Gaussian Approximation and Multiplier Bootstrap for Stochastic Gradient Descent

Marina Sheshukova[1], Sergey Samsonov[1] Denis Belomestny[2,1], Eric Moulines[3,4], Qi-Man Shao [5], Zhuo-Song Zhang[5] Alexey Naumov[1,6],

[1]HSE University [2]Duisburg–Essen University [3]CMAP, UMR 7641, École Polytechnique [4]Mohamed Bin Zayed University of AI
[5]Southern University of Science and Technology [6]Steklov Mathematical Institute of the Russian Academy of Sciences

Stochastic Gradient Descent (SGD)

- We aim to estimate $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^d} f(\theta)$ with access only to the noisy gradients $\nabla F(\theta,\xi)$ such that $\nabla f(\theta) = \mathbb{E}_{\xi\sim\mathbb{P}_\xi}[\nabla F(\theta,\xi)]$. Here ξ is a noise variable with the distribution $\mathbb{P}_\xi$. We assume that $\theta^\star$ is the unique minimizer.
- SGD with Polyak–Ruppert (PR) averaging:

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

$$\bar\theta_n = n^{-1}\sum_{k=0}^{n-1}\theta_k\,. \tag{2}$$

- Polyak-Juditsky central limit theorem (see [2]) implies asymptotic normality

$$\sqrt{n}(\bar\theta_n - \theta^\star) \xrightarrow{d} \mathcal{N}(0,\Sigma_\infty), \qquad n\to\infty\,, \tag{3}$$

where the covariance matrix Σ_∞ is minimax-optimal.

Key questions

- What is the rate of convergence in (3)?
- How can (3) be leveraged to construct confidence sets for $\theta^\star$, given that Σ_∞ is unknown in practice?

To quantify convergence rates in (3), we employ convex distance, which is defined for random vectors $X, Y \in \mathbb{R}^d$ as

$\mathsf{d}_{\mathsf{C}}(X,Y) = \sup_{B\in\mathsf{C}(\mathbb{R}^d)}\left|\mathbb{P}(X\in B) - \mathbb{P}(Y\in B)\right|$, where $\mathsf{C}(\mathbb{R}^d)$ denotes the collection of convex subsets of $\mathbb{R}^d$.

Confidence sets based on Multiplier bootstrap procedure

- Let $\mathcal{W}^{n-1} = \{w_\ell\}_{1\le\ell\le n-1}$ be i.i.d. random variables with distribution $\mathbb{P}_w$, each with mean $\mathbb{E}[w_1]=1$ and variance $\mathrm{Var}[w_1]=1$. Assume $\mathcal{W}^{n-1}$ is independent of $\Xi^{n-1} = \{\xi_\ell\}_{1\le\ell\le n-1}$. Procedure is based on perturbing the trajectory (1)(see [1])

$$\begin{aligned}\theta_k^{\mathsf{b}} &= \theta_{k-1}^{\mathsf{b}} - \alpha_k w_k\{\nabla f(\theta_{k-1}^{\mathsf{b}}) + g(\theta_{k-1}^{\mathsf{b}},\xi_k) + \eta(\xi_k)\}\,, \quad k\ge 1\,, \quad \theta_0^{\mathsf{b}} = \theta_0\,,\\ \bar\theta_n^{\mathsf{b}} &= n^{-1}\sum_{k=0}^{n-1}\theta_k^{\mathsf{b}}\,, \quad n\ge 1\,.\end{aligned} \tag{4}$$

Note that, when generating different weights w_k, we can draw samples from the conditional distribution of $\bar\theta_n^{\mathsf{b}}$ given the data Ξ^{n-1}.
- The core principle: "bootstrap world" probabilities $\mathbb{P}\big(\sqrt{n}(\bar\theta_n^{\mathsf{b}} - \bar\theta_n) \in B \mid \Xi^{n-1}\big)$ are close to $\mathbb{P}\big(\sqrt{n}(\bar\theta_n - \theta^\star)\in B\big)$ for $B\in\mathsf{C}(\mathbb{R}^d)$.

Assumptions

A1. The function f is two times continuously differentiable and L_1-smooth on $\mathbb{R}^d$. Moreover, we assume that f is μ-strongly convex on $\mathbb{R}^d$.

A2. For each $k\ge 1$, ζ_k admits the decomposition $\zeta_k = \eta(\xi_k) + g(\theta_{k-1},\xi_k)$, where

1. $\{\xi_k\}_{k=1}^{n-1}$ is a sequence of i.i.d. random variables on $(\mathsf{Z},\mathcal{Z})$ with distribution $\mathbb{P}_\xi$, $\eta:\mathsf{Z}\to\mathbb{R}^d$ is a function such that $\mathbb{E}[\eta(\xi_1)]=0$ and $\mathbb{E}[\eta(\xi_1)\eta(\xi_1)^\top] = \Sigma_\xi$. Moreover, $\lambda_{\min}(\Sigma_\xi) > 0$.
2. The function $g:\mathbb{R}^d\times\mathsf{Z}\to\mathbb{R}^d$ satisfies $\mathbb{E}[g(\theta,\xi_1)]=0$ for any $\theta\in\mathbb{R}^d$. Moreover, there exists $L_2>0$ such that for any $\theta,\theta'\in\mathbb{R}^d$, it holds that

$$\|g(\theta,\xi) - g(\theta',\xi)\| \le L_2\|\theta-\theta'\| \quad \text{and} \quad g(\theta^\star,z) = 0 \;\; \text{for all } z\in\mathsf{Z}\,. \tag{5}$$

3. There exist $C_{1,\xi}, C_{2,\xi} > 0$ such that $\mathbb{P}_\xi$-almost surely that $\|\eta(\xi)\|\le C_{1,\xi}$ and $\sup_\theta\|g(\theta,\xi)\|\le C_{2,\xi}$.

A3. There exist $L_3,\beta>0$ such that for all θ with $\|\theta-\theta^\star\|\le\beta$, it holds

$$\|\nabla^2 f(\theta) - \nabla^2 f(\theta^\star)\| \le L_3\|\theta-\theta^\star\|\,. \tag{6}$$

A4.The stochastic gradient $F(\theta,\xi) := \nabla f(\theta) + g(\theta,\xi) + \eta(\xi)$ is almost surely L_4-co-coercive, that is, for any $\theta,\theta'\in\mathbb{R}^d$, it holds $\mathbb{P}_\xi$-almost surely that

$$L_4\langle F(\theta,\xi) - F(\theta',\xi), \theta-\theta'\rangle \ge \|F(\theta,\xi) - F(\theta',\xi)\|^2\,. \tag{7}$$

Assumptions

A5.There exist constants $0 < W_{\min} < W_{\max} < +\infty$, such that $W_{\min}\le w_1\le W_{\max}$ a.s.
A6.Let $\alpha_k = c_0\{k_0+k\}^{-\gamma}$, where $\gamma\in(1/2,1)$, an c_0 satisfies $c_0 W_{\max}\max(2L_4,\mu)\le 1$ and $k_0 \ge (\frac{2\gamma}{\mu c_0 W_{\min}})^{1/(1-\gamma)}$.
A7.Number of observations n is large enough.

Non-asymptotic multiplier bootstrap validity

Theorem 1. Assume **A1** - **A7**. Then with $\mathbb{P}$ - probability at least $1-1/n$, it holds

$$\sup_{B\in\mathsf{C}(\mathbb{R}^d)}\left|\mathbb{P}(\sqrt{n}(\bar\theta_n^{\mathsf{b}} - \bar\theta_n)\in B\mid\Xi^{n-1}) - \mathbb{P}(\sqrt{n}(\bar\theta_n - \theta^\star)\in B)\right| \le \frac{\mathrm{C}_1\log n}{n^{\gamma-1/2}}\,, \tag{8}$$

where C_1 are some problem-specific constants.

- The proof is based on the gaussian approximation in real and boostrap world.
- Key observation: instead of $\mathcal{N}(0,\Sigma_\infty)$ we use $\mathcal{N}(0,\Sigma_n)$, where Σ_n is the covariance of the linearized recursion which correspond to (1) with additive noise $\eta(\xi_k)$ (see [3] for details).

Rate of convergence in the Polyak–Juditsky central limit theorem

A8(p) Conditions (i) and (ii) from **A2** holds. Moreover, there exists $\sigma_p>0$ such that $\mathbb{E}^{1/p}[\|\eta(\xi_1)\|^p]\le\sigma_p$.
A9 Suppose that $\alpha_k = c_0/(k_0+k)^\gamma$, where $\gamma\in(1/2,1)$, $k_0\ge 1$, and c_0 satisfies $2c_0L_1\le 1$.
Theorem 2. Assume **A1**, **A3**, **A8**(4),**A9**. Then, with $Y\sim\mathcal{N}(0,\mathrm{I}_d)$ it holds that

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar\theta_n - \theta^\star), \Sigma_\infty^{1/2}Y) \le \frac{\mathrm{C}_2}{n^{\gamma-1/2}} + \frac{C_\infty}{n^{1-\gamma}}\,, \tag{9}$$

where C_2 and C_∞ are some problem-specific constants.

Lower bounds

Theorem 3. There exists the problem satisfying conditions **A1**, **A3**, **A8**(4),**A9**, such that with $Y\sim\mathcal{N}(0,\mathrm{I}_d)$ for n large enough it holds that

$$\mathsf{d}_{\mathsf{C}}(\sqrt{n}(\bar\theta_n - \theta^\star), \Sigma_\infty^{1/2}Y) \ge \frac{C_3}{n^{1-\gamma}}\,. \tag{10}$$

The bound (10) implies that $\sqrt{n}(\bar\theta_n - \theta^\star)$ cannot be approximated by $\mathcal{N}(0,\Sigma_\infty)$ faster than $1/n^{1-\gamma}$, and the rate in Equation (9) is tight for $\gamma\in[3/4,1)$. This highlights the necessity of using Σ_n in the bootstrap result (Equation (8)).

Takeaways

- We prove the first non-asymptotic validity result for multiplier bootstrap in averaged SGD, with rate $n^{-(\gamma-1/2)}$ for $\gamma\in(1/2,1)$;
- For strongly convex problems, we show a tight Gaussian approximation rate $n^{-1/4}$ in (3) with $\gamma = 3/4$.

Acknowledgement

This work was supported by the Ministry of Economic Development of the Russian Federation (code 25-139-66879-1-0003).

References

[1] Yixin Fang, Jinfeng Xu, and Lei Yang.
Online bootstrap confidence intervals for the stochastic gradient descent estimator.
Journal of Machine Learning Research, 19(78):1–21, 2018.

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

[3] Marina Sheshukova, Sergey Samsonov, Denis Belomestny, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov.
Gaussian approximation and multiplier bootstrap for stochastic gradient descent.
arXiv preprint arXiv:2502.06719, 2025.