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

Marina Sheshukova[1], Denis Belomestny[2,1], Alain Durmus[3], Eric Moulines[3,4], Alexey Naumov[1,5], Sergey Samsonov[1]

[1]HSE University [2]Duisburg–Essen University [3]CMAP, UMR 7641, École Polytechnique [4]Mohamed Bin Zayed University of AI
[5]Steklov Mathematical Institute of the Russian Academy of Sciences

## Problem setting and known results

- Stochastic gradient methods aim to minimize a function $f$ with access only to the noisy gradients:
$$\min_{\theta\in\mathbb{R}^d} f(\theta), \qquad \nabla f(\theta) = \mathsf{E}_{\xi\sim\mathbb{P}_\xi}\big[\nabla F(\theta,\xi)\big]. \tag{1}$$
Here $\xi$ is a noise variable with the distribution $\mathbb{P}_\xi$ and $\theta^\star$ is the unique minimizer. The standard SGD algorithm writes as
$$\theta_{k+1} = \theta_k - \gamma_{k+1}\nabla F(\theta_k,\xi_{k+1}), \qquad \theta_0\in\mathbb{R}^d.$$
- A well-known technique for training stabilization is the Polyak-Ruppert averaging:
$$\bar\theta_{n_0,n} = n^{-1}\sum_{k=n_0+1}^{n+n_0}\theta_k\,,$$
where $n_0$ is the burn-in period. In is known that the averaged iterate is asymptotically normal:
$$\sqrt{n}(\bar\theta_{n_0,n}-\theta^\star) \xrightarrow{d} \mathrm{N}(0,\Sigma_\infty), \qquad n\to\infty\,,$$
where the covariance matrix $\Sigma_\infty$ is minimax-optimal, see [1]. Important research direction for the first-order optimization methods is to obtain the finite-sample bounds of the form
$$\mathsf{E}^{1/2}\big[\|\bar\theta_{n_0,n}-\theta^\star\|^2\big] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\infty}}{n^{1/2}} + \frac{C(f,d)}{n^{1/2+\delta}} + \mathcal{R}(\|\theta_0-\theta^\star\|,n), \tag{2}$$
where $\mathcal{R}(\|\theta_0-\theta^\star\|,n)$ is a (transient) term corresponding to the initial condition. Our aim is to obtain a version of the above bound with the best-known constant $\delta>0$. The best known counterpart of (2) is the bound of [2], which obtains the result with $\delta=1/4$ with the Root-SGD algorithm.

## Assumptions

We aim to solve the problem (1) using SGD with a constant step size $\gamma>0$,, starting from initial distribution $\nu$:
$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma\nabla F(\theta_k^{(\gamma)},\xi_{k+1})\,, \quad \theta_0^{(\gamma)}=\theta_0\sim\nu\,. \tag{3}$$
We focus on the convergence to $\theta^\star$ of the Polyak-Ruppert averaged estimator defined for any $n\ge1$, by
$$\bar\theta_n^{(\gamma)} = n^{-1}\sum_{k=n+1}^{2n}\theta_k^{(\gamma)}\,.$$
**A1**. The function $f$ is continuously differentiable and $\mu$-strongly convex on $\mathbb{R}^d$.
**A2**. The function $f$ is 4 times continuously differentiable and $\mathrm{L}_2$-smooth on $\mathbb{R}^d$, that is, there is a constant $\mathrm{L}_2\ge0$, such that for any $\theta,\theta'\in\mathbb{R}^d$, $\|\nabla f(\theta)-\nabla f(\theta')\|\le \mathrm{L}\|\theta-\theta'\|$. Moreover, $f$ has bounded 3-rd and 4-th derivatives.
**A3**($p$). Random variables $\{\xi_k\}_{k\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) random variables with distribution $\mathbb{P}_\xi$, such that $\xi_i$ and $\theta_0$ 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 $\tau_p$, such that $\mathsf{E}^{1/p}[\|\nabla F(\theta^\star,\xi)\|^p]\le\tau_p$, and for any $q=2,\dots,p$ it holds with some $\mathrm{L}>0$ that for any $\theta_1,\theta_2\in\mathbb{R}^d$,
$$\mathrm{L}^{q-1}\|\theta_1-\theta_2\|^{q-2}\langle\nabla f(\theta_1)-\nabla f(\theta_2),\theta_1-\theta_2\rangle \ge \mathsf{E}_{\xi\sim\mathbb{P}_\xi}[\|\nabla F(\theta_1,\xi)-\nabla F(\theta_2,\xi)\|^q]\,.$$

## Our contributions

- We show that a version of SGD algorithm with constant step size, Polyak-Ruppert averaging, and Richardson-Romberg extrapolation lead to the root-MSE bound (2) with $\delta=1/4$ when applied to strongly convex minimization problems. This result requires that the number of samples $n$ is known a priori to optimize the step size $\gamma$.
- We obtain high-order moment bounds on the error, that is, we obtain for $p\ge2$ the bounds of the form
$$\mathsf{E}^{1/p}[\|\bar\theta_n^{(RR)}-\theta^\star\|^p] \le \frac{c_0 p^{1/2}\sqrt{\operatorname{Tr}\Sigma_\infty}}{n^{1/2}} + \frac{C(f,d,p)}{n^{3/4}} + \mathcal{R}(\|\theta_0-\theta^\star\|,n,p)\,,$$
where $c_0$ is a universal constant, and $\bar\theta_n^{(RR)}$ is a counterpart of $\bar\theta_{n_0,n}$ when using Richardson-Romberg extrapolation.

## SGD iterates as a Markov chain

- The sequence $\{\theta_k^{(\gamma)}\}$ defined by the relation (3) is a time-homogeneous Markov chain with the Markov kernel
$$\mathrm{Q}_\gamma(\theta,\mathsf{A}) = \int_{R^d}\mathbb{1}_{\mathsf{A}}(\theta-\gamma\nabla F(\theta,z))\mathsf{P}_\xi(\mathrm{d}z)\,.$$
- Key ingredient of our result is to show that the Markov chain $\{\theta_k^{(\gamma)}\}$ converges to its invariant distribution $\pi_\gamma$ in a Wasserstein distance $\mathbf{W}_c$, associated with the non-standard distance-like function:
$$c(\theta,\theta') = \|\theta-\theta'\|\big(\|\theta-\theta^\star\|+\|\theta'-\theta^\star\|+\frac{2\sqrt{2}\tau_2\sqrt{\gamma}}{\sqrt{\mu}}\big), \qquad \theta,\theta'\in\mathbb{R}^d\,.$$

**Theorem 1**. Assume **A1**, **A2**, **A3**(4). Then for $n$ large enough, any initial distribution $\nu$, it holds setting $\gamma=n^{-2/3}$ that
$$\mathsf{E}_\nu^{1/2}[\|\mathrm{H}^\star(\bar\theta_n^{(\gamma)}-\theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{\mathsf{C}(\mathrm{L},\mu)}{n^{2/3}} + \mathcal{R}(n,1/n^{2/3},\|\theta_0-\theta^\star\|)\,,$$
where the remainder term $\mathcal{R}(n,\gamma,\|\theta_0-\theta^\star\|) = (1-\mu\gamma)^{n/2}\operatorname{poly}(\gamma,n,\|\theta_0-\theta^\star\|)$.

## Richardson-Romberg Extrapolation

Instead of considering a single SGD trajectory $\{\theta_k^{(\gamma)}\}_{k\in\mathbb{N}}$, construct two parallel chains based on the same sequence $\{\xi_k\}_{k\in\mathbb{N}}$:
$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma\nabla F(\theta_k^{(\gamma)},\xi_{k+1})\,, \quad \bar\theta_n^{(\gamma)} = n^{-1}\sum_{k=n+1}^{2n}\theta_k^{(\gamma)}\,,$$
$$\theta_{k+1}^{(2\gamma)} = \theta_k^{(2\gamma)} - 2\gamma\nabla F(\theta_k^{(2\gamma)},\xi_{k+1})\,, \quad \bar\theta_n^{(2\gamma)} = n^{-1}\sum_{k=n+1}^{2n}\theta_k^{(2\gamma)}\,.$$
Then construct the Richardson-Romberg estimator, which allows to reduce the steady-state bias of SGD:
$$\bar\theta_n^{(RR)} := 2\bar\theta_n^{(\gamma)} - \bar\theta_n^{(2\gamma)}\,.$$

**Theorem 2**. Assume **A1**, **A2**, **A3**(6). Then for $n$ large enough, any initial distribution $\nu$, it holds setting $\gamma=n^{-1/2}$ that
$$\mathsf{E}_\nu^{1/2}[\|\mathrm{H}^\star(\bar\theta_n^{(RR)}-\theta^\star)\|^2] \le \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}}{n^{1/2}} + \frac{\mathsf{C}(\mathrm{L},\mu)}{n^{3/4}} + \mathcal{R}(n,1/\sqrt{n},\|\theta_0-\theta^\star\|)\,,$$
where the remainder term $\mathcal{R}(n,\gamma,\|\theta_0-\theta^\star\|) = (1-\mu\gamma)^{n/2}\operatorname{poly}(\gamma,n,\|\theta_0-\theta^\star\|)$. Here the power $n^{-3/4}$ is the best known among the first-order methods. This result can be generalized to $p$-th moment bounds with $p\ge2$:

**Theorem 3**. Let $p\ge2$ and assume **A1**, **A2**, **A3**($3p$). Then for $n$ large enough, any initial distribution $\nu$, it holds setting $\gamma=n^{-1/2}$ that
$$\mathsf{E}_\nu^{1/p}[\|\mathrm{H}^\star(\bar\theta_n^{(RR)}-\theta^\star)\|^p] \le \frac{c_1\sqrt{\operatorname{Tr}\Sigma_\varepsilon^\star}p^{1/2}}{n^{1/2}} + \frac{\mathsf{C}(\mathrm{L},\mu,p)}{n^{3/4}} + \mathcal{R}(n,1/\sqrt{n},\|\theta_0-\theta^\star\|)\,,$$
where $c_1=60\mathrm{e}$ is an absolute constant. Note that the Richardson-Romberg procedure allows for more aggressive choice of step size ($\gamma$ of order $n^{-1/2}$), compared to $n^{-2/3}$, which corresponds to $\bar\theta_n^{(\gamma)}$.

## References

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

[2] 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.