A systematic approach to Lyapunov analyses of continuous-time models in convex optimization versions

Céline Moucer

PEP Talks, February 2023

Joint work with

Adrien Taylor

Francis Bach

Motivations

- A principled approach to worst-case analysis to continuous-time limit of optimization methods
- A tool for constructing suitable Lyapunov functions for ODEs and SDEs
- A simple insight to what can be expected from (stochastic) optimization methods

First-order methods in convex optimization

A very popular setting:

$$f(x_\star) = \min_{x \in \mathbf{R}^d} f(x),$$

where f is convex, differentiable, and $x_\star \in \mathbf{R}^d$ an optimal point.

- **First-order methods**: low-cost per iteration, accuracy is not critical (machine learning, signal processing, etc.)

$$x_{k+1} \in \mathbf{Span}(x_0, \nabla f(x_0), ..., \nabla f(x_{k+1}))$$

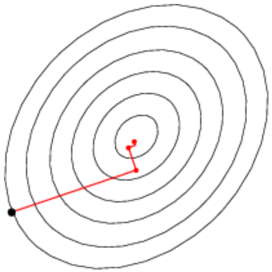


Figure: Convex function and optimization algorithm

First-order methods in convex optimization

Gradient descent with fixed step size $\gamma > 0$:

$$x_{k+1} = x_k - \gamma \nabla f(x_k).$$

First-order methods in convex optimization

Gradient descent with fixed step size $\gamma > 0$:

$$x_{k+1} = x_k - \gamma \nabla f(x_k).$$

- **Ordinary differential equations (ODEs)**: When taking the step size γ to 0, it is directly related to the gradient flow,

$$\dot{X}_t = -\nabla f(X_t), \; X_0 = x_0 \in \mathbf{R}^d,$$

where X_t verifies $X_{t_k} \approx x_k$ with the identification $t_k = \gamma k$.

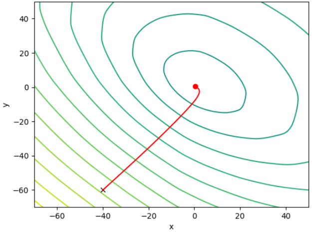


Figure: Integration of the gradient flow for a logistic regression problem.

Optimization methods and ODEs: convergence guarantees

- **First-order methods**: given a class of functions $\mathcal{F}$, a starting point $x_0 \in \mathbf{R}^d$, and given gradient descent with step size $\gamma > 0$

$$x_{k+1} = x_k - \gamma \nabla f(x_k),$$

the goal is to quantify the convergence speed to an optimum $x_\star$ in a small number of steps k,

$$\|x_k - x_\star\|^2 \leqslant \tau(k, \mathcal{F}, \gamma)\|x_0 - x_\star\|^2.$$

Optimization methods and ODEs: convergence guarantees

- **First-order methods**: given a class of functions $\mathcal{F}$, a starting point $x_0 \in \mathbf{R}^d$, and given gradient descent with step size $\gamma > 0$

$$x_{k+1} = x_k - \gamma \nabla f(x_k),$$

the goal is to quantify the convergence speed to an optimum $x_\star$ in a small number of steps k,

$$\|x_k - x_\star\|^2 \leqslant \tau(k, \mathcal{F}, \gamma)\|x_0 - x_\star\|^2.$$

- **ODEs** : given a class of function $\mathcal{F}$, a starting point $x_0 \in \mathbf{R}^d$, the gradient flow starting is given by,

$$\frac{d}{dt} X_t = -\nabla f(X_t),$$

the goal is to quantify the convergence speed to an $x_\star$,

$$\|X_t - x_\star\|^2 \leqslant \tau(t, \mathcal{F})\|x_0 - x_\star\|^2.$$

Convex optimization setting

Common assumptions:

- f is convex and differentiable,
- A differentiable function f is L**-smooth** if and only if it satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

- A convex differentiable function f is μ**-strongly convex** if and only if it satisfies

$$\|\nabla f(x) - \nabla f(y)\| \geqslant \mu\|x - y\|.$$

$\mathcal{F}_{\mu,L}$ is the family of a L-smooth μ-strongly convex functions, with $0 \leq \mu \leq L \leq +\infty$.

Performance estimation problems (PEPs)

Main ideas:

1. Optimization methods and associated ODEs are usually studied via worst-case analyses.
2. Convergence proofs are combinations of inequalities (from methods and problem class).
3. Automated search for combinations of inequalities.

References:

- Initiated by Drori and Teboulle (2012) [2]
- Analyses of first-order methods and design of proofs by Taylor et al. (2017) [10]

An example: the gradient flow

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from differentiable functions f:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

An example: the gradient flow

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from differentiable functions f:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Lyapunov functions: given a trajectory X_t, many proofs construct a Lyapunov function $\mathcal{V} : x, t \in \mathbf{R}^d, \mathbf{R}^+ \to \mathbf{R}$, such that,

1. $\mathcal{V}(x, t) = 0 \iff x = x_\star$,
2. $\mathcal{V}(X_t, t) \geqslant 0$,
3. $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$.

An example: the gradient flow

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from differentiable functions f:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Lyapunov functions: given a trajectory X_t, many proofs construct a Lyapunov function $\mathcal{V} : x, t \in \mathbf{R}^d, \mathbf{R}^+ \to \mathbf{R}$, such that,

1. $\mathcal{V}(x, t) = 0 \iff x = x_\star$,
2. $\mathcal{V}(X_t, t) \geqslant 0$,
3. $\frac{d}{dt}\mathcal{V}(X_t, t) \leqslant 0$.

For example, let us consider the function $\mathcal{V}(X_t, t) = f(X_t)$:

$$\frac{d}{dt}\mathcal{V}(X_t, t) = \dot{X_t}^T \nabla f(X_t) = -\|\nabla f(X_t))\|^2 \leqslant 0.$$

Worst-case guarantee using Lyapunov functions

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Worst-case guarantee: given a Lyapunov function $\mathcal{V}$, we look for (the largest) values $\tau(\mu) \geqslant 0$ such that

$$\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\tau(\mu)\mathcal{V}(X_t),$$

is true for all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t.

Worst-case guarantee using Lyapunov functions

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Worst-case guarantee: given a Lyapunov function $\mathcal{V}$, we look for (the largest) values $\tau(\mu) \geqslant 0$ such that

$$\frac{d}{dt} \mathcal{V}(X_t) \leqslant -\tau(\mu) \mathcal{V}(X_t),$$

is true for all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t.

Integrating between 0 and t: $\mathcal{V}(X_t) \leqslant e^{-\tau(\mu)t} \mathcal{V}(x_0)$.

Worst-case guarantee using Lyapunov functions

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Worst-case guarantee: given a Lyapunov function $\mathcal{V}$, we look for (the largest) values $\tau(\mu) \geqslant 0$ such that

$$\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\tau(\mu)\mathcal{V}(X_t),$$

is true for all functions $f \in \mathcal{F}_{\mu,\infty}$, and all trajectories X_t.

Integrating between 0 and t: $\mathcal{V}(X_t) \leqslant e^{-\tau(\mu)t}\mathcal{V}(x_0)$.

Reformulation as an optimization problem:

$$\begin{aligned} -\tau(\mu) = \max_{X_t \in \mathbf{R}^d,\ f \in \mathcal{F}_{\mu,\infty}} & \frac{d}{dt}\mathcal{V}(X_t), \\ \text{subject to } & \mathcal{V}(X_t) = 1, \\ & \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

Worst-case guarantee using Lyapunov functions

We consider the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from strongly convex functions $f \in \mathcal{F}_{\mu,\infty}$:

$$\frac{d}{dt} X_t = -\nabla f(X_t).$$

Given the Lyapunov function $\mathcal{V}(X_t) = f(X_t) - f_\star$,

$$\begin{aligned} -\tau(\mu) = \max_{X_t,\, f \in \mathcal{F}_{\mu,\infty}} & \ \dot{X}_t^T \nabla f(X_t), \\ \text{subject to } & \ f(X_t) - f_\star = 1, \\ & \ \dot{X}_t = -\nabla f(X_t). \end{aligned}$$

This infinite dimensional problem can be reformulated as an SDP.

A reformulation into an SDP

Formulation into an SDP,

$$\begin{aligned}\max_{G\succeq 0, F} & \operatorname{Tr}(A_0 G),\\ \text{subject to } & b_0^T F = 1,\\ & b_1^T F + \operatorname{Tr}(A_1 G) \geqslant 0,\\ & b_2^T F + \operatorname{Tr}(A_2 G) \geqslant 0,\end{aligned}$$

where $A_0 = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$, $A_1 = \begin{pmatrix} -\mu/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, $A_2 = \begin{pmatrix} -\mu/2 & 0 \\ 0 & 0 \end{pmatrix}$, $b_1 = -1$ and $b_2 = b_0 = 1$, and $F = f_t - f_\star$, $G = \begin{pmatrix} \|X_t - x_\star\|^2 & \langle X_t - x_\star, g_t\rangle \\ \langle X_t - x_\star, g_t\rangle & \|g_t\|^2 \end{pmatrix} \succeq 0$ is a Gram matrix.

Linear SDP → can be solved numerically.

Generalization to a family of Lyapunov functions

Given the gradient flow, it is reasonable to search for **quadratic Lyapunov functions**, for $a, c \geqslant 0$:

$$\mathcal{V}_{a,c}(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2.$$

Generalization to a family of Lyapunov functions

Given the gradient flow, it is reasonable to search for **quadratic Lyapunov functions**, for $a, c \geqslant 0$:

$$\mathcal{V}_{a,c}(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2.$$

Goal: verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau \mathcal{V}_{a,c}(X_t)$, is satisfied for all $d \in \mathbf{N}$, for all $f \in \mathcal{F}_{\mu,\infty}$ and all X_t solutions to the gradient flow.

Generalization to a family of Lyapunov functions

Given the gradient flow, it is reasonable to search for **quadratic Lyapunov functions**, for $a, c \geqslant 0$:

$$\mathcal{V}_{a,c}(X_t) = a \cdot (f(X_t) - f_\star) + c \cdot \|X_t - x_\star\|^2.$$

Goal: verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a,c}(X_t) \leqslant -\tau \mathcal{V}_{a,c}(X_t)$, is satisfied for all $d \in \mathbf{N}$, for all $f \in \mathcal{F}_{\mu,\infty}$ and all X_t solutions to the gradient flow.

It is equivalent with **the existence of** $\lambda_1, \lambda_2 \geqslant 0$ **such that:**

$$S = \begin{pmatrix} \tau c - \frac{\mu}{2}(\lambda_1 + \lambda_2) & -c + \frac{\lambda_1}{2} \\ -c + \frac{\lambda_1}{2} & -a \end{pmatrix} \preccurlyeq 0, \ \tau a = \lambda_1 - \lambda_2.$$

This is a **Linear Matrix Inequality (LMI)**, that allows to **verify a Lyapunov function.**

Numerical VS known bounds

A numerical bound:

The LMI is jointly convex in $\lambda_1, \lambda_2, a, c$ and linear in τ. A **bisection search** allows to optimize over τ and a, c at the same time.

A closed-form upper bound in the worst-case:

Lemma

Let f be a μ-strongly convex function, $x_0 \in \mathbf{R}^d$, and $x_\star$ the minimizer of f. The solution X_t to the gradient flow verifies

$$\frac{d}{dt}\left(f(X_t) - f(x_\star)\right) \leqslant -2\mu\left(f(X_t) - f(x_\star)\right),$$

and after integrating between 0 *and* t, $f(X_t) - f(x_\star) \leqslant e^{-2\mu t}(f(x_0) - f(x_\star))$.

Numerical VS known upper bound: gradient flow

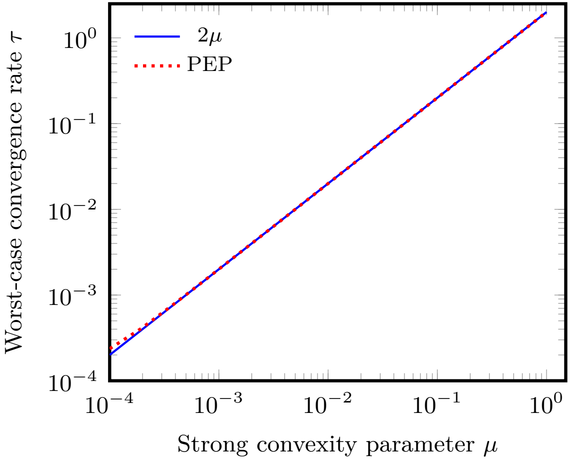


Figure: Worst-case rate $\tau_\star$ for the class of quadratic Lyapunov functions

.

Gradient flow originating from convex functions

Let X_t be the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from **convex functions** $f \in \mathcal{F}_{0,\infty}$, worst-case convergence guarantees are often **sublinear**. Typically:

$$f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t}.$$

Gradient flow originating from convex functions

Let X_t be the **gradient flow** starting from $x_0 \in \mathbf{R}^d$, and originating from **convex functions** $f \in \mathcal{F}_{0,\infty}$, worst-case convergence guarantees are often **sublinear**. Typically:

$$f(X_t) - f_\star \leqslant \frac{\|x_0 - x_\star\|^2}{2t}.$$

A corresponding Lyapunov function is given by:

$$\mathcal{V}(X_t, t) = t(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2.$$

proof: $\frac{d}{dt}\mathcal{V}(X_t, t) = t\langle \nabla f(X_t), \dot{X}_t\rangle + f(X_t) - f_\star + \langle \dot{X}_t, X_t - x_\star\rangle =$
$-t\|\nabla f(X_t)\|^2 + f(X_t) - f_\star - \langle \nabla f(X_t), X_t - x_\star\rangle \leqslant -t\|\nabla f(X_t)\|^2 \leqslant 0$, using convexity.

A time-dependent Lyapunov function

Let us adapt the techniques by considering **quadratic Lyapunov functions:**

$$\mathcal{V}_{a_t,c_t}(X_t, t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2,$$

where $c_t, a_t \geqslant 0$ are functions differentiable with respect to time such that the function $\mathcal{V}_{a_t,c_t}$ verifies:

- $\mathcal{V}_{a_t,c_t}(X_t, t) \geqslant 0$,
- $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t, t) \leqslant 0$.

A time-dependent Lyapunov function

Let us adapt the techniques by considering **quadratic Lyapunov functions:**

$$\mathcal{V}_{a_t,c_t}(X_t, t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2,$$

where $c_t, a_t \geqslant 0$ are functions differentiable with respect to time such that the function $\mathcal{V}_{a_t,c_t}$ verifies:

- $\mathcal{V}_{a_t,c_t}(X_t, t) \geqslant 0$,
- $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t, t) \leqslant 0$.

After integrating between 0 and t, a convergence guarantee in function values is given by

$$f(X_t) - f_\star \leqslant \frac{\mathcal{V}_{a_0,c_0}(x_0, 0)}{a_t} = \frac{a_0(f(x_0) - f_\star) + c_0\|x_0 - x_\star\|^2}{a_t}.$$

A time-dependent Lyapunov function

Let us adapt the techniques by considering **quadratic Lyapunov functions:**

$$\mathcal{V}_{a_t,c_t}(X_t,t) = a_t(f(X_t) - f_\star) + c_t\|X_t - x_\star\|^2,$$

where $c_t, a_t \geqslant 0$ are functions differentiable with respect to time such that the function $\mathcal{V}_{a_t,c_t}$ verifies:

- $\mathcal{V}_{a_t,c_t}(X_t,t) \geqslant 0,$
- $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t) \leqslant 0.$

After integrating between 0 and t, a convergence guarantee in function values is given by

$$f(X_t) - f_\star \leqslant \frac{\mathcal{V}_{a_0,c_0}(x_0,0)}{a_t} = \frac{a_0(f(x_0) - f_\star) + c_0\|x_0 - x_\star\|^2}{a_t}.$$

Remark

The strongly convex case as defined above is a particular case of the convex one*, using a specific Lyapunov function $\Phi(\cdot)$, such that $\mathcal{V}(X_t,t) = e^{\tau t}\Phi(X_t)$. Then,*

$$\frac{d}{dt}\mathcal{V}(X_t,t) \leqslant 0 \iff \frac{d}{dt}\Phi(X_t) \leqslant -\tau\Phi(X_t).$$

A differential LMI

Verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t,t) \leqslant 0$, is satisfied for all $d \in \mathbf{N}$, all $f \in \mathcal{F}_{0,\infty}$ and all X_t generated by the gradient flow, is equivalent with **the existence of** $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ **such that:**

$$S = \begin{pmatrix} \dot{c}_t & -c_t + \frac{\lambda_t^{(1)}}{2} \\ -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preccurlyeq 0, \;\; \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

[1] see implementation in PEPit [3]

A differential LMI

Verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a_t,c_t}(X_t, t) \leqslant 0$, is satisfied for all $d \in \mathbf{N}$, all $f \in \mathcal{F}_{0,\infty}$ and all X_t generated by the gradient flow, is equivalent with **the existence of** $\lambda_t^{(1)}, \lambda_t^{(2)} \geqslant 0$ **such that:**

$$S = \begin{pmatrix} \dot{c}_t & -c_t + \frac{\lambda_t^{(1)}}{2} \\ -c_t + \frac{\lambda_t^{(1)}}{2} & -a_t \end{pmatrix} \preccurlyeq 0, \;\; \dot{a}_t = \lambda_t^{(1)} - \lambda_t^{(2)}.$$

- Choosing $\lambda_t^{(1)} = 1$, $\lambda_t^{(2)} = 0$, together with $c_t = \frac{1}{2}$ and $a_t = t$, $\mathcal{V}(x,t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2$ is **a feasible point of the LMI.**
- A problem that is jointly convex in $\lambda_t^{(1)}$, $\lambda_t^{(2)}$, c_t, a_t, $\dot{a}_t$, $\dot{c}_t$, allowing numerical verification[1].

[1] see implementation in PEPit [3]

Accelerated methods and higher-order gradient flows

An accelerated gradient method [2],

$$x_{k+1} = y_k - \gamma \nabla f(y_k),$$
$$y_{k+1} = x_{k+1} + \alpha_k (x_{k+1} - x_k),$$

where $\gamma, \alpha_k \geqslant 0$ depend on the class of functions to minimize.

This method happens to be closely related to

- **Polyak damped oscillator** [3](strongly convex functions)

$$\ddot{X}_t + 2\sqrt{\mu}\dot{X} + \nabla f(X_t) = 0, \quad (\text{conv. in } \mathcal{O}(e^{-\sqrt{\mu}t})),$$

- **Nesterov's accelerated gradient flow** [4] (convex functions)

$$\ddot{X}_t + \frac{3}{t}\dot{X} + \nabla f(X_t) = 0, \quad (\text{conv. in } \mathcal{O}(\frac{1}{t^2})).$$

---

[2] Nesterov, [5]

[3] introduced by Polyak in [6]

[4] see Su et al. [9, Theorem 3]

Higher-order and non-autonomous gradient flows

More generally, we study **non-autonomous second-order gradient flows**, for $\beta_t \geqslant 0$:

$$\ddot{X}_t + \beta_t \dot{X} + \nabla f(X_t) = 0,$$

with a family of quadratic Lyapunov functions, where a_t, P_t are differentiable functions:

$$\mathcal{V}_{a_t,P_t}(X_t, t) = a_t(f(X_t) - f_\star) + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top (P_t \otimes I_d) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}.$$

After integration between 0 and t, it leads to a convergence guarantee in function values

$$f(X_t) - f_\star \leqslant \frac{\mathcal{V}(x_0)}{a_t}.$$

Polyak's damped oscillator

Let $f \in \mathcal{F}_{\mu,\infty}$. Given the Polyak damped oscillator

$$\ddot{X}_t + 2\sqrt{\mu}\dot{X} + \nabla f(X_t) = 0,$$

Verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a,P}(X_t) \leqslant -\tau \mathcal{V}_{a,P}(X_t)$, is satisfied for all $d \in \mathbf{N}$, all $f \in \mathcal{F}_{\mu,\infty}$ and all X_t is equivalent with the existence of $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$ such that

$$\begin{pmatrix} -\frac{\mu}{2}(\lambda_1+\lambda_2)+\tau p_{11} & p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & -p_{12}+\frac{\lambda_1}{2} \\ p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & 2(p_{12}-2\sqrt{\mu}p_{22})+\tau p_{22} & -p_{22}+\frac{a}{2} \\ -p_{12}+\frac{\lambda_1}{2} & -p_{22}+\frac{a}{2} & 0 \end{pmatrix} \preccurlyeq 0,$$

$$\tau a = \lambda_1 - \lambda_2,$$

Polyak's damped oscillator

Let $f \in \mathcal{F}_{\mu,\infty}$. Given the Polyak damped oscillator

$$\ddot{X}_t + 2\sqrt{\mu}\dot{X} + \nabla f(X_t) = 0,$$

Verifying that the inequality $\frac{d}{dt}\mathcal{V}_{a,P}(X_t) \leqslant -\tau \mathcal{V}_{a,P}(X_t)$, is satisfied for all $d \in \mathbf{N}$, all $f \in \mathcal{F}_{\mu,\infty}$ and all X_t is equivalent with the existence of $\lambda_1, \lambda_2, \nu_1, \nu_2 \geqslant 0$ such that

$$\begin{pmatrix} -\frac{\mu}{2}(\lambda_1+\lambda_2)+\tau p_{11} & p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & -p_{12}+\frac{\lambda_1}{2} \\ p_{11}-2\sqrt{\mu}p_{12}+\tau p_{12} & 2(p_{12}-2\sqrt{\mu}p_{22})+\tau p_{22} & -p_{22}+\frac{a}{2} \\ -p_{12}+\frac{\lambda_1}{2} & -p_{22}+\frac{a}{2} & 0 \end{pmatrix} \preccurlyeq 0,$$

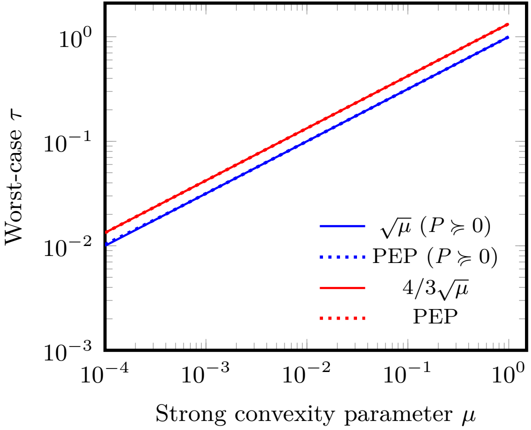
$$\tau a = \lambda_1 - \lambda_2,$$

$$\begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{\mu}{2}(\nu_1+\nu_2) & 0 & \frac{-\nu_1}{2} \\ 0 & 0 & 0 \\ \frac{-\nu_1}{2} & 0 & 0 \end{pmatrix} \succcurlyeq 0,$$

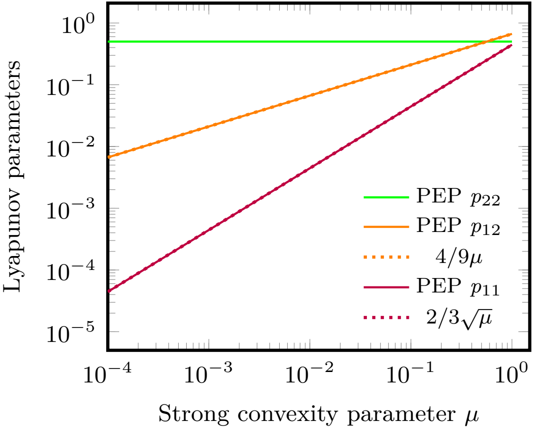
$$a = \nu_2 - \nu_1.$$

Usually, Lyapunov functions are defined for $P \succcurlyeq 0$ so that $\mathcal{V}_{a,P}(x) \geqslant 0$, which is here replaced with a relaxed nonnegativity condition $\mathcal{V}_{a,P}(X_t) \geqslant 0$.

Numerical help for computing Lyapunov parameters



(a) Best guarantees found within the class of quadratic Lyapunov functions.



(b) Lyapunov parameters P for $\tau = 4/3\sqrt{\mu}$ and $a = 1$, as a function of the condition number μ.

An improved convergence guarantee for the Polyak's damped oscillator

A classical Lyapunov function is given by[5]

$$\mathcal{V}(X_t) = f(X_t) - f_\star + \frac{1}{2}\begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \left(\begin{pmatrix} \mu & \sqrt{\mu} \\ \sqrt{\mu} & 1 \end{pmatrix} \otimes I_d \right) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix},$$

that verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -\sqrt{\mu}\mathcal{V}(X_t)$ for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{\mu,\infty}$, and all trajectory X_t generated by the Polyak damped oscillator.

[5] see [8, Theorem 4.3], [7]

An improved convergence guarantee for the Polyak's damped oscillator

Using this framework, we show the function

$$\mathcal{V}(X_t) = f(X_t) - f_\star + \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix}^\top \left(\begin{pmatrix} 4/9\mu & 2/3\sqrt{\mu} \\ 2/3\sqrt{\mu} & 1/2 \end{pmatrix} \otimes I_d \right) \begin{pmatrix} X_t - X_\star \\ \dot{X}_t \end{pmatrix},$$

verifies $\frac{d}{dt}\mathcal{V}(X_t) \leqslant -4/3\sqrt{\mu}\mathcal{V}(X_t)$ for all dimension $d \in \mathbf{N}$, all function $f \in \mathcal{F}_{\mu,\infty}$, and all trajectory X_t generated by the Polyak damped oscillator.

A connection between SDEs and SGD

Stochastic gradient descent (SGD) is given by:

$$x_{k+1} = x_k - h_k \nabla \tilde{f}(x_k, \xi_{i_k}),$$

where $h_k > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, ..., \xi_n)$, and where $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$.

A connection between SDEs and SGD

Stochastic gradient descent (SGD) is given by:

$$x_{k+1} = x_k - h_k \nabla \tilde{f}(x_k, \xi_{i_k}),$$

where $h_k > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, ..., \xi_n)$, and where $\nabla \tilde{f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$.

A connection to **stochastic differential equation (SDEs)** was proven by Li et al. [4]:

$$dX_t = -h_t \nabla f(X_t)\, dt + h_t (\gamma \Sigma(X_t))^{1/2}\, dB_t,$$

where B_t is a standard Brownian motion, h_t the step size and Σ_t is a stochastic covariance matrix.

A connection between SDEs and SGD

Stochastic gradient descent (SGD) is given by:

$$x_{k+1} = x_k - h_k \tilde{\nabla f}(x_k, \xi_{i_k}),$$

where $h_k > 0$ is the step size, ξ_{i_k} are uniformly drawn in $(\xi_1, ..., \xi_n)$, and where $\tilde{\nabla f}(x_k, \xi_{i_k})$ is an unbiased estimate of full gradient $\nabla f(x_k)$.

A connection to **stochastic differential equation (SDEs)** was proven by Li et al. [4]:

$$dX_t = -h_t \nabla f(X_t)\, dt + h_t (\gamma \Sigma(X_t))^{1/2}\, dB_t,$$

where B_t is a standard Brownian motion, h_t the step size and Σ_t is a stochastic covariance matrix.

Lemma (Ito's Lemma)

Let g be a twice continuously differentiable function, and X_t be a stochastic process solution to the SDE (29)*, then*

$$dg(X_t, t) = \frac{\partial}{\partial t} g(X_t, t)\, dt + \frac{\partial}{\partial x} g(X_t, t)\, dX_t + \frac{1}{2}\gamma \mathrm{Tr}(\frac{\partial^2}{\partial x^2} g(X_t, t) \Sigma(X_t))\, dt.$$

Lyapunov functions from deterministic setting?

Let $f \in \mathcal{F}_{0,\infty}$ be convex and twice differentiable, X_t be generated by the SDE above for $h_t = 1$, and consider the Lyapunov approach from the deterministic setting for the gradient flow:

$$\mathcal{V}(x,t) = t(f(x) - f_\star) + \frac{1}{2}\|x - x_\star\|^2.$$

Ito's formula and convexity lead to:

$$\frac{d}{dt}\mathbf{E}\mathcal{V}(X_t,t) \leqslant -t\mathbf{E}\|\nabla f(X_t)\|^2 + \mathbf{E}\frac{1}{2}\mathrm{Tr}((t\nabla_x^2 f(X_t) + I)\Sigma(X_t))$$

After integrating between 0 and t, assuming f to be L-smooth and $\Sigma_t \preccurlyeq \Sigma$,

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{2t} + \frac{1}{2}(L\frac{t}{2} + 1)\mathrm{Tr}(\Sigma).$$

Diminishing step sizes is the key to succes

Corollary

Let $f \in \mathcal{F}_{0,\infty}$ be a twice continuously differentiable function, and $X_t \in \mathbf{R}^d$ be generated by the SDE. The quadratic function

$$\mathcal{V}(X_t, t) = a_t^{(1)}(f(X_t) - f_\star) + \frac{1}{2}\|X_t - x_\star\|^2,$$

with $\dot{a}_t^{(1)} = 2h_t$ *verifies* $\frac{d}{dt}\mathbf{E}[\mathcal{V}(X_t, t)] \leqslant h_t^2 \mathbf{E}\mathrm{Tr}((\nabla_{xx}^2 f(X_t) a_t^{(1)} + \frac{1}{2}I_d)\Sigma(X_t))$.
Furthermore, it holds that:

$$\mathbf{E}[f(X_t) - f_\star] \leqslant \frac{\|x_0 - x_\star\|^2}{a_t^{(1)}} + \frac{\gamma}{2a_t^{(1)}}\int_0^t h_s^2 \mathbf{E}\mathrm{Tr}((\nabla_{xx}^2 f(X_s) a_s^{(1)} + \frac{1}{2}I_d)\Sigma(X_s))\, ds.$$

- A term that **forgets the initial conditions**
- A variance term due to **noise**

Choosing the best step size

Let the step size be defined for $\alpha \geqslant 0$:

$$h_t = \frac{1}{(t+1)^\alpha}.$$

Then, assuming bounded covariance and smoothness of f, the term $\mathbf{E}[f(X_t) - f_\star]$ is

- bounded by $\mathcal{O}(\frac{1}{t^{2\alpha-1}})$ if $\alpha \in (1/2, 2/3)$,
- bounded by $\mathcal{O}(\frac{1}{t^{1-\alpha}})$ if $\alpha \in (2/3, 1)$
- unbounded otherwise.

The convergence regime changes at $\alpha = \frac{2}{3}$ with a global convergence rate in $\mathcal{O}(\frac{1}{t^{1/3}})$, as for SGD [6], but using **simpler formulations and fewer assumptions**.

[6] see Bach and Moulines [1, Theorem 3]

Extensions

Other techniques were developed to improve convergence, and can be handled using this framework, such as:

- Polyak-Ruppert averaging

$$\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

- Non-uniform averaging
- Higher-order stochastic differential equations

$$d^2 X_t + \beta_t dX_t + h_t \nabla f(X_t)\, dt + h_t \sqrt{\gamma \Sigma(X_t)}\, dB_t = 0.$$

Concluding remarks

Conclusion:

- Verifying a Lyapunov function can be cast as the feasibility of a small-sized LMI
- A systematic approach to finding quadratic Lyapunov functions for families of ODEs
- May be extended in the stochastic setting for SDEs
- Similar guarantees to the discrete setting requiring less assumptions on the problem classes, and shorter proofs

Concluding remarks

Conclusion:

- Verifying a Lyapunov function can be cast as the feasibility of a small-sized LMI
- A systematic approach to finding quadratic Lyapunov functions for families of ODEs
- May be extended in the stochastic setting for SDEs
- Similar guarantees to the discrete setting requiring less assumptions on the problem classes, and shorter proofs

Future work:

- Extension of the family of quadratic Lyapunov functions
- Analyzing differential and monotone inclusion problems
- Analyzing higher order methods and assumptions (already implied by the variance term in the stochastic setting)

A systematic approach to Lyapunov analyses of continuous-time models in convex optimization versions

Céline Moucer

PEP Talks, February 2023

Thanks!
Any question?

ENS | PSL
École des Ponts ParisTech

References I

[1] Francis Bach and Eric Moulines. "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning". In: *Neural Information Processing Systems (NIPS)*. 2011.

[2] Yoel Drori and Marc Teboulle. "Performance of first-order methods for smooth convex minimization: a novel approach". In: *Mathematical Programming* 145.1 (2014), pp. 451–482.

[3] Baptiste Goujaud et al. *PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python*. 2022.

[4] Qianxiao Li, Cheng Tai, and Weinan E. "Stochastic modified equations and adaptive stochastic gradient algorithms". In: *International Conference on Machine Learning (ICML)*. 2017.

[5] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

References II

[6] B.T. Polyak. *Some methods of speeding up the convergence of iteration methods.* 1964.

[7] Jesús María Sanz Serna and Konstantinos C Zygalakis. "The connections between Lyapunov functions for some optimization algorithms and differential equations". In: *SIAM Journal on Numerical Analysis* 59.3 (2021), pp. 1542–1565.

[8] Bin Shi et al. "Acceleration via Symplectic Discretization of High-Resolution Differential Equations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[9] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *The Journal of Machine Learning Research (JMLR)* 17.153 (2016), pp. 1–43.

[10] Adrien B Taylor, Julien M Hendrickx, and François Glineur. "Smooth strongly convex interpolation and exact worst-case performance of first-order methods". In: *Mathematical Programming* 161.1 (2017), pp. 307–345.