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

Alexey Naumov

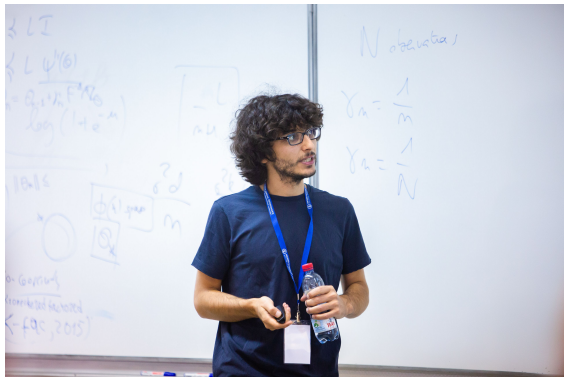
International Laboratory of Stochastic Algorithms and High-Dimensional Inference
HSE University

October 10, 2024

Joint work with

Denis Belomestny

(Duisburg-Essen University & HSE)



Alain Durmus

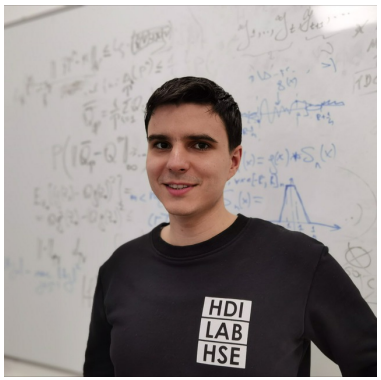
(ENS Paris-Saclay)

Marina Sheshukova

(HSE)

Eric Moulines

(Ecole Polytechnique)



Sergey Samsonov

(HSE)

Stochastic minimization problem

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

$$f(\theta) \;\; ! \;\; \min_{\theta 2 \mathrm{R}^d} . \tag{1}$$

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

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

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

$$\varepsilon_k(\theta) = r\, F(\theta, \xi_k) \quad r\, f(\theta), \tag{3}$$

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

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,

$$p_{\overline{n}}(\bar{\theta}_{n_0,n} \quad \theta^\star) \; \overset{d}{!} \; \mathrm{N}(0, H^{\;1} \Sigma_\varepsilon^\star H^{\;1}) , \quad n \; ! \; 1 , \tag{5}$$

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

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

- 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

$$\mathrm{E}^{1/2}[k\bar{\theta}_n \quad \theta \ k^2] \ . \ \frac{\sqrt{\mathrm{Tr}\,(H \ ^1\Sigma^\star_\varepsilon H \ ^1)}}{p\overline{n}} + O(n \ ^{7/12})$$

    .
  - 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

$$\mathrm{E}^{1/2}[k\bar{\theta}_n \quad \theta \ k^2] \ . \ \frac{\sqrt{\mathrm{Tr}\,(H \ ^1\Sigma^\star_\varepsilon H \ ^1)}}{p\overline{n}} + O(n \ ^{5/8})$$

    .

Analysis of SGD: Assumptions on function f

Assumption A1

The function f is μ-strongly convex on R^d, that is, it is continuously differentiable and there exists a constant $\mu > 0$, such that for any $\theta, \theta^0 \; 2 \; \mathrm{R}^d$, it holds that

$$\frac{\mu}{2} k\theta \quad \theta^0 k^2 \quad f(\theta) \quad f(\theta^0) \quad h \, r \, f(\theta^0), \theta \quad \theta^0 i \, . \tag{7}$$

Assumption A2

The function f is 4 times continuously differentiable and L_2-smooth on R^d, i.e., it is continuously differentiable and there is a constant $L_2 > 0$, such that for any $\theta, \theta^0 \; 2 \; \mathrm{R}^d$,

$$k \, r \, f(\theta) \quad r \, f(\theta^0) k \quad \mathrm{L}_2 \, k\theta \quad \theta^0 k \, . \tag{8}$$

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

$$k \, r^{\, i} f(\theta) k \quad \mathrm{L}_i \, , \text{ for } i \; 2 \; f3, 4g \, . \tag{9}$$

Analysis of SGD: Assumptions on the noisy gradient $r\,F$

Assumption A3(p)

$f\xi_k g_{k2\mathbb{N}}$ is a sequence of independent and identically distributed (i.i.d.) random variables with distribution P_ξ, such that ξ_i and θ_0 are independent and for any $\theta\ 2\ \mathrm{R}^d$ it holds that

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

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

$$\mathrm{L}_1^q\ \ {}^1\,k\theta_1\quad \theta_2 k^q\ \ {}^2h\,r\,f(\theta_1)\quad r\,f(\theta_2),\theta_1\quad \theta_2 i$$
$$\mathrm{E}_{\xi\ \ \mathrm{P}_\xi}[k\,r\,F(\theta_1,\xi)\quad r\,F(\theta_2,\xi)k^q]\,. \qquad (10)$$

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 P_ξ-a.s. convex with respect to $\theta\ 2\ \mathrm{R}^d$.

Kantorovich - Wasserstein distance

Definition

The function $c : Z \quad Z \; ! \; R_+$ 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^0 we denote by $\mathscr{C}(\xi, \xi^0)$ the set of couplings of two probability measures, that is, for any $C \; 2 \; \mathscr{C}(\xi, \xi^0)$ and any $A \; 2 \; Z$ it holds $C(Z \quad A) = \xi^0(A)$ and $C(A \quad Z) = \xi(A)$. We define

$$\mathbf{W}_c(\xi, \xi^0) = \inf_{C 2 \mathscr{C}(\xi, \xi^0)} \int_Z \; _Z \, c(z, z^0) C(\mathrm{d}z, \mathrm{d}z^0) \, . \tag{11}$$

Analysis of SGD: Bias

- Under assumptions A1-A3(2) the sequence $f\theta_k^{(\gamma)} g_{k\,2\,\mathrm{N}}$ is a homogeneous Markov chain with the Markov kernel

$$\mathrm{Q}_\gamma(\theta, \mathrm{A}) = \int_{\mathrm{R}^d} \mathbb{1}_{\mathrm{A}}(\theta \quad \gamma\, r\, F(\theta, z)) \mathrm{P}_\xi(\mathrm{d}z)\,, \quad \theta\ 2\ \mathrm{R}^d\,,\ \mathrm{A}\ 2\ \mathrm{B}(\mathrm{R}^d)\,; \quad (12)$$

see Dieuleveut et al. [2020a].

- Introduce distance-like function

$$c(\theta, \theta^0) = k\theta \quad \theta^0 k(k\theta \quad \theta\ k + k\theta^0 \quad \theta\ k + c_0 \gamma^{1/2})\,, \qquad (13)$$

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

Lemma 1

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

$$\mathbf{W}_c(\nu \mathrm{Q}_\gamma^k, \pi_\gamma) \quad 4(1/2)^{k/m(\gamma)} \mathbf{W}_c(\nu, \pi_\gamma), \text{ where } m(\gamma) = d2\frac{\log 4}{\gamma\mu}e \qquad (14)$$

- 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 $D_{\text{last},p}, C_{\text{step},p}$ 2 depending only on p, such that for any step size γ 2 $(0, 1/(L\, C_{\text{step},p})]$, and any initial distribution ν it holds that

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

Moreover, for the stationary distribution π_γ it holds that

$$\mathrm{E}_{\pi_\gamma}^{2/p}\left[k\theta_0^{(\gamma)} \quad \theta^\star k^p\right] \quad D_{\text{last},p}\gamma\tau_p^2/\mu\,. \tag{16}$$

- 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 2, the bound (16) holds for γ 2 $(0, 1/(L\, C_{\text{step},p})]$ with some constants $D_{\text{last},p}$ and $C_{\text{step},p}$, 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 γ *2* $(0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$, the following bias expansion holds:

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

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

where Δ_1 *2* R^d, Δ_2 *2* $\mathrm{R}^{d\ \ d}$ are constants independent of the step size γ. Moreover, for any starting point θ_0 *2* R^d, it holds that

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

where $k\Delta_1(k\theta_0 \quad \theta^\star k, \gamma, n)k \quad \frac{e^{\ \gamma\mu(n+1)/2}}{n\gamma\mu}(k\theta_0 \quad \theta^\star k + \frac{p\overline{2\gamma}\tau_2}{p\overline{\mu}})$.

Bound for Polyak-Ruppert averaged estimator

Theorem 3

Assume A1-A3(6), C1(6). Then for any γ *2* $(0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$ and any *n 2* N, the sequence of Polyak-Ruppert estimates (4) satisfies

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

where

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

Theorem 3: sketch of proof

- Note that

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

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

$$\eta(\theta) = r\, f(\theta) \quad H(\theta \quad \theta^\star)\,,$$

- We get

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

- Rearranging the terms, we obtain

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

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

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

- Since ξ_k are i.i.d, we have

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

- Using A3, we get

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

- 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 $\gamma \quad n^{-2/3}$, which yields an error bound of order:

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

- 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

$$\mathrm{E}^{1/2}[\|H(\bar{\theta}_n \quad \theta^\star)\|^2] \;.\; \frac{\sqrt{\mathrm{Tr}\,\Sigma_\varepsilon^\star}}{n^{1/2}} + O\left(\frac{1}{n^{3/4}}\right) + R_1^0(\|\theta_0 \quad \theta^\star\|, n)\,. \quad (26)$$

- 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 $f\xi_k g_{k\,2N}$:

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

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

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

- 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 $f\xi_k g_{k\,2N}$ 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 γ *2* $(0, 1/(\mathrm{L}\,\mathrm{C}_{\mathrm{step},6})]$, and any starting point θ_0 *2* R^d, it holds that

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

where $k\Delta_2(k\theta_0 \quad \theta^\star k, \gamma, n)k \quad 3\frac{e^{\quad \gamma\mu(n+1)/2}}{n\gamma\mu}(\mathrm{E}_\nu^{1/2}[k\theta_0 \quad \theta^\star k^2] + \frac{p\overline{2\gamma\tau_2}}{\overline{p}\overline{\mu}})$.

Bound for Richardson-Romberg estimator

Theorem 5

Assume A1-A4(6). Then for any γ 2 $(0; \min(\frac{1}{9L}, \frac{1}{C_3 L})]$ and any n 2 N, the estimator defined in 28 satisfies

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

where

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

Theorem 5: sketch of proof

- Using the recursion 21, we obtain that

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

$$\frac{2}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta_k^{\gamma}) \quad \varepsilon_{k+1}(\theta^\star)] + \frac{1}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta_k^{2\gamma}) \quad \varepsilon_{k+1}(\theta^\star)]$$

$$+\frac{1}{n}\sum_{k=n+1}^{2n}[\varepsilon_{k+1}(\theta^\star)] \quad \frac{1}{n}\sum_{k=n+1}^{2n}[2\eta(\theta_k^{\gamma}) \quad \eta(\theta_k^{2\gamma})]$$

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

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

- Using A2, we obtain

$$k\frac{1}{6}\left(\int_0^1 r^{\,4} f(t\theta^\star + (1 \quad t)\theta)\, dt\right)(\theta \quad \theta^\star)^{\ 3} k \quad \frac{1}{6}L_4 k\theta \quad \theta^\star k^3. \quad (31)$$

Theorem 5: sketch of proof

- The key technical element of the proof is to bound

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

- Using coupling technique, it can be shown that

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

- It can be shown that for any θ, θ^0 *2* R^d, it holds that

$$k\psi(\theta) \quad \psi(\theta^0)k \quad \frac{1}{2}L_3 c(\theta, \theta^0). \tag{32}$$

- Moreover, see Douc et al. [2018], for any start point θ_0 *2* R^d, it holds

$$jQ^k\psi(\theta_0) \quad \pi_\gamma(\psi)j \;.\; L_3(1/2)^{k/m(\gamma)}\mathbf{W}_c(\delta_{\theta_0}, \pi_\gamma).$$

Theorem 5: sketch of proof

- For covariance term, we get

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

- For variance term, we have

$$\mathrm{E}_{\pi_\gamma}[k\psi(\theta_0) \quad \pi_\gamma(\psi)k^2] \quad \mathrm{E}_{\pi_\gamma}[k\psi(\theta_0)k^2] \ . \ \gamma^2$$

- Combining results, we obtain

$$\frac{1}{n}\mathrm{E}^{1/2}_{\pi_\gamma}[k\sum_{k=n+1}^{2n} f\psi(\theta_k) \quad \pi_\gamma(\psi)gk^2] \ . \ \frac{\gamma}{n^{1/2}} + \frac{\gamma^{1/2}}{n^{1/2}}$$

- It remains to note that from Proposition 2 $k2\pi_\gamma(\psi) \quad \pi_{2\gamma}(\psi)k \ . \ \gamma^{3/2}$
- 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 $\gamma \quad n^{-1/2}$, which yields an error bound of order:

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

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

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

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

where $\psi(\theta) = (1/2)\, r^{\,3} f(\theta\ )(\theta \quad \theta^\star)^{\ 2}$.

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

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

where

$$R_3(n,\gamma,k\theta_0 \quad \theta^\star k) \;.\; \frac{e^{\ \gamma\mu(n+1)/2}}{\gamma n}\big(\mathrm{E}_\nu^{1/p}[k\theta_0 \quad \theta^\star k^p] + \mathrm{E}_\nu^{1/p}[k\theta_0 \quad \theta^\star k^{2p}] + \mathrm{E}_\nu^{1/p}[k\theta_0 \quad \theta^\star k^{3p}] + \gamma\big)$$

Theorem 7: Discussion

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

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

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.