Bias-Optimal Bounds for SGD

A Computer-Aided Lyapunov Analysis

Daniel Cortild, L. Ketels, J. Peypouquet, G. Garrigos

PGMO Days 2025

EDF Paris, France, November 18th, 2025

Stochastic Gradient Descent

Consider the problem

$$\min \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \colon x \in \mathbb{R}^d \right\},$$

where all $f_i \colon \mathbb{R}^d \to \mathbb{R}$ are convex and L-smooth, and f has minimizers.

Stochastic Gradient Descent (SGD) iterates

$$x_0 \in \mathbb{R}^d, \quad x_{t+1} = x_t - \gamma \nabla f_{i_k}(x_t) \quad \text{for } t = 0, 1, \ldots,$$

where i_k is chosen i.i.d. from the uniform distribution on $\{1, \ldots, n\}$.

Solution Variance is defined as

$$\sigma_*^2 := \mathbb{E}[\|\nabla f_{i_k}(x_*)\|^2] \quad \text{for some } x_* \in \operatorname{argmin} f.$$

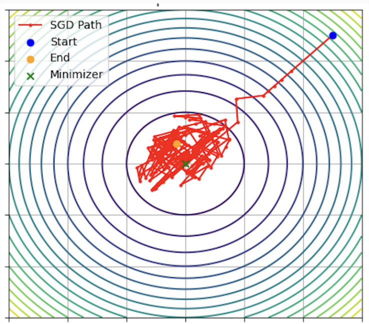
Note this is automatically finite in our setting.

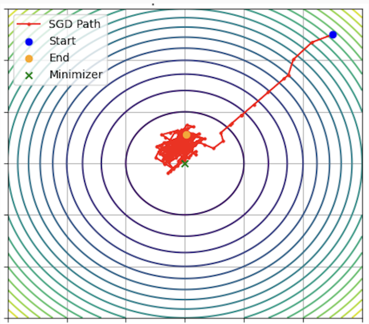
Type of Results for SGD

Convergence results for SGD are usually presented as

$$\text{Performance}(t) \leq \text{Bias}(t) + \text{Variance}(t),$$

where $\text{Bias}(t) \to 0$ as $t \to \infty$, and (ideally) $\text{Variance}(t)$ remains bounded.





Goal: Minimize the bias term first, and the variance term second.

Our Results in the Convex Setting

We obtain a result on the Cesàro average $\overline{x}_T = \frac{x_0 + \cdots + x_{T-1}}{T}$ of the form

$$\mathbb{E}[f(\overline{x}_T) - \min f] \leq \text{Bias}(T) \cdot \|x_0 - x_*\|^2 + \text{Variance}(T) \cdot \sigma_*^2,$$

where

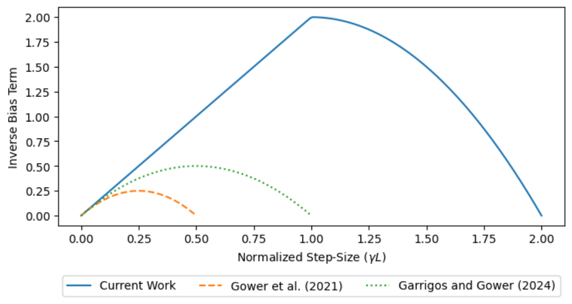
	$\gamma \mathbf{L} \in (\mathbf{0}, \mathbf{1})$	$\gamma \mathbf{L} = \mathbf{1}$	$\gamma \mathbf{L} \in (\mathbf{1}, \mathbf{2})$
Bias(T)	$\frac{1}{2\gamma T + 2(1/L - \gamma)}$	$\frac{1}{(2-\varepsilon)\gamma T}$	$\frac{1-(2-\gamma L)^{2T}}{2\gamma(2-\gamma L)T}$
Variance(T)	$\frac{\gamma}{2(1-\gamma L)}$	$\frac{\gamma(2+\varepsilon)}{\varepsilon(2-\varepsilon)}$	$\frac{\exp(T)}{2-\gamma L}$

Observation 1: Singularity at $\gamma L = 1$ for critical step-size.
Observation 2: No uniform bound in T for $\gamma L > 1$. This can be fixed by slightly hurting the bias.
Observation 3: If $\sigma_*^2 = 0$, these are not problems.

Comparison to State-of-the-Art

Comparison to

- Gower et al. (2021)[1],
- Garrigos and Gower (2024)[2].



[1]Gower, Sebbouh, and Loizou, "SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation", 2021.

[2]Garrigos and Gower, *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*, 2024.

Tightness of Bias and Variance Terms

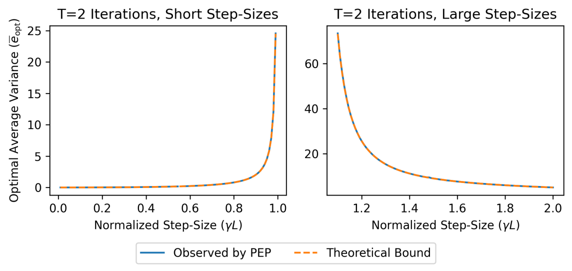
Bias-Optimality.

Our results for $\gamma L \in (0,2)\backslash\{1\}$ are **bias-optimal**, in the sense that there exists a problem that attains that bias:

- If $\gamma L \in (0,1)$; Pick a Huber function $f(x) = \mathcal{H}_\eta(x)$.
- If $\gamma L \in (1,2)$; Pick a quadratic $f(x) = \frac{L}{2}\|x\|^2$.

Variance-Optimality.

Constraint to the optimal bias, our variance is empirically optimal.



Proof Strategy 1/2

Our proofs are based on a Lyapunov analysis with an energy of the form

$$E_t := a_t\|x_t - x_*\|^2 + \rho\sum_{s=0}^{t-1}[f(x_s) - \min f] - \sum_{s=0}^{t-1} e_s\sigma_*^2,$$

where $(a_t), (e_t), \rho \geq 0$.

If we can prove a decrease in energy, namely $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$, then;

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(x_t) - \min f] \leq \frac{a_0}{\rho T}\cdot\|x_0 - x_*\|^2 + \frac{1}{\rho T}\sum_{t=0}^{T-1} e_t\sigma_*^2.$$

We aim at solving

$$\mathrm{Bias}_{\mathrm{opt}}(T) = \inf\left\{\frac{a_0}{\rho T} : (a_t), (e_t), \rho \text{ are Lyapunov parameters}\right\}.$$

$$\inf_{(a_t),(e_t),\rho}\left\{\frac{a_0}{\rho T} : \mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t], \text{ for all convex smooth functions}\right\}$$

Proof Strategy 2/2

$$\inf_{(a_t),(e_t),\rho} \left\{ \frac{a_0}{\rho T} : \mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t],\ \text{for all convex smooth functions} \right\}$$

- Using standard tools from the *Performance Estimation Problem* methodology,[3][4][5] we transform the problem into a finite-dimensional optimization problem.
- This problem may be solved numerically.
- The dual problem of the equivalent SDP provides dual variables that help us inspire the proof.

---

[3] Drori and Teboulle, "Performance of first-order methods for smooth convex minimization: a novel approach", 2014.

[4] Taylor, Hendrickx, and Glineur, "Smooth Strongly Convex Interpolation and Exact Worst-case Performance of First-order Methods", 2017.

[5] Taylor and Bach, "Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions", 2019.

Results in Strongly Convex Setting

We obtain a bound of the form

$$\mathbb{E}[\|x_T - x_*\|^2] \le \mathrm{Bias}(T) \cdot \|x_0 - x_*\|^2 + \mathrm{Variance}(T) \cdot \sigma_*^2,$$

where, for $\varepsilon \ge 0$ arbitrary,

$$\mathrm{Bias}(T) = (\max\{1 - \gamma\mu, \gamma L - 1\}^2 + \varepsilon)^T,$$

and

$$\mathrm{Variance}(T) = \mathcal{O}\left(\gamma^2 + \frac{\gamma^4}{\left|\gamma - \frac{2}{L+\mu}\right| + \varepsilon}\right).$$

Observation: If γ approaches $\frac{2}{L+\mu}$, the variance explodes if $\varepsilon = 0$. To get finite variance at $\gamma = \frac{2}{L+\mu}$, we need to impose $\varepsilon > 0$.

Tightness in the Strongly Convex Setting

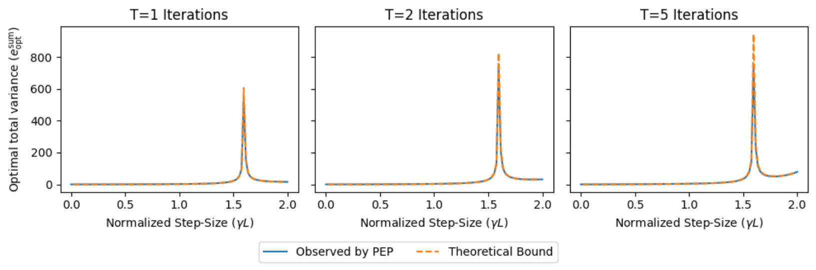
Bias-Optimality.

Our result for $\gamma \in (0,2) \setminus \{\frac{2}{L+\mu}\}$ is **bias-optimal**, in the sense that there exists a problem that attains that bias:

- If $\gamma \in (0, \frac{2}{L+\mu})$; Pick a quadratic $f(x) = \frac{\mu}{2}\|x\|^2$.
- If $\gamma \in (\frac{2}{L+\mu}, 2)$; Pick a quadratic $f(x) = \frac{L}{2}\|x\|^2$.

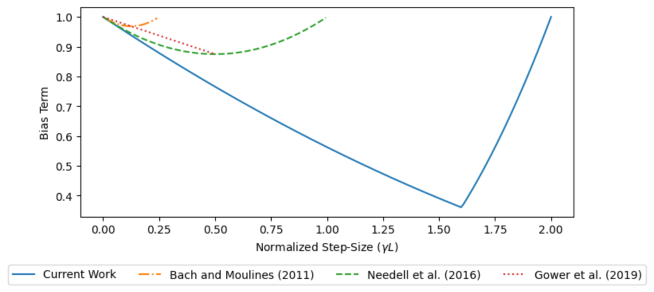
Variance-Optimality.

Constraint to the optimal bias, our variance is empirically optimal.



Results in Strongly Convex Setting

Comparison to state-of-the-art[6,7,8]:



[6] Bach and Moulines, "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning", 2011.

[7] Needell, Srebro, and Ward, "Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm", 2016.

[8] Gower, Loizou, et al., "SGD: General Analysis and Improved Rates", 2019.

Conclusion

- We provided the first study of SGD without variance assumptions for $\gamma L \in (0, 2)$, for convex and strongly convex functions, and improved the current results.
- There is a previously unobserved singularity at critical step-sizes.
- We provided matching lower bounds for the variance term, showing bias-optimality of our results.
- Our proofs are computer-inspired and numerically shown to be tight within our Lyapunov framework.

Based on: Daniel Cortild, Lucas Ketels, Juan Peypouquet, and Guillaume Garrigos. **New Tight Bounds for SGD without Variance Assumption: A Computer-Aided Lyapunov Analysis.** arXiv preprint arXiv:2505.17965. May 2025

Thank you!