Gaussian Approximation and Multiplier Bootstrap for Stochastic Gradient Descent

Marina Sheshukova

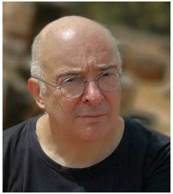
HSE AI institute, HSE University

April 30, 2026

Joint work with

Alexey Naumov
(HSE)

Denis Belomestny
(Duisburg-Essen University & HSE)



Eric Moulines
(LRE EPITA & MBZUAI CMS Division)

Qi-Man Shao
(SUSTech)

Zhuo-Song Zhang
(SUSTech)

Sergey Samsonov
(HSE)

Problem statement

Optimization problem

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

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

SGD recursion

We solve the problem using SGD with decreasing step size α_k, starting from $\theta_0 \in \mathbb{R}^d$:

$$\theta_{k+1} = \theta_k - \alpha_{k+1} F(\theta_k, \xi_{k+1}), \qquad \xi_k \text{ are i.i.d.}$$

Polyak–Ruppert averaging and CLT

For $\bar{\theta}_n = n^{-1} \sum_{k=0}^{n-1} \theta_k$, under appropriate assumptions on f and α_k,

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) \xrightarrow{d} \mathrm{N}(0, \Sigma_\infty), \qquad n \to \infty. \tag{1}$$

This result raises two key question

(i) what is the rate of convergence in (1)
(ii) how can (1) be leveraged to construct confidence sets for $\theta^\star$?

Confidence Sets: Plug-in vs Batch Mean vs Bootstrap

Setup: step size $\alpha_k = c_0 k^{-\gamma}$

Plug-in methods

- Estimate Σ_∞ via $\hat{\Sigma}_n$ using a second-order oracle
- Error: $\delta_n = \mathbb{E}\|\hat{\Sigma}_n - \Sigma_\infty\|$
- Rate [Chen et al., 2020]: $\delta_n \lesssim n^{-\gamma/2}$

Batch mean methods

- Estimate Σ_∞ via $\hat{\Sigma}_n$ using batching
- Error: $\delta_n = \mathbb{E}\|\hat{\Sigma}_n - \Sigma_\infty\|$
- Rate [Zhu et al., 2023]: $\delta_n \lesssim n^{-\frac{1-\gamma}{4}}$

Limitation

Focus on estimating Σ_∞, without taking into account CLT convergence.

Our approach: bootstrap-based method [Fang et al., 2018]

- No estimation of Σ_∞
- Fully non-asymptotic guarantees
- Achieves better approximation rates than plug-in methods for some step sizes α_k

Multiplier Bootstrap for SGD

Bootstrap recursion

$$\theta_k^{\mathrm{b}} = \theta_{k-1}^{\mathrm{b}} - \alpha_k w_k F(\theta_{k-1}^{\mathrm{b}}, \xi_k), \qquad \bar{\theta}_n^{\mathrm{b}} = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k^{\mathrm{b}}$$

where $\{\xi_k\}_{k=1}^{n-1}$ are the same observations as in the original SGD trajectory, and $\{w_\ell\}_{1\le \ell \le n-1}$ are i.i.d. bootstrap weights such that

$$\mathbb{E}[w_1] = 1, \qquad \mathrm{Var}(w_1) = 1.$$

Key idea

Bootstrap probabilities approximate the real ones:

$$\mathbb{P}^{\mathrm{b}}\Big(\sqrt{n}(\bar{\theta}_n^{\mathrm{b}} - \bar{\theta}_n) \in B\Big) \approx \mathbb{P}\big(\sqrt{n}(\bar{\theta}_n - \theta^\star) \in B\big)\,,$$

for convex sets $B \subset \mathbb{R}^d$, where

$$\mathbb{P}^{\mathrm{b}} = \mathbb{P}(\,\cdot \mid \xi_1, \ldots, \xi_{n-1}).$$

In practice

Use Monte Carlo to simulate M perturbed trajectories to approximate $\mathbb{P}^{\mathrm{b}}$.

Bootstrap Validity

Main result

Under suitable assumptions, it holds with $\mathbb{P}$-probability at least $1 - 1/n$ that

$$\sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}^{\mathsf{b}}(\sqrt{n}(\bar{\theta}_n^{\mathsf{b}} - \bar{\theta}_n) \in B) - \mathbb{P}(\sqrt{n}(\bar{\theta}_n - \theta^\star) \in B)| \leq C_1 \left(\frac{d^2 (\log n)^{1/2}}{n^{1/2}} + \frac{(d \log n)^{3/2}}{n^{\gamma - 1/2}} \right),$$

where C_1 is a problem-specific constant.

Main message

We achieve approximation rates of up to $\mathcal{O}(n^{-1/2})$.

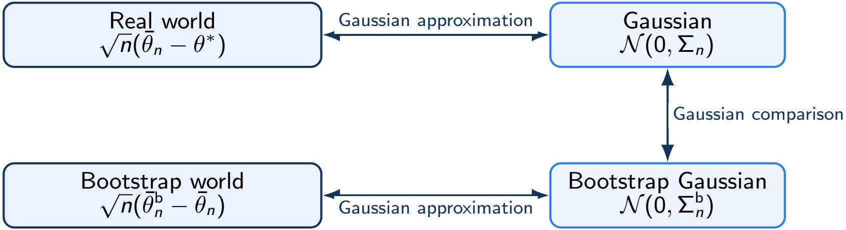
Comparison with previous works

- Samsonov et al. [2024] show the rate $\mathcal{O}(n^{-1/4})$ for LSA problem using $\mathcal{N}(0, \Sigma_\infty)$ for Gaussian approximation when showing bootstrap validity.
- Wu et al. [2024]: Improved to $\mathcal{O}(n^{-1/3})$ using direct estimation of Σ_∞.

Sketch of proof

Main idea

Following Spokoiny and Zhilova [2015], relate the real and bootstrap worlds through Gaussian approximation and Gaussian comparison.



- Use Σ_n from the linearized SGD instead of the limiting covariance Σ_∞.
- Apply Gaussian comparison to control the real–bootstrap gap.

Additional results

- Explicit convergence rates in the CLT; see Theorem 4 in our paper.
- A lower bound for the Gaussian approximation of $\sqrt{n}(\bar{\theta}_n - \theta^\star)$ by $\mathrm{N}(0, \Sigma_\infty)$.

Paper QR Code

Thank you!

I would be happy to discuss more during the poster session.

References I

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. The Annals of Statistics, 48(1):251 – 273, 2020.

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. URL http://jmlr.org/papers/v19/17-370.html.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian Approximation and Multiplier Bootstrap for Polyak-Ruppert Averaged Linear Stochastic Approximation with Applications to TD Learning. In Advances in Neural Information Processing Systems, volume 37, pages 12408–12460. Curran Associates, Inc., 2024.

Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. The Annals of Statistics, 43(6): 2653 – 2675, 2015. doi: 10.1214/15-AOS1355. URL https://doi.org/10.1214/15-AOS1355.

Weichen Wu, Gen Li, Yuting Wei, and Alessandro Rinaldo. Statistical Inference for Temporal Difference Learning with Linear Function Approximation. arXiv preprint arXiv:2410.16106, 2024.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. Journal of the American Statistical Association, 118(541):393–404, 2023.