SGD without Variance Assumption

New Tight Bounds via a Computer-Aided Lyapunov Analysis

Daniel Cortild, **Lucas Ketels**, J. Peypouquet, G. Garrigos

OBI 2 - Dynamics, Optimization and Control

Groningen, Netherlands, June 16th, 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

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

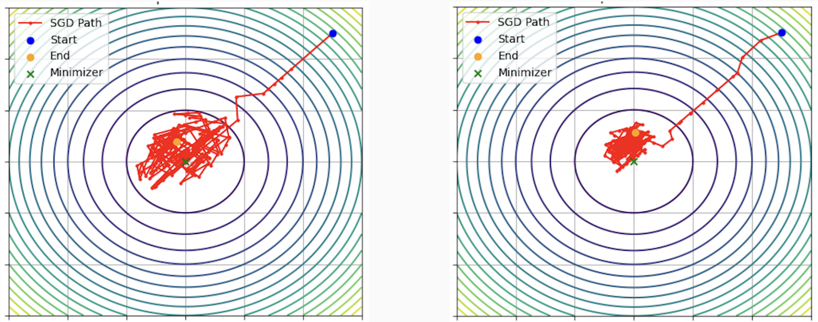
where i_k is chosen i.i.d. from the uniform distribution on $\{1, \ldots, n\}$.

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.

Variance Assumption

Typical Assumption: Uniformly bounded gradient variance;

$$\sup_{x \in \mathbb{R}^d} \mathbb{E}[\|\nabla f_{i_k}(x) - \nabla f(x)\|^2] < +\infty.$$

However, this is unrealistic in practice.[1]

Alternative Assumptions: Weak growth, Strong growth, Maximal strong growth, Relaxed growth, etc.

Our setting: We define

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

Note this is automatically finite in our setting.

[1] Bottou, Curtis, and Nocedal, "Optimization Methods for Large-Scale Machine Learning", 2018.

Results in 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

$$\text{Bias}(T) = \begin{cases} \frac{1}{2\gamma T} & \text{if } \gamma L \in (0,1), \\ \frac{1}{(2-\varepsilon)\gamma T} & \text{if } \gamma L = 1, \varepsilon > 0, \\ \frac{1}{2\gamma(2-\gamma L)T} & \text{if } \gamma L \in (1,2), \end{cases}$$

and

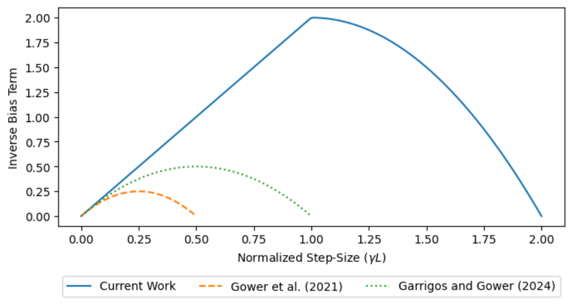
$$\text{Variance}(T) = \begin{cases} \frac{\gamma}{2(1-\gamma L)} & \text{if } \gamma L \in (0,1), \\ \frac{\gamma(2+\varepsilon)}{\varepsilon(2-\varepsilon)} & \text{if } \gamma L = 1, \varepsilon > 0, \\ \frac{\exp(T)}{2-\gamma L} & \text{if } \gamma L \in (1,2). \end{cases}$$

Observation 1: Singularity at $\gamma L = 1$ for optimal step-size.
Observation 2: No uniform bound in T for $\gamma L > 1$.
Observation 3: If $\sigma_*^2 = 0$, these are not problems.

Comparison to State-of-the-Art

Comparison to

- Gower et al. (2021)[2],
- Garrigos and Gower (2024)[3].



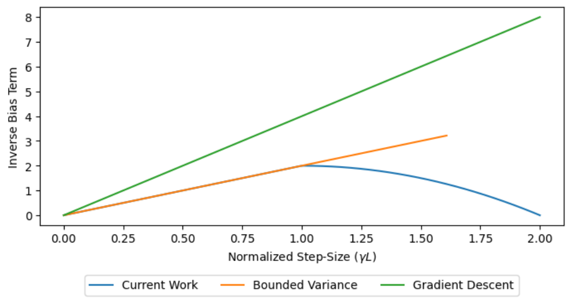
[2] Gower, Sebbouh, and Loizou, "SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation", 2021.

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

Comparison to Other Methods

Comparison to

- SGD with Uniformly Bounded Variance[4],
- Gradient Descent[5].



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

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

Proof Strategy 1/3

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 then aim to minimize $\mathrm{Bias}(T) = \frac{a_0}{\rho T}$.

Proof Strategy 2/3

Obtaining the best bias may then be formulated as

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

- Using standard tools from the *Performance Estimation Problem* methodology,[6][7][8] we transform the problem into a finite-dimensional optimization problem.
- This problem may be solved numerically.

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

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

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

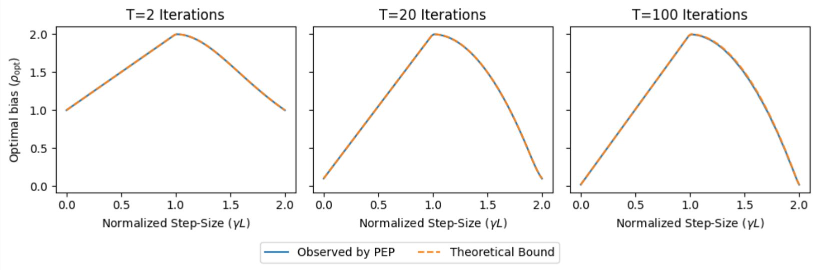
Proof Strategy 3/3

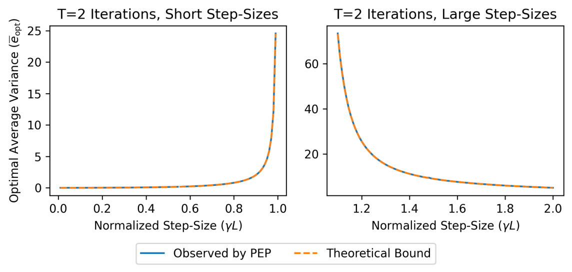
The PEP methodology provides the following:

- An estimation of the minimal bias.
- Numerical values of the coefficients (a_t), (e_t) and ρ.
- Dual variables that help us inspire the proof.

Which helped us getting a *theoretical bias* term $\mathrm{Bias}_{\mathrm{theory}}(T)$.

Tightness of Bias and Variance Terms



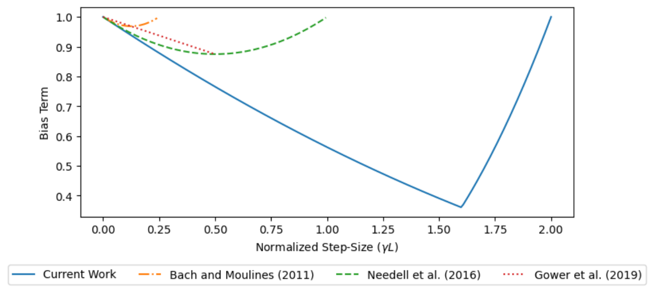


Note: We only claim tightness within our framework.

Results in Strongly Convex Setting

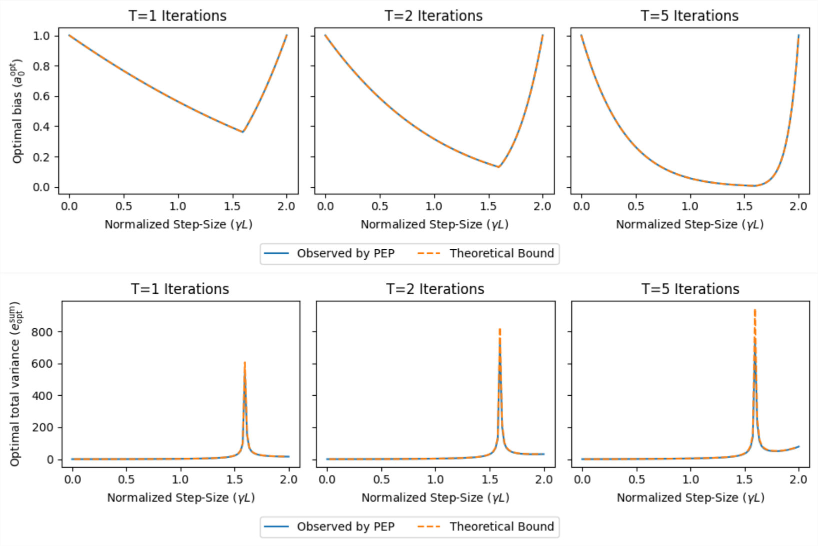
- We obtain improved results for $\gamma L \in (0,2)$,
- The variance remains bounded for all step-sizes,
- There is a similar singularity at the optimal step-size $\gamma_{\text{opt}} = \frac{2}{L+\mu}$.

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



[9] Bach and Moulines, "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning", 2011; Needell, Srebro, and Ward, "Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm", 2016; Gower, Loizou, et al., "SGD: General Analysis and Improved Rates", 2019.

Tightness in the Strongly Convex Setting



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 optimal step-sizes.
- 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!