Improving our understanding of SGD

with Lyapunov energy arguments

Guillaume Garrigos

04 Mai 2025 - SMAI Bordeaux

Université Paris Cité

Introduction: Lyapunov energies

Lyapunov for continuous gradient dynamics

$$\dot{x}(t) + \nabla f(x(t)) = 0$$

A small tour of classical energies (convex smooth case):

- $\|x(t) - x_*\|^2$ whose derivative is $-\langle \nabla f(x(t)), x(t) - x_* \rangle \leqslant 0$
- $f(x(t)) - \inf f$ whose derivative is $-\|\nabla f(x(t))\|^2 \leqslant 0$
- $\|x(t) - x_*\|^2 + t(f(x(t)) - \inf f)$ $\longrightarrow$ gives $O(1/t)$ rates

Second order dynamics (Nesterov, Heavy Ball) may also feature terms like

$$\|\dot{x}(t)\|^2, \quad \langle \nabla f(x(t)), x(t) - x_* \rangle$$

Lyapunov for discrete gradient dynamics

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

A small tour of classical energies (convex smooth case):

- $\|x_t - x_*\|^2$
- $f(x_t) - \inf f$
- $\|x_t - x_*\|^2 + t(f(x_t) - \inf f) \quad \longrightarrow$ gives $O(1/t)$ rates

Second order dynamics (Nesterov, Heavy Ball) may also feature terms like

$$\|x_t - x_{t-1}\|^2, \quad \langle \nabla f(x_t), x_t - x_* \rangle, \quad \langle \nabla f(x_t), x_t - x_{t-1} \rangle$$

Let us now define $\lambda = \lambda_f/m$. We finally obtain the (HBF) system

$$\ddot{x} + \lambda\dot{x} + g\nabla\Phi(x) = 0\,. \tag{HBF}$$

This equation only possesses a mechanical sense when $\dot{x}$ is small.

On the other hand, we can define along every trajectory of (3.1) the energy by

$$E(t) = \frac{1}{2}|\dot{x}(t)|^2 + g\Phi(x(t)).$$

We first observe that (3.1) and the regularity assumptions on Φ automatically imply that $x(\cdot)$ is $\mathcal{C}^2$ on $[0, T_{\max})$. By differentation of $E(t)$, and by using (3.1), we obtain

$$\dot{E}(t) = \langle \dot{x}(t), \ddot{x}(t) + g\nabla\Phi(x(t))\rangle = -\lambda|\dot{x}(t)|^2\,. \tag{3.4}$$

Thus, the function $E(\cdot)$ is decreasing and for all $t \in [0, T_{\max})$,

A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics

F. Alvarez [a], H. Attouch [b,*], J. Bolte [b], P. Redont [b]

Given two parameters $\alpha > 0$ and $\beta > 0$, consider the following second-order in time system in H:

$$\text{(DIN)} \quad \ddot{x} + \alpha\dot{x} + \beta\nabla^2\Phi(x)\dot{x} + \nabla\Phi(x) = 0.$$

Along every trajectory of (DIN) and for $\lambda > 0$ define:

$$E_\lambda(t) = \lambda\Phi\big(x(t)\big) + \frac{1}{2}\big|\dot{x}(t) + \beta\nabla\Phi\big(x(t)\big)\big|^2. \tag{1}$$

Theorem 3.3. *Let us assume that* (h_1)–(h_6) *hold and let* $u \in C([0,\infty[;V) \cap C^1([0,\infty[;H)$ *be a solution of*

$$\frac{d^2u}{dt^2}(t) + \alpha \frac{du}{dt}(t) + Au(t) + f(u(t)) + \varepsilon(t)u(t) = 0, \ t > 0,$$

where $\alpha > 0$ *and* $\varepsilon : [0,\infty[\to [0,\infty[$ *is a given differentiable function such that for all* $t \geq 0$, $\dot{\varepsilon}(t) \leq 0$. *Suppose that the energy*

$$E_{\varepsilon(t)}(t) := \frac{1}{2}\left|\frac{du}{dt}(t)\right|^2 + \frac{1}{2}a(u(t),u(t)) + F(u(t)) + \frac{\varepsilon(t)}{2}|u(t)|^2$$

is absolutely continuous with $\frac{d}{dt}[E(t)] \leq -\alpha|\frac{du}{dt}(t)|^2 + \frac{\dot{\varepsilon}(t)}{2}|u(t)|^2$ *for a.e.* $t > 0$. *Under these conditions, we have* $\frac{du}{dt} \in L^2(0,\infty;H)$, *there exists* $C \geq 0$ *such that* $E_{\varepsilon(t)}(t) \leq C/t$, $\frac{du}{dt}(t) \to 0$ *strongly in* H *and if* $\int_0^\infty \varepsilon(\tau)d\tau = \infty$ *then* $u(t) \to 0$ *strongly in* V *as* $t \to \infty$.

Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions

Vassilis Apidopoulos, Jean-François Aujol, Charles H Dossal, Aude Rondepierre

$$y_n = x_n + \frac{n}{n+b}(x_n - x_{n-1})$$
$$x_{n+1} = T_\gamma(y_n) := y_n - \gamma \nabla F(y_n).$$

$$E_n = \left(t_n^2 + \lambda\beta t_n\right) w_n + \frac{1}{2\gamma}\|\lambda(x_{n-1} - x^*) + t_n(x_n - x_{n-1})\|^2 + \frac{\lambda t_n}{2\gamma}\|x_n - x_{n-1}\|^2 + \frac{\xi}{2\gamma}\|x_{n-1} - x^*\|^2$$

$$w_n = F(x_n) - F(x^*), \quad \delta_n = \|x_n - x_{n-1}\|^2 \quad \text{and} \quad h_n = \|x_n - x^*\|^2.$$

Introduction #2 : SGD and standard results

The problem, the algorithm

Let $f_i : \mathbb{R}^N \to \mathbb{R}$ be convex, and minimize

$$\min_{x \in \mathbb{R}^N} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

with the Stochastic Gradient Descent (SGD) algorithm

$$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t), \quad \gamma_t > 0, \quad i_t \sim \mathcal{U}(1, \dots, m)$$

Rk: You can consider $f(x) = \mathbb{E}_\xi [f(\xi, x)]$ with $\xi \sim \mathcal{D}$ if you want

Rk: You can do minibatches if you want, the story will remain the same

Smooth case: Known results

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

Smooth case: Known results

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

- γ_t can go up to $\frac{1}{2L}$, requires knowing L
- $\sigma_*^2 = 0$ in the deterministic or *interpolation* cases
- We do not assume bounded variance : $\mathbb{V}[\nabla f_i(x_t)] \leqslant C + D\|\nabla f(x_t)\|^2$
- Results also available for the strongly convex regime

Smooth case: Known results

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

- SGD does *not* converge with constant stepsizes (complexity available)
- $\gamma \propto \frac{1}{\sqrt{T}}$ gives a *finite horizon* rate of $O(\frac{D^2+\sigma_*^2}{\sqrt{T}})$, not *adaptive* to σ_*^2
- $\gamma \propto \frac{1}{\sqrt{t}}$ gives a similar but *asymptotic* rate $O(\frac{D^2+\log(T)}{T})$

A Lyapunov argument

Complexity bounds for SGD can be derived from a simple Lyapunov argument!

$$E_t = \|x_t - x_*\|^2 + \gamma \sum_{s=0}^{t-1} f(x_s) - \inf f - 2\gamma t \sigma_*^2.$$

If E_t decreases, then

$$\sum_{s=0}^{T-1} f(x_s) - \inf f - 2\gamma T \sigma_*^2 \leqslant E_T \leqslant E_0 = \|x_0 - x_*\|^2$$

from which we deduce

$$f(\bar{x}_T) - \inf f \leqslant \frac{1}{T} \sum_{s=0}^{T-1} f(x_s) - \inf f \leqslant \frac{\|x_0 - x_*\|^2}{\gamma T} + 2\gamma \sigma_*^2.$$

$\mathbb{E}[E_t]$ decreases for $\gamma L \leqslant 1/4$, see *Garrigos, Gower, Handbook of convergence theorems for (stochastic) gradient methods, 2023. arXiv:2301.11235.*

III : Better bounds with computer help

A work (in progress) in collaboration with

Daniel Cortild
U. Groningen

Lucas Ketels
U. Groningen & UPC

Juan Peypouquet
U. Groningen

New Tight Bounds for SGD without Variance Assumption: A Computer-Aided Lyapunov Analysis, 2025. arXiv:2505.17965.

III : Better bounds with computer help

1 : Exploring the Lyapunov landscape

Better Lyapunov $\Rightarrow$ Better bounds

$$E_t = a_t \|x_t - x_*\|^2 + \rho \sum_{s=0}^{t-1} (f(x_s) - \inf f) - \sum_{s=0}^{t-1} e_s \sigma_*^2.$$

If E_t decreases, then

$$\sum_{s=0}^{T-1} (f(x_s) - \inf f) - \sum_{s=0}^{T-1} e_s \sigma_*^2 \leqslant E_T \leqslant E_0 = \|x_0 - x_*\|^2$$

from which we deduce (note $\bar{e}_T = \frac{1}{T} \sum_{t=0}^{T-1} e_t$)(usually $a_0 = 1$)

$$f(\bar{x}_T) - \inf f \leqslant \frac{1}{T} \sum_{s=0}^{T-1} f(x_s) - \inf f \leqslant \frac{a_0 D^2}{\rho T} + \bar{e}_T \sigma_*^2.$$

PEP: Finding Lyapunovs with SDP

$$E_t = a_t \|x_t - x_*\|^2 + \rho \sum_{s=0}^{t-1} (f(x_s) - \inf f) - \sum_{s=0}^{t-1} e_s \sigma_*^2.$$

$(a_t, \rho, e_t)_{t=0}^{T}$ are *feasible* Lyapunov parameters if $\mathbb{E}[E_t] \searrow$ for every *problem*

PEP: Finding Lyapunovs with SDP

$$E_t = a_t \|x_t - x_*\|^2 + \rho \sum_{s=0}^{t-1} (f(x_s) - \inf f) - \sum_{s=0}^{t-1} e_s \sigma_*^2.$$

$(a_t, \rho, e_t)_{t=0}^T$ are *feasible* Lyapunov parameters if $\mathbb{E}[E_t] \searrow$ for every *problem*

Theorem

There exists a (SDP) feasibility problem such that

$$(a_t, \rho, e_t)_{t=0}^T \text{ are feasible} \Leftrightarrow \text{(SDP) is feasible.}$$

Rk: $O(m^2 + mT)$ variables, $O(T)$ constraints : easy for $m \simeq 2$ and $T \simeq 100$

PEP: Finding Lyapunovs with SDP: How to

$(a_t, \rho, e_t)_{t=0}^{T}$ are feasible $\Leftrightarrow$ (SDP) is feasible.

PEP: Finding Lyapunovs with SDP: How to

$(a_t, \rho, e_t)_{t=0}^{T}$ are feasible $\Leftrightarrow$ (SDP) is feasible.

- Main idea # 1: reformulate all constraints $f \in \Gamma_0(\mathbb{R}^N) \cap C^{1,1}(\mathbb{R}^N)$

$$\Leftrightarrow \begin{cases} f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geqslant \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \\ f_y - f_x - \langle g_x, y - x \rangle \geqslant \frac{1}{2L} \|g_y - g_x\|^2, \quad f_y, f_x \in \mathbb{R}, g_x \in \mathbb{R}^N \end{cases}$$

Proposed by Drori, Teboulle (2014), made formal by Taylor et al (2017)

PEP: Finding Lyapunovs with SDP: How to

$(a_t, \rho, e_t)_{t=0}^{T}$ are feasible $\Leftrightarrow$ (SDP) is feasible.

- Main idea # 1: reformulate all constraints $f \in \Gamma_0(\mathbb{R}^N) \cap C^{1,1}(\mathbb{R}^N)$
- Main idea # 2: replace iterates x_t with any point
  Idea from Taylor, Bach (2019)

PEP: Finding Lyapunovs with SDP: How to

$(a_t, \rho, e_t)_{t=0}^{T}$ are feasible $\Leftrightarrow$ (SDP) is feasible.

- Main idea # 1: reformulate all constraints $f \in \Gamma_0(\mathbb{R}^N) \cap C^{1,1}(\mathbb{R}^N)$
- Main idea # 2: replace iterates x_t with any point
- Main idea # 3: quadratic constraints $\Rightarrow$ SDP with Gram matrices
  Relaxation is exact because dimension N is as large as we want

III : Better bounds with computer help

2 : Our results

SGD with short step-sizes

Theorem (Short step-size)

If $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\gamma L < 1$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{2\gamma T} + \frac{\gamma \sigma_*^2}{2(1-\gamma L)}.$$

- First bounds for $\gamma L \geqslant 1/2$ with no variance assumption
- With variance assumptions, bias term is the "same"
- Numerically : sharp
- Variance explodes when $\gamma L = 1$?? Need to relax the bias

SGD with optimal step-size

Theorem (Optimal step-size)

If $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\gamma L = 1$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{(2-\varepsilon)\gamma T} + \frac{\gamma(2+\varepsilon)\sigma_*^2}{\varepsilon(2-\varepsilon)}, \qquad \text{for } \varepsilon \in (0,2).$$

- First bound for $\gamma L = 1$ with no variance assumption
- Variance explodes when $\varepsilon \to 0$. Topology of feasible parameters?
- $\varepsilon = 0$ ok if interpolation holds

SGD with large step-sizes

Theorem (Large step-size)

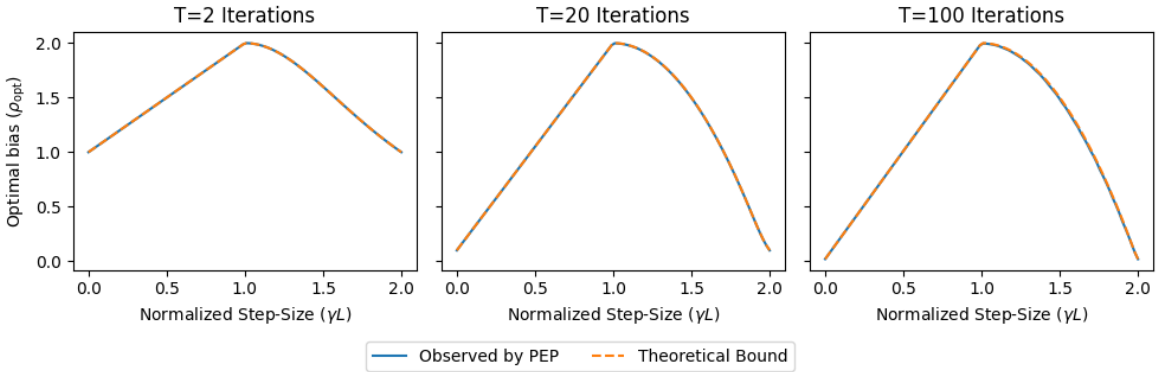
If $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\gamma L \in (1, 2)$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{2\gamma(2-\gamma L)T} + \frac{\gamma e^{O(T)}\sigma_*^2}{2(2-\gamma L)^3}.$$

- Variance explodes when $T \to +\infty$
- Numerics suggest this cannot be avoided
- Big difference with "bounded variance" settings !!

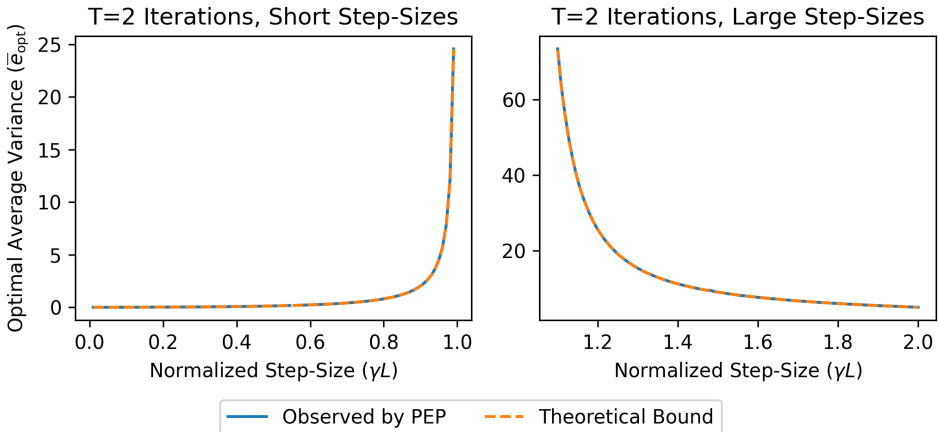
Numerical validation with PEP

Bias term is sharp within this Lyapunov framework



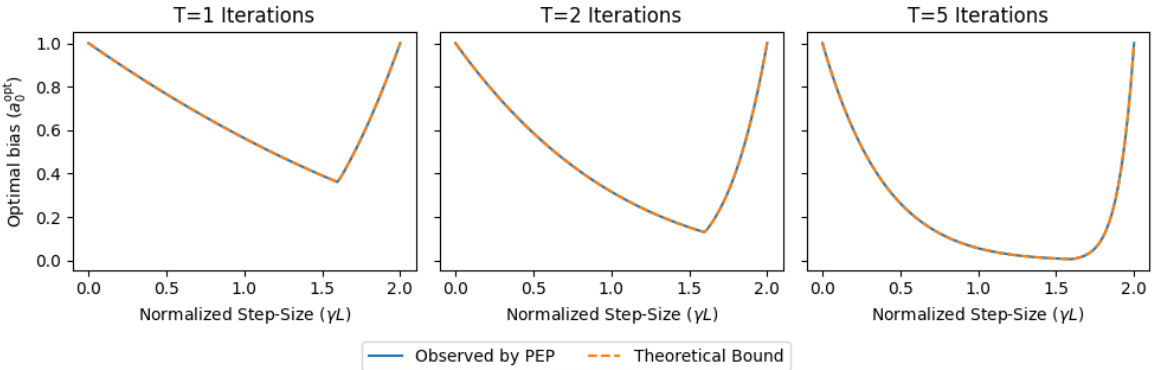
Numerical validation with PEP

Variance term is sharp once bias term is fixed



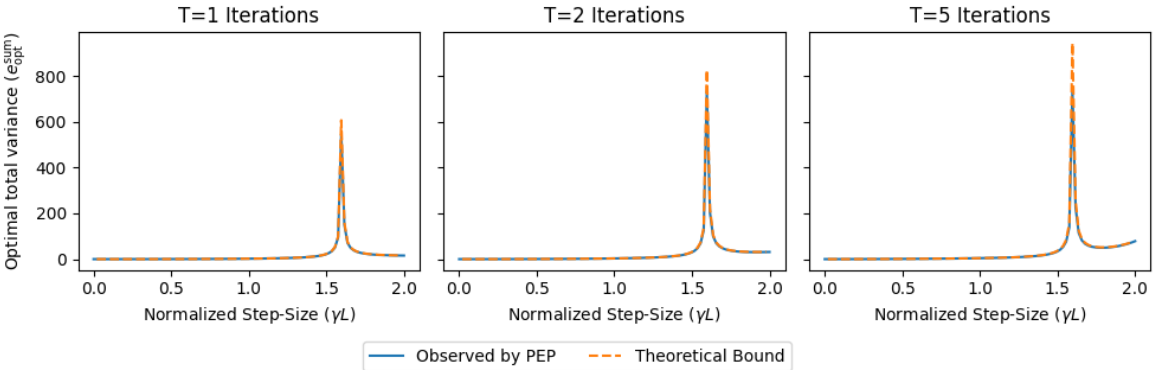
Numerical validation with PEP

We also have rates for the strongly convex case!



Numerical validation with PEP

We also have rates for the strongly convex case! With again a singularity!



Take-home messages

- Bounds available for full rage $\gamma L \in (0, 2)$
- Tricky things happen starting from $1/L$
- Having a bias as good as the best GD one seems impossible
- Considering other Lyapunov elements seem to not improve results
- Results seem invariant with number of functions m

Thanks for your attention !

Any questions?