsf{E}^{2/p}_{\pi_\gamma}[\|\mathrm{Q}^m_\gamma g_m(\theta_0)\|^p]\right)$$
$$\lesssim q\mathsf{E}^{2/p}_{\pi_\gamma}[\|g_m(\theta_0)\|^p]$$

It remains to upper bound the moment $\mathsf{E}^{2/p}_{\pi_\gamma}[\|g_m(\theta_0)\|^p]$. In order to do this, we first note that due to the duality theorem [DMPS18, Theorem 20.1.2.], we get that for any $k \in \mathbb{N}$,

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

where the last inequality is due to Proposition 1. Hence, applying the definition of $g_m(\theta)$ in (91), we obtain that

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

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 $\mathbf{C}1(2p)$, we get

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

Combining now the bounds above in (95), we get that

$$\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=1}^{q}\big\{g_m(\theta_{km}) - \mathrm{Q}^m_\gamma g_m(\theta_{(k-1)m})\big\}\|^p\big] \lesssim \frac{\mathsf{D}_{\mathrm{last},2p}p\,\mathrm{L}_3\,\tau^2_{2p}\gamma\sqrt{q}}{\mu} \tag{96}$$

and, hence, substituting into (89), we get

$$\mathsf{E}^{1/p}_{\pi_\gamma}\big[\|\sum_{k=0}^{n-1}\overline{\psi}(\theta_k)\|^p\big] \lesssim \frac{\mathsf{D}_{\mathrm{last},2p}p\,\mathrm{L}_3\,\tau^2_{2p}\gamma\sqrt{q}m}{\mu} + \frac{\mathrm{L}\,\mathsf{D}_{\mathrm{last},2p}\tau^2_{2p}\gamma m}{2\mu} \tag{97}$$

Now the statement follows from the definition of $m = m(\gamma)$ in (88) and $q = \lfloor n/m \rfloor \leq n/m$.