Nonasymptotic Analysis of Stochastic Gradient Descent with the Richardson–Romberg Extrapolation

Alexey Naumov

International Laboratory of Stochastic Algorithms and High-Dimensional Inference
HSE University

October 10, 2024

Joint work with

Denis Belomestny

(Duisburg-Essen University & HSE)

Alain Durmus

(ENS Paris-Saclay)

Marina Sheshukova

(HSE)

Eric Moulines

(Ecole Polytechnique)

Sergey Samsonov

(HSE)

Stochastic minimization problem

- Consider the strongly convex minimization problem, which admits a unique solution $\theta^\star$

$$f(\theta) \to \min_{\theta \in \mathbb{R}^d} . \tag{1}$$

- The access to $\nabla f(\theta)$ is available only through the (unbiased) noisy observations $\nabla F(\theta, \xi)$, where ξ is a random variable on $(\mathsf{Z}, \mathcal{Z})$.
- We solve the problem (1) using the SGD with constant step size γ, starting from initial distribution ν:

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma \nabla F(\theta_k^{(\gamma)}, \xi_{k+1}) , \quad \theta_0 \sim \nu, \tag{2}$$

  where ξ_k is a sequence of i.i.d. random variables.
- Define noise function as

$$\varepsilon_k(\theta) = \nabla F(\theta, \xi_k) - \nabla f(\theta) , \tag{3}$$

  and noise covariance matrix as $\Sigma_\varepsilon^\star = \mathrm{E}[\nabla F(\theta^\star, \xi)^{\otimes 2}]$

Polyak-Ruppert averaged estimator

Consider the Polyak-Ruppert averaged estimator

$$\bar{\theta}_n^{(\gamma)} = \frac{1}{n} \sum_{k=n+1}^{2n} \theta_k^{(\gamma)} \,. \tag{4}$$

- Under appropriate assumptions on f and γ_k,

$$\sqrt{n}(\bar{\theta}_{n_0,n} - \theta^\star) \xrightarrow{d} \mathrm{N}(0, H^{-1}\Sigma_\varepsilon^\star H^{-1}) \,, \quad n \to \infty, \tag{5}$$

  where $H = \nabla^2 f(\theta^\star)$; e.g. Fort [2015]).
- We examine the mean-squared error bounds in the following form:

$$\mathrm{E}^{1/2}[\|\bar{\theta}_{n_0,n} - \theta^\star\|^2] \le \frac{\sqrt{\mathrm{Tr}\, H^{-1}\Sigma_\varepsilon^\star H^{-1}}}{n^{1/2}} + \frac{C(f,d)}{n^{1/2+\delta}} + \dots \tag{6}$$

- The goal of the work is to obtain the result in the form (6) with the best possible constant δ.

Related works: Polyak-Ruppert averaged estimator

- Previous studies considered decreasing step size in the dynamics (2):
  - In Moulines and Bach [2011] for strongly convex functions it was shown that

$$\mathrm{E}^{1/2}[\|\bar{\theta}_n - \theta^*\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\left(H^{-1}\Sigma_\varepsilon^\star H^{-1}\right)}}{\sqrt{n}} + \mathcal{O}(n^{-7/12})$$

    .
  - Gadat and Panloup [2023] improved the result of Moulines and Bach [2011] and for a certain class of functions f, including strongly convex functions it was shown that

$$\mathrm{E}^{1/2}[\|\bar{\theta}_n - \theta^*\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\left(H^{-1}\Sigma_\varepsilon^\star H^{-1}\right)}}{\sqrt{n}} + \mathcal{O}(n^{-5/8})$$

    .

Analysis of SGD: Assumptions on function f

Assumption A1

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

$$\frac{\mu}{2}\|\theta - \theta'\|^2 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \,. \tag{7}$$

Assumption A2

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

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq \mathrm{L}_2 \|\theta - \theta'\| \,. \tag{8}$$

Moreover, f has uniformly bounded 3-rd and 4-th derivatives, such that

$$\|\nabla^i f(\theta)\| \leq \mathrm{L}_i \,, \text{ for } i \in \{3, 4\} \,. \tag{9}$$

Analysis of SGD: Assumptions on the noisy gradient ∇F

Assumption A3(p)

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

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

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

$$\begin{aligned}\mathrm{L}_1^{q-1}\,\|\theta_1-\theta_2\|^{q-2}\langle\nabla f(\theta_1)-\nabla f(\theta_2),\theta_1-\theta_2\rangle \\ \ge \mathrm{E}_{\xi\sim\mathbb{P}_\xi}[\|\nabla F(\theta_1,\xi)-\nabla F(\theta_2,\xi)\|^q]\,. \qquad (10)\end{aligned}$$

Note that, A3(p) generalizes the well-known L_1-co-coercivity assumption, see Dieuleveut et al. [2020b]. A sufficient condition is to assume that $F(\theta,\xi)$ is $\mathbb{P}_\xi$-a.s. convex with respect to $\theta \in \mathbb{R}^d$.

Kantorovich - Wasserstein distance

Definition

The function $c : \mathsf{Z} \times \mathsf{Z} \to \mathbb{R}_+$ is called distance-like if it is symmetric, lower semi-continuous and $c(x, y) = 0$ if and only if $x = y$.

Definition

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

$$\mathbf{W}_c(\xi, \xi') = \inf_{\mathcal{C} \in \mathscr{C}(\xi, \xi')} \int_{\mathsf{Z} \times \mathsf{Z}} c(z, z') \mathcal{C}(\mathrm{d}z, \mathrm{d}z') \,. \tag{11}$$

Analysis of SGD: Bias

- Under assumptions A1-A3(2) the sequence $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$ is a homogeneous Markov chain with the Markov kernel

$$\mathrm{Q}_\gamma(\theta, \mathrm{A}) = \int_{\mathbb{R}^d} \mathbb{1}_{\mathrm{A}}(\theta - \gamma\nabla F(\theta, z))\mathrm{P}_\xi(\mathrm{d}z)\,, \quad \theta \in \mathbb{R}^d\,,\, \mathrm{A} \in \mathrm{B}(\mathbb{R}^d)\,; \quad (12)$$

see Dieuleveut et al. [2020a].

- Introduce distance-like function

$$c(\theta, \theta') = \|\theta - \theta'\|(\|\theta - \theta^*\| + \|\theta' - \theta^*\| + c_0\gamma^{1/2})\,, \quad (13)$$

where we set $c_0 = 2^{3/2}\tau_2/\mu^{1/2}$

Lemma 1

Assume A1-A3(2). Then, for any $\gamma \in (0; \frac{1}{2L}]$ the Markov kernel Q_γ admits a unique invariant probability measure π_γ. Moreover, for all $\theta \in \mathbb{R}^d$ and $k \in \mathbb{N}$

$$\mathbf{W}_c(\nu\mathrm{Q}_\gamma^k, \pi_\gamma) \le 4(1/2)^{k/m(\gamma)}\mathbf{W}_c(\nu, \pi_\gamma), \text{ where } m(\gamma) = \lceil 2\frac{\log 4}{\gamma\mu}\rceil \quad (14)$$

- However, unless the function f is quadratic,

$$\mathrm{E}_{\pi_\gamma}[\theta] \neq \theta^\star\,.$$

Analysis of SGD: Bias

- We consider the following condition.

Assumption C1(p)

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

$$\mathrm{E}_{\nu}^{2/p}\left[\|\theta_k^{(\gamma)} - \theta^\star\|^p\right] \leq (1-\gamma\mu)^k \mathrm{E}_{\nu}^{2/p}\left[\|\theta_0 - \theta^\star\|^p\right] + \mathrm{D}_{\mathrm{last},p}\gamma\tau_p^2/\mu\,. \tag{15}$$

Moreover, for the stationary distribution π_γ it holds that

$$\mathrm{E}_{\pi_\gamma}^{2/p}\left[\|\theta_0^{(\gamma)} - \theta^\star\|^p\right] \leq \mathrm{D}_{\mathrm{last},p}\gamma\tau_p^2/\mu\,. \tag{16}$$

- It is important to recognize that C1(p) is not independent from the previous assumptions A1-A3(p). In particular, [Dieuleveut et al., 2020a, Lemma 13] implies that, under A1-A3(p) with $p \geq 2$, the bound (16) holds for $\gamma \in (0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},p})]$ with some constants $\mathrm{D}_{\mathrm{last},p}$ and $\mathrm{C}_{\mathrm{step},p}$, which depends only upon p.

Analysis of SGD: Bias

Proposition 2, Theorem 4 in Dieuleveut et al. [2020a]

A1-A3(6), C1(6). Then, for any $\gamma \in (0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$, the following bias expansion holds:

$$\mathrm{E}_{\pi_\gamma}[\theta] := \int_{\mathbb{R}^d} u \pi_\gamma(\mathrm{d}u) = \theta^\star + \gamma \Delta_1 + \mathcal{O}(\gamma^{3/2}) \,, \tag{17}$$

$$\mathrm{E}_{\pi_\gamma}[(\theta - \theta^\star)^{\otimes 2}] := \int_{\mathbb{R}^d} (u - \theta^\star)^{\otimes 2} \pi_\gamma(\mathrm{d}u) = \gamma \Delta_2 + \mathcal{O}(\gamma^{3/2}) \,, \tag{18}$$

where $\Delta_1 \in \mathbb{R}^d, \Delta_2 \in \mathbb{R}^{d \times d}$ are constants independent of the step size γ. Moreover, for any starting point $\theta_0 \in \mathbb{R}^d$, it holds that

$$\mathrm{E}[\bar{\theta}_n] = \theta^\star + \gamma \Delta_1 + \mathcal{O}(\gamma^{3/2}) + \Delta_1(\|\theta_0 - \theta^\star\|, \gamma, n) \,, \tag{19}$$

where $\|\Delta_1(\|\theta_0 - \theta^\star\|, \gamma, n)\| \leq \frac{e^{-\gamma\mu(n+1)/2}}{n\gamma\mu}(\|\theta_0 - \theta^\star\| + \frac{\sqrt{2\gamma}\tau_2}{\sqrt{\mu}})$.

Bound for Polyak-Ruppert averaged estimator

Theorem 3

Assume A1-A3(6), C1(6). Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$ and any $n \in \mathbb{N}$, the sequence of Polyak-Ruppert estimates (4) satisfies

$$\mathrm{E}_\nu^{1/2}[\|H(\bar{\theta}_n^{(\gamma)} - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{1}{\gamma^{1/2} n} + \gamma + \frac{\gamma^{1/2}}{n^{1/2}} + R_1(n, \gamma, \|\theta_0 - \theta^\star\|)\,, \tag{20}$$

where

$$R_1(n, \gamma, \|\theta_0 - \theta^\star\|) \lesssim \frac{e^{-\gamma\mu(n+1)/2}}{\gamma n}(\mathrm{E}_\nu^{1/2}\left[\|\theta_0 - \theta^\star\|^2\right] + \mathrm{E}_\nu^{1/2}\left[\|\theta_0 - \theta^\star\|^4\right]$$

Theorem 3: sketch of proof

- Note that

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

  where $\varepsilon_{k+1}(\theta_k)$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$.
- Set

$$\eta(\theta) = \nabla f(\theta) - H(\theta - \theta^\star)\,,$$

- We get

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

- Rearranging the terms, we obtain

$$H(\theta_k - \theta^\star) = \frac{\theta_k - \theta_{k+1}}{\gamma} - \varepsilon_{k+1}(\theta_k) - \eta(\theta_k)\,. \tag{21}$$

Theorem 3: sketch of proof

- Summing the recurrence (21), we obtain that

$$H(\bar{\theta}_n - \theta^\star) = \frac{\theta_{n+1} - \theta^\star}{\gamma n} - \frac{\theta_{2n} - \theta^\star}{\gamma n} - \frac{1}{n}\sum_{k=n+1}^{2n} \{\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\} + \frac{1}{n}\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta^\star) - \frac{1}{n}\sum_{k=n+1}^{2n} \eta(\theta_k)\,. \tag{22}$$

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

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

- Using A2, we thus obtain that

$$\|\eta(\theta_k)\| \leq \frac{1}{2} L_3 \|\theta_k - \theta^\star\|^2\,.$$

- Hence, applying Minkowski's inequality, we get

$$\mathrm{E}_\nu^{1/2}[\|H(\bar{\theta}_n - \theta^\star)\|^2] \leq \frac{\mathrm{E}_\nu^{1/2}[\|\theta_{n+1} - \theta^\star\|^2]}{\gamma n} + \frac{\mathrm{E}_\nu^{1/2}[\|\theta_{2n} - \theta^\star\|^2]}{\gamma n}$$
$$+ \frac{1}{n}\mathrm{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^2] + \frac{1}{n}\mathrm{E}_\nu^{1/2}[\|\sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta^\star)\|^2]$$
$$+ \frac{L_3}{2n}\sum_{k=n+1}^{2n} \mathrm{E}_\nu^{1/2}[\|\theta_k - \theta^\star\|^4]$$

Theorem 3: sketch of proof

- We have

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

- Since ξ_k are i.i.d, we have

$$\mathrm{E}_\nu^{1/2}[\| \sum_{k=n+1}^{2n} \varepsilon_{k+1}(\theta^\star)\|^2] = \sqrt{n \operatorname{Tr} \Sigma_\varepsilon^\star} \tag{23}$$

- Using A3, we get

$$\mathrm{E}_\nu[\|\varepsilon_{k+1}(\theta_k) - \varepsilon_{k+1}(\theta^\star)\|^2] \leq L^2 \mathrm{E}_\nu[\|\theta_k - \theta^\star\|^2]. \tag{24}$$

- Applying C1[4], we complete the proof.

Theorem 3: Discussion

- Setting optimal step size γ depending on the number of samples n, we arrive at $\gamma \approx n^{-2/3}$, which yields an error bound of order:

$$\mathrm{E}_{\nu}^{1/2}[\|H(\bar{\theta}_n - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\left(\frac{1}{n^{2/3}}\right) + R_1(\|\theta_0 - \theta^\star\|, n)\,. \quad (25)$$

- There are different results in the literature that provide various decay rates of the second-order term in (25). However, all these results are known to be suboptimal for the first-order methods.
- In fact, the recent result of Li et al. [2022] shows that a second-order error can be achieved by modifying the SGD algorithm with averaging and control variates

$$\mathrm{E}^{1/2}[\|H(\bar{\theta}_n - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\left(\frac{1}{n^{3/4}}\right) + R_1'(\|\theta_0 - \theta^\star\|, n)\,. \quad (26)$$

- Our goal is to obtain an analogue of the 2-moment bound (26), using a simpler algorithm.

Richardson-Romberg estimator

- We construct two parallel chains based on the same sequence of noise variables $\{\xi_k\}_{k\in\mathbb{N}}$:

$$\begin{aligned}\theta_{k+1}^{(\gamma)} &= \theta_k^{(\gamma)} - \gamma\nabla F(\theta_k^{(\gamma)}, \xi_{k+1})\,, \quad \bar{\theta}_n^{(\gamma)} = \tfrac{1}{n}\textstyle\sum_{k=n+1}^{2n}\theta_k^{(\gamma)}\,,\\ \theta_{k+1}^{(2\gamma)} &= \theta_k^{(2\gamma)} - 2\gamma\nabla F(\theta_k^{(2\gamma)}, \xi_{k+1})\,, \quad \bar{\theta}_n^{(2\gamma)} = \tfrac{1}{n}\textstyle\sum_{k=n+1}^{2n}\theta_k^{(2\gamma)}\,.\end{aligned} \tag{27}$$

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

$$\bar{\theta}_n^{(RR)} := 2\bar{\theta}_n^{(\gamma)} - \bar{\theta}_n^{(2\gamma)}\,. \tag{28}$$

- Note that in the decomposition (22), the linear statistics $W = n^{-1}\sum_{k=n+1}^{2n}\varepsilon_{k+1}(\theta^\star)$ does not depend upon γ. Hence, using the same sequence $\{\xi_k\}_{k\in\mathbb{N}}$ of noise variables in (27) yields an estimator $\bar{\theta}_n^{(RR)}$, such that its leading component of the variance still equals W.

Richardson-Romberg estimator: Bias

Proposition 4

Assume A1-A3(6), C1(6). Then, for any $\gamma \in (0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$, and any starting point $\theta_0 \in \mathbb{R}^d$, it holds that

$$\mathrm{E}_\nu[\bar{\theta}_n^{(RR)}] = \theta^\star + \mathcal{O}(\gamma^{3/2}) + \Delta_2(\|\theta_0 - \theta^\star\|, \gamma, n)\,, \tag{29}$$

where $\|\Delta_2(\|\theta_0 - \theta^\star\|, \gamma, n)\| \leq 3\frac{e^{-\gamma\mu(n+1)/2}}{n\gamma\mu}(\mathrm{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^2] + \frac{\sqrt{2\gamma}\tau_2}{\sqrt{\mu}})$.

Bound for Richardson-Romberg estimator

Theorem 5

Assume A1-A4(6). Then for any $\gamma \in (0; \min(\frac{1}{9L}, \frac{1}{C_3 L})]$ and any $n \in \mathbb{N}$, the estimator defined in 28 satisfies

$$\mathrm{E}_\nu^{1/2}[\|H(\bar{\theta}_n^{(RR)} - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{\gamma^{1/2}}{n^{1/2}} + \frac{1}{\gamma^{1/2} n} + \frac{\gamma}{n^{1/2}} + \gamma^{3/2} + R_2(n, \gamma, \|\theta - \theta^\star\|),$$

where

$$R_2(n, \gamma, \|\theta_0 - \theta^\star\|) \lesssim \frac{e^{-\gamma\mu(n+1)/2}}{\gamma n}\big(\mathrm{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^2] + \mathrm{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^4] + \mathrm{E}_\nu^{1/2}[\|\theta_0 - \theta^\star\|^6] + \gamma\big)$$

Theorem 5: sketch of proof

- Using the recursion 21, we obtain that

$$H(\bar{\theta}_n^{(RR)} - \theta^\star) = 2\frac{\theta_{n+1}^\gamma - \theta^\star}{\gamma n} - 2\frac{\theta_{2n}^\gamma - \theta^\star}{\gamma n} - \frac{\theta_{n+1}^{2\gamma} - \theta^\star}{2\gamma n} + \frac{\theta_{2n}^{2\gamma} - \theta^\star}{2\gamma n}$$
$$- \frac{2}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta_k^\gamma) - \varepsilon_{k+1}(\theta^\star)] + \frac{1}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta_k^{2\gamma}) - \varepsilon_{k+1}(\theta^\star)]$$
$$+ \frac{1}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta^\star)] - \frac{1}{n}\sum_{k=n+1}^{2n}[2\eta(\theta_k^\gamma) - \eta(\theta_k^{2\gamma})]$$

- Define the function $\psi(\theta) = (1/2)\nabla^3 f(\theta^*)(\theta - \theta^\star)^{\otimes 2}$
- Applying the 4-rd order Taylor expansion with integral remainder, we get that

$$\eta(\theta) = \psi(\theta) + \frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\, dt\right)(\theta - \theta^\star)^{\otimes 3}, \quad (30)$$

- Using A2, we obtain

$$\|\frac{1}{6}\left(\int_0^1 \nabla^4 f(t\theta^\star + (1-t)\theta)\, dt\right)(\theta - \theta^\star)^{\otimes 3}\| \leq \frac{1}{6}L_4\|\theta - \theta^\star\|^3. \quad (31)$$

Theorem 5: sketch of proof

- The key technical element of the proof is to bound

$$\frac{1}{n}\mathrm{E}_{\nu}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2]$$

- Using coupling technique, it can be shown that

$$\frac{1}{n}\mathrm{E}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \lesssim \frac{1}{n}\mathrm{E}_{\pi_\gamma}^{1/2}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] + \frac{e^{-\gamma\mu(n+1)/2}}{\gamma n}(\mathrm{E}_{\nu}^{1/2}[\|\theta_0 - \theta^\star\|^4] + \gamma)$$

- It can be shown that for any $\theta, \theta' \in \mathbb{R}^d$, it holds that

$$\|\psi(\theta) - \psi(\theta')\| \leq \frac{1}{2} L_3 c(\theta, \theta'). \tag{32}$$

- Moreover, see Douc et al. [2018], for any start point $\theta_0 \in \mathbb{R}^d$, it holds

$$|Q^k \psi(\theta_0) - \pi_\gamma(\psi)| \lesssim L_3 (1/2)^{k/m(\gamma)} \mathbf{W}_c(\delta_{\theta_0}, \pi_\gamma).$$

Theorem 5: sketch of proof

- For covariance term, we get

$$\mathrm{E}_{\pi_\gamma}[(\psi(\theta_0) - \pi_\gamma(\psi))^T(\psi(\theta_k) - \pi_\gamma(\psi))] \lesssim (1/2)^{k/m(\gamma)}\gamma^2$$

- For variance term, we have

$$\mathrm{E}_{\pi_\gamma}[\|\psi(\theta_0) - \pi_\gamma(\psi)\|^2] \leq \mathrm{E}_{\pi_\gamma}[\|\psi(\theta_0)\|^2] \lesssim \gamma^2$$

- Combining results, we obtain

$$\frac{1}{n}\mathrm{E}^{1/2}_{\pi_\gamma}[\| \sum_{k=n+1}^{2n} \{\psi(\theta_k) - \pi_\gamma(\psi)\}\|^2] \lesssim \frac{\gamma}{n^{1/2}} + \frac{\gamma^{1/2}}{n^{1/2}}$$

- It remains to note that from Proposition 2 $\|2\pi_\gamma(\psi) - \pi_{2\gamma}(\psi)\| \lesssim \gamma^{3/2}$
- To obtain result of Theorem 5 it remains to apply Minkowski's inequality to the decomposition 30.

Theorem 5: Discussion

- Setting optimal step size γ depending on the number of samples n, we arrive at $\gamma \approx n^{-1/2}$, which yields an error bound of order:

$$\mathrm{E}_{\nu}^{1/2}[\|H(\bar{\theta}_n^{RR} - \theta^\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + \mathcal{O}\left(\frac{1}{n^{3/4}}\right) + R_2(\|\theta_0 - \theta^\star\|, n)\,.$$

- Now we aim to generalize this result for the p-th moment bounds with $p \geq 2$.

Richardson-Romberg estimator, p-th moment

- The key technical element of our proof for the p-th moment bound is the following statement, which can be viewed as a version of Rosenthal's inequality [Rosenthal, 1970, Pinelis, 1994].

Proposition 6

Let $p \geq 2$ and assume A1-A3($2p$), and C1($2p$). Then for any $\gamma \in (0, 1/(\mathsf{L}\,\mathsf{C}_{\mathsf{step},2p})]$, it holds that

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

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

Richardson-Romberg estimator, p-th moment

Repeating the proof of Theorem 5, and using Proposition 6, we obtain the following bound:

Theorem 7

Let $p \geq 2$ and assume A1-A3($3p$) and C1($3p$). Then for any $\gamma \in (0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},3p})]$ and any $n \in \mathbb{N}$, the estimator defined in 28 satisfies

$$\mathrm{E}_\nu^{1/p}[\|H(\bar{\theta}_n^{(RR)} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}p^{1/2}} + \frac{1}{n^{1-1/p}} + \frac{\gamma^{1/2}}{n^{1/2}} + \frac{1}{\gamma^{1/2}n} + \frac{\gamma}{n^{1/2}} + \gamma^{3/2} + R_3(n, \gamma, \|\theta - \theta^\star\|),$$

where

$$R_3(n, \gamma, \|\theta_0 - \theta^\star\|) \lesssim \frac{e^{-\gamma\mu(n+1)/2}}{\gamma n}\left(\mathrm{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^p] + \mathrm{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^{2p}] + \mathrm{E}_\nu^{1/p}[\|\theta_0 - \theta^\star\|^{3p}] + \gamma\right)$$

Theorem 7: Discussion

- Setting optimal step size γ depending on the number of samples n, we arrive at $\gamma \approx n^{-1/2}$, which yields an error bound of order:

$$\mathrm{E}_{\nu}^{1/p}[\|H(\bar{\theta}_n^{RR} - \theta^\star)\|^p] \lesssim \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}\, p^{1/2}}{n^{1/2}} + \mathcal{O}\left(\frac{1}{n^{3/4}}\right) + R_3(\|\theta_0 - \theta^\star\|, n)\,.$$

Thank you!

References I

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, 2020a. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

Aymeric Dieuleveut, Alain Durmus, and Bach Francis. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Annals of Statistics*, 2020b.

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

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

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

Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, pages 909–981. PMLR, 2022.

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

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

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