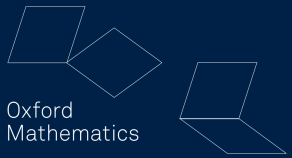
Bias-Optimal Bounds for SGD: A Computer-Aided Lyapunov Analysis

Daniel Cortild, L. Ketels,
J. Peypouquet, G. Garrigos
Mathematical Institute, University of Oxford

SIAM Conference on Optimization - June 5th, 2026



Mathematical
Institute

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) with constant step-size $\gamma > 0$ iterates, for $x_0 \in \mathbb{R}^d$,

$$x_{t+1} = x_t - \gamma \nabla f_{i_k}(x_t) \quad \text{for } t = 0, 1, \ldots,$$

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

Historical Assumption

Uniformly Bounded Variance: $\forall x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\nabla f_{i_k}(x) - \nabla f(x)\|^2\right] \leq \sigma^2 < +\infty.$$

This is unrealistic in practice.

Our "Assumption"

Variance-at-Solution: for some $x_* \in \operatorname{argmin} f$

$$\sigma_*^2 := \mathbb{E}[\|\nabla f_{i_k}(x_*)\|^2] < +\infty.$$

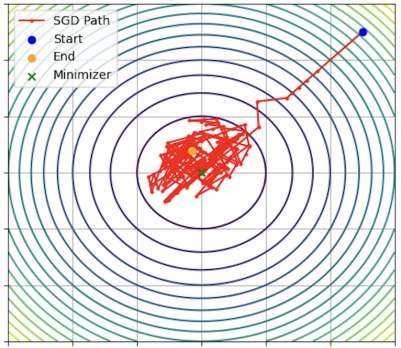
This is automatically finite in our setting.

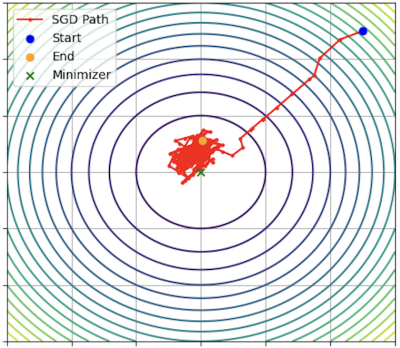
Type of Results for SGD

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)	$\frac{\gamma}{2(1-\gamma L)}$	$\frac{\gamma(2+\varepsilon)}{\varepsilon(2-\varepsilon)}$	$\frac{\exp(T)}{2-\gamma L}$

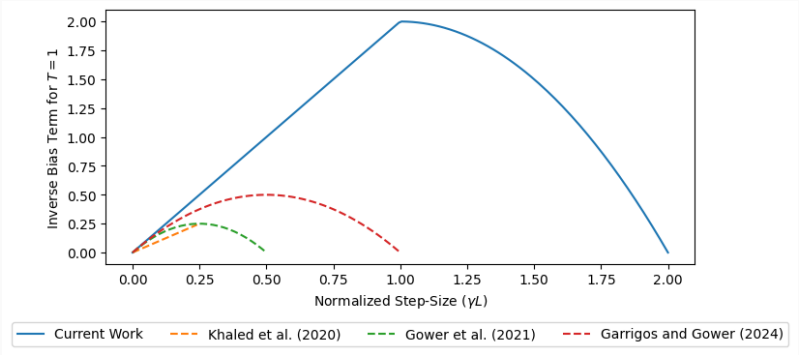
Observation 1: Singularity at $\gamma L = 1$ for critical step-size.

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 Khaled et al. (2020)[1], Gower et al. (2021)[2], and Garrigos and Gower (2024)[3].



[1] Khaled et al., *Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization*, 2020.

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

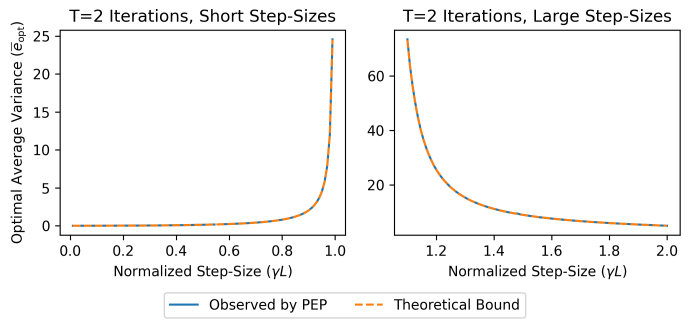
[3] Garrigos and Gower, *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*, 2024.

Tightness of Bias and Variance Terms

Bias-Optimality.
Our results for $\gamma L \in (0,2)\backslash\{1\}$ are **bias-optimal** (there is 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

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

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 ideas from the *Performance Estimation Problem*,[4][5][6] we get a finite-dimensional problem.

Theorem (Interpolation Condition for Convex and Smooth Functions)

For a given set of tuplets $\{(x_i, f_i, g_i)\}$, there exists a convex L-smooth function f such that $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$ if, and only if, a set of quadratic inequalities are satisfied.

- The resulting problem has a tight SDP reformulation, and may be solved numerically.
- The dual variables help us inspire the proof. We provide an analytical proof.

[4] Drori and Teboulle, "Performance of First-Order Methods for Smooth Convex Minimization: A Novel Approach", 2014.

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

[6] 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

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

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

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

and

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

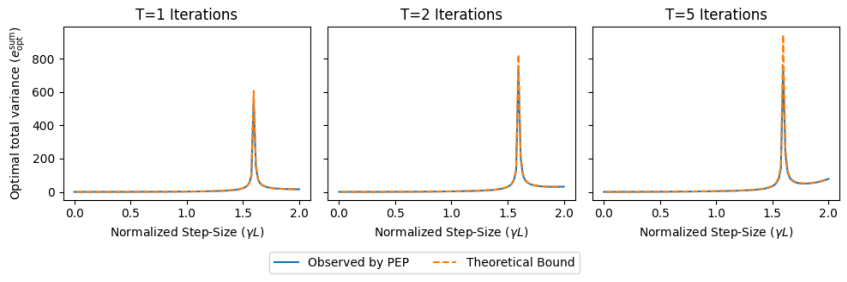
Observation: If γ approaches $\frac{2}{L+\mu}$, the variance explodes if $\varepsilon = 0$. To get finite variance at $\gamma = \frac{2}{L+\mu}$, we need to impose $\varepsilon > 0$.

Tightness in the Strongly Convex Setting

Bias-Optimality.
Our result for $\gamma \in (0,2)\backslash\{\frac{2}{L+\mu}\}$ is **bias-optimal** (there is a problem that attains that bias):

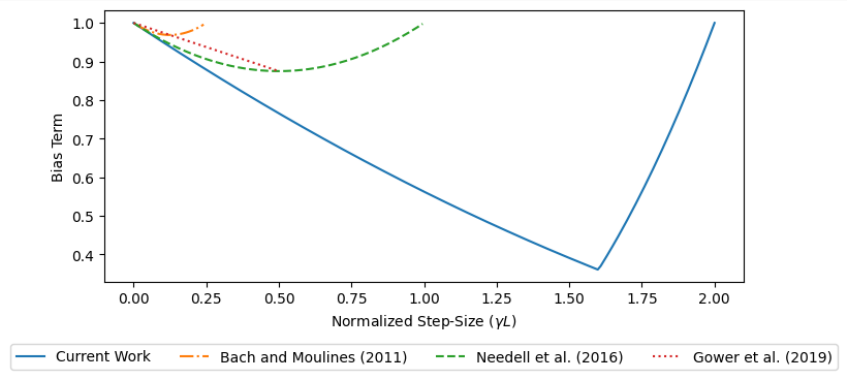
- If $\gamma \in (0, \frac{2}{L+\mu})$; Pick a quadratic $f(x) = \frac{\mu}{2}\|x\|^2$.
- If $\gamma \in (\frac{2}{L+\mu}, 2)$; Pick a quadratic $f(x) = \frac{L}{2}\|x\|^2$.

Variance-Optimality. Constraint to the optimal bias, our variance is empirically optimal.



Results in Strongly Convex Setting

Comparison to state-of-the-art[7][8][9]:



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

[8] Needell, Srebro, and Ward, "Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz Algorithm", 2016.

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

Beyond SGD: Stochastic Krasnoselskii–Mann Iterations[10]

Given $T_\xi \colon \mathcal{H} \to \mathcal{H}$ nonexpansive with $T = \mathbb{E}_{\mathcal{D}}[T_\xi]$, we seek $p \in \mathrm{Fix}(T)$.

Relation to SGD: This reduces to SGD if $T_\xi = I - \gamma \nabla f_\xi$.

Stochastic Krasnoselskii-Mann Iterations:

$$x_{k+1} = \lambda_k x_k + (1 - \lambda_k) T_{\xi_k} x_k, \quad \text{where } \xi_k \sim \mathcal{D}.$$

Variance-at-Solution Assumption:

$$\mathbb{E}_{\mathcal{D}}\left[\|(T - T_\xi)p\|^2\right] \leq \sigma_*^2 < +\infty \quad \text{for some } p \in \mathrm{Fix}(T).$$

Under suitable step-sizes, we can guarantee, almost surely:

1. Weak convergence of the iterates (x_k);
2. Convergence of $\|Tx_k - x_k\|$ to 0, and a $O(\varepsilon^{-4})$ complexity rate of $\mathbb{E}[\|T\hat{x}_k - \hat{x}_k\|]$

[10] Cortild, Cartis, and Peypouquet, *Stochastic Krasnoselskii-Mann Iterations: Convergence without Uniformly Bounded Variance*, 2026.

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.
- We observe a previously unobserved singularity at critical step-sizes.
- We provided matching lower bounds, showing bias-optimality of our results.
- Our proofs are computer-inspired and numerically tight within our Lyapunov framework.

Based on: Daniel Cortild, Lucas Ketels, Juan Peypouquet, and Guillaume Garrigos. **Bias-Optimal Bounds for SGD: A Computer-Aided Lyapunov Analysis.** arXiv preprint arXiv:2505.17965. May 2025. arXiv: 2505.17965 [math]