Étude de méthodes inertielles en optimisation et leur comportement sous conditions de géométrie

Hippolyte Labarrière

Directeurs de thèse : Jean-François Aujol, Charles Dossal et Aude Rondepierre

Soutenance de thèse
Amphithéâtre Sophie Germain
INSA Toulouse
20 Septembre 2023

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

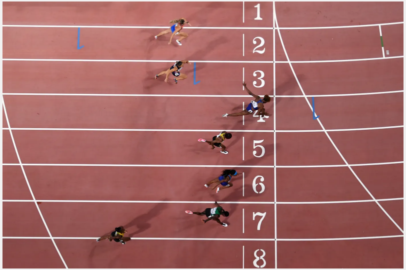
École Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Framework and motivations

Optimization, what is this?

→ Find a set of parameters that minimizes a quantity.

Find the route that minimizes journey time.



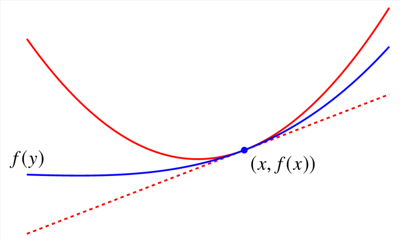
Find the training that leads to the best 100-meter time.

Minimization problem

$$\min_{x\in\mathbb{R}^N} F(x) = f(x) + h(x),$$

where:

- f is a convex differentiable function having a L-Lipschitz gradient,



- h is a convex proper lower semicontinuous function,
- F has a non-empty set of minimizers X^*.

Framework and motivations

Motivations

$$\min_{x \in \mathbb{R}^N} F(x),$$

Which algorithm is the most efficient according to the **assumptions** satisfied by F and the **expected accuracy**?

$\rightarrow$ **Convergence analysis** of the numerical schemes:

How fast does $F(x_k) - F^*$ decreases?

Outline

A classical algorithm: the proximal gradient method (Combettes and Wajs, '05)

$$\forall k > 0,\ x_k = \text{prox}_{sh}\left(x_{k-1} - s\nabla f(x_{k-1})\right).$$

Composite version of the **Gradient Descent method**:

$$\forall k > 0,\ x_k = x_{k-1} - s\nabla F(x_{k-1}).$$

A classical algorithm: the proximal gradient method (Combettes and Wajs, '05)

$$\forall k > 0, \; x_k = \text{prox}_{sh}\left(x_{k-1} - s\nabla f(x_{k-1})\right).$$

Composite version of the **Gradient Descent method**:

$$\forall k > 0, \; x_k = x_{k-1} - s\nabla F(x_{k-1}).$$

Convergence guarantees

If F is convex and s is sufficiently small:

$$F(x_k) - F^* = \mathcal{O}\left(k^{-1}\right)$$

$\rightarrow$ Simple but slow!

A classical algorithm: the proximal gradient method

$$\forall k > 0,\ x_k = \text{prox}_{sh}\left(x_{k-1} - s\nabla f(x_{k-1})\right).$$

Inertia

A classical algorithm: the proximal gradient method

$$\forall k > 0, \; x_k = \text{prox}_{sh}\left(x_{k-1} - s\nabla f(x_{k-1})\right).$$

Introducing inertia

→ Apply the same transformation to a shifted point.

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}\left(y_{k-1} - s\nabla f(y_{k-1})\right), \\ y_k = x_k + \alpha_k(x_k - x_{k-1}), \end{cases}$$

How to chose α_k?

Inertia

A classical algorithm: the proximal gradient method

$$\forall k > 0,\ x_k = \text{prox}_{sh}\left(x_{k-1} - s\nabla f(x_{k-1})\right).$$

Introducing inertia

→ Apply the same transformation to a shifted point.

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}\left(y_{k-1} - s\nabla f(y_{k-1})\right), \\ y_k = x_k + \alpha_k(x_k - x_{k-1}), \end{cases}$$

How to chose α_k?

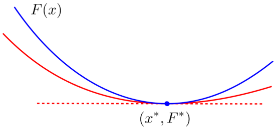
- **Heavy-Ball schemes** (Polyak,'64, Nesterov,'03, ...): constant friction $\rightarrow \alpha_k = \alpha$.
- **FISTA** (Beck and Teboulle,'09, Nesterov,'83): vanishing friction $\rightarrow \alpha_k = \frac{k-1}{k+\alpha-1}$. If F is convex, $\alpha \geqslant 3$ and s is sufficiently small:

$$F(x_k) - F^* = \mathcal{O}\left(k^{-2}\right)$$

Classical geometry assumptions

- **Strong convexity ($\mathcal{SC}_\mu$)**:
  F is μ-strongly convex if for all $x \in \mathbb{R}^N$, $g : x \mapsto F(x) - \frac{\mu}{2}\|x\|^2$ is convex.
- **Quadratic growth condition ($\mathcal{G}^2_\mu$)**:
  F has a quadratic growth around its set of minimizers if

$$\exists \mu > 0, \ \forall x \in \mathbb{R}^N, \ \frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*.$$



Example: LASSO problem:

$$F(x) = \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1.$$

Convergence rate of $F(x_k) - F^*$

Algorithm	Convex	$\mathcal{SC}_\mu$
Proximal gradient method	$\mathcal{O}\left(k^{-1}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
Heavy-Ball methods	$\mathcal{O}\left(k^{-1}\right)$	$\mathcal{O}\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right)$
FISTA	$\mathcal{O}\left(k^{-2}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$

The continuous setting

→ **Key tool in convergence analysis**: Link numerical schemes to dynamical systems.

Gradient descent→ Gradient flow

$$x_k = x_{k-1} - s\nabla F(x_{k-1})$$

The continuous setting

→ **Key tool in convergence analysis**: Link numerical schemes to dynamical systems.

Gradient descent→ Gradient flow

$$x_k = x_{k-1} - s\nabla F(x_{k-1})$$

$$\iff \quad \frac{x_k - x_{k-1}}{s} = -\nabla F(x_{k-1})$$

The continuous setting

$\rightarrow$ **Key tool in convergence analysis**: Link numerical schemes to dynamical systems.

Gradient descent $\rightarrow$ Gradient flow

$$x_k = x_{k-1} - s\nabla F(x_{k-1})$$

$$\iff \quad \frac{x_k - x_{k-1}}{s} = -\nabla F(x_{k-1})$$

$$\downarrow$$

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

The continuous setting

Nesterov's accelerated gradient→Asymptotic vanishing damping system (Su, Boyd and Candès,2014)

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}\left(y_{k-1} - s\nabla f(y_{k-1})\right), \\ y_k = x_k + \dfrac{k-1}{k+\alpha-1}(x_k - x_{k-1}) \end{cases}$$

$$\downarrow$$

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0$$

Heavy-Ball schemes→ Heavy-Ball Friction system

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}\left(y_{k-1} - s\nabla f(y_{k-1})\right), \\ y_k = x_k + \alpha(x_k - x_{k-1}), \end{cases}$$

$$\downarrow$$

$$\ddot{x}(t) + \alpha_C\dot{x}(t) + \nabla F(x(t)) = 0$$

The continuous setting

Why is this relevant?

- easier computations (derivatives),
- most of the time, convergence properties of the trajectories can be extended to the iterates of the related scheme.

Back to the discrete setting

Challenging for the following reasons:

- no more derivative,
- several possible discretization choices,
- which condition on the stepsize?

Outline

1 Key concepts and mathematical tools
   Inertia
   Geometry of convex functions
   The continuous setting: a guideline for the discrete analysis

2 Restart strategies

3 Attenuating oscillations introducing Hessian-driven damping

4 Inertia without uniqueness of the minimizers

5 Conclusion

Restart strategies

Framework

$$\min_{x\in\mathbb{R}^N} F(x),$$

where F satisfies a growth condition ($\mathcal{SC}_\mu$ or $\mathcal{G}_\mu^2$) and the growth parameter μ is not known.

First-order methods

In this setting:

- **proximal gradient method**: $F(x_k)-F^*=\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$,
- **Heavy-Ball methods**: $F(x_k)-F^*=\mathcal{O}\left(e^{-K\sqrt{\frac{\mu}{L}}k}\right)$ if μ is known,
- **FISTA** (Beck and Teboulle,'09, Nesterov,'83):

$$\forall k>0,\begin{cases} x_k=\text{prox}_{sh}(y_{k-1}-s\nabla f(y_{k-1})),\\ y_k=x_k+\dfrac{k-1}{k+2}(x_k-x_{k-1})\end{cases}$$

$$\to F(x_k)-F^*=\mathcal{O}\left(k^{-2}\right)$$

Restarting FISTA, why?

- to take advantage of inertia,
- to avoid oscillations.

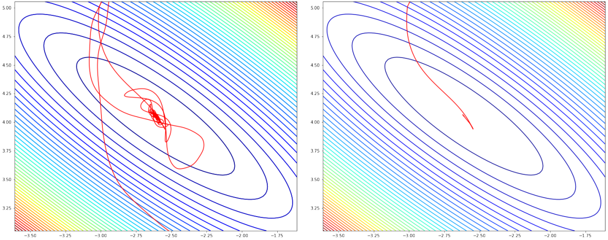


Figure: Projection of the trajectory of the iterates of FISTA (left) and FISTA restart (right) for a least-squares problem $(N = 20)$.

Restart strategies

Restarting FISTA, how?

Algorithm 1 : FISTA restart

Require: $x_0 \in \mathbb{R}^N, \; y_0 = x_0, \; k = 0, \; i = 0.$
  repeat
    $k = k + 1, i = i + 1$
    $x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1}))$
    if Restart condition is $True$ **then**
      $i = 1$
    end if
    $y_k = x_k + \frac{i-1}{i+2}(x_k - x_{k-1})$
  until Exit condition is $True$

$\rightarrow$ Cutting inertia is equivalent to restarting the algorithm from the last iterate.

Restart strategies

Empiric FISTA restart (O'Donoghue and Candès, '15, Beck and Teboulle, '09)

Restart under some exit condition

- on F:

$$F(x_k) > F(x_{k-1}),$$

- on ∇F:

$$\langle \nabla F(x_k), x_k - x_{k-1} \rangle > 0.$$

Restart strategies

Empiric FISTA restart (O'Donoghue and Candès, '15, Beck and Teboulle, '09)

Restart under some exit condition

- on F:

$$F(x_k) > F(x_{k-1}),$$

- on ∇F:

$$\langle \nabla F(x_k), x_k - x_{k-1} \rangle > 0.$$

Fixed FISTA restart (Nesterov, '13, O'Donoghue and Candès, '15...)

Restart every k^* iterations where k^* is defined according to the growth parameter μ. If $k^* = \left\lfloor 2e\sqrt{\frac{L}{\mu}} \right\rfloor$:

$$F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k} \right).$$

Generalization: Scheduled restarts, Roulet and D'Aspremont '17.

Adaptive FISTA restart

Restart according to the geometry of F and previous iterations.

- Fercoq and Qu, '19: $F(x_k) - F^* = o\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}\left(2-\sqrt{\frac{\mu}{\mu_0}}\right)}\sqrt{\frac{\mu}{L}}k} \right)$.
- Alamo et al., '19: $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$.
- Alamo et al., '22: $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{\ln(15)}{4e}\sqrt{\frac{\mu}{L}}k}\right)$, where $\frac{\ln(15)}{4e} \approx \frac{1}{4}$.
- Renegar and Grimmer, '22: $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{2\sqrt{2}}\sqrt{\frac{\mu}{L}}k}\right)$.

Introduction of an automatic restart scheme [1]:

Features: a restart condition that

- does not require to know the growth parameter μ,
- ensures a fast convergence of the method: $F(x_k) - F^* = \mathcal{O}(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k})$,
- is not computationnaly expensive,
- is easy to implement.

Strategy

- to estimate μ at each restart,
- to adapt the number of iterations of the following restart according to this estimation.

[1] FISTA restart using an automatic estimation of the growth parameter. Aujol, Dossal, L., Rondepierre, '21.

Algorithm 2 : Automatic FISTA restart

Require: $r_0 \in \mathbb{R}^N,\ j = 1,\ C = 6.38.$

$n_0 = \lfloor 2C \rfloor$

$r_1 = \text{FISTA}(r_0, n_0)$

$n_1 = \lfloor 2C \rfloor$

repeat

$j = j + 1$

$r_j = \text{FISTA}(r_{j-1}, n_{j-1})$

$$\tilde{\mu}_j = \min_{\substack{i \in \mathbb{N}^* \\ i < j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}$$

Estimation of the parameter μ.

if $n_{j-1} \leqslant C\sqrt{\frac{L}{\tilde{\mu}_j}}$ **then**

$n_j = 2n_{j-1}$

Update of the number of iterations per restart.

end if

until *exit condition is satisfied*

Restart strategies

Theorem (Aujol, Dossal, L., Rondepierre,'21)

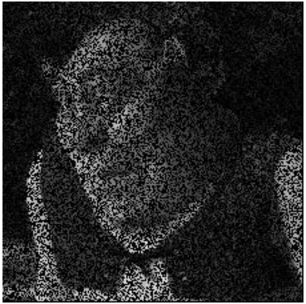
If F satisfies the assumptions stated before, then

$$F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right).$$

Image inpainting:

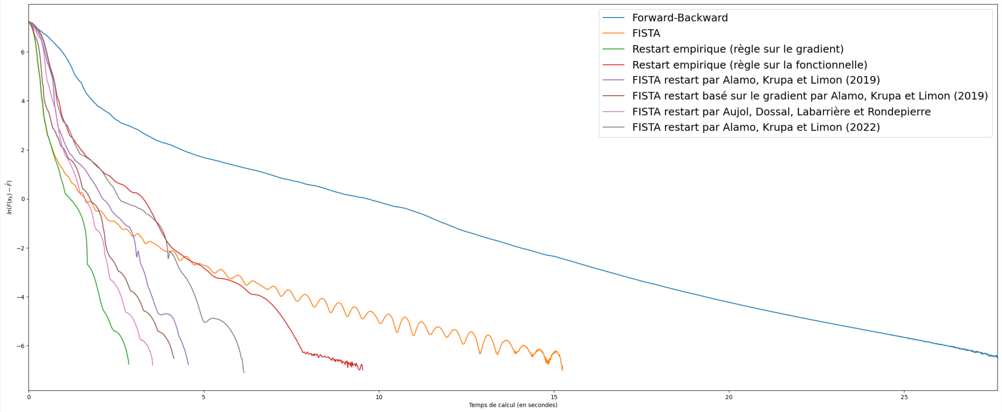
$$\min_x F(x) := \frac{1}{2}\|Mx - y\|^2 + \lambda\|Tx\|_1,$$

where M is a mask operator and T is an orthogonal transformation ensuring that Tx^0 is sparse.



Restart strategies

Image inpainting:



Introducing Free-FISTA, a parameter-free restart scheme [2]:

Combining backtracking and restarting

By combining a **backtracking strategy** and a **restarting strategy**, Free-FISTA automatically estimates μ and L.

- Still efficient if L **is not known**.
- Adaptation to the **local geometry of** F.
- **Convergence guarantees**: $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{\sqrt{\rho}}{12}\sqrt{\frac{\mu}{L}}k}\right)$.

[2] Parameter-Free FISTA by Adaptive Restart and Backtracking. Aujol, Calatroni, Dossal, L., Rondepierre, '23.

Outline

Asymptotic vanishing damping system (Su et al., '14)

FISTA can be seen as a discretization of the following ODE:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0.$$

Both the dynamical system and the numerical scheme exhibit an **oscillatory behavior**.

Convergence properties for convex functions

Convergence rate of the error:

$$F(x(t)) - F^* = \mathcal{O}\left(t^{-2}\right)$$

Attenuating oscillations via Hessian-driven damping

Hessian-driven damping

(DIN-AVD) system (**Attouch, Peypouquet and Redont,'16**)

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta H_F(x(t))\dot{x}(t) + \nabla F(x(t)) = 0.$$

- Can be discretized to define **first-order schemes**.
- Attenuation of the oscillations through the introduction of a geometry-driven damping term.

Convergence properties for C^2 convex functions

- Convergence rate of the error:

$$F(x(t)) - F^* = \mathcal{O}\left(t^{-2}\right)$$

- Integrability of the gradient:

$$\int_{t_0}^{+\infty} t^2 \|\nabla F(x(t))\|^2 dt < +\infty,$$

What happens if F satisfies $\mathcal{G}_\mu^2$?

Improved integrability of the gradient

- **Theorem ([3])**: if F is convex and C^2, satisfies $\mathcal{G}_\mu^2$ and has a unique minimizer. Then, for $\alpha \geqslant 3$ and $\beta > 0$:

$$\int_{t_0}^{+\infty} t^{\alpha-\varepsilon}\|\nabla F(x(t))\|^2 dt < +\infty, \forall \varepsilon \in (0,1).$$

[3] Fast convergence of inertial dynamics with Hessian-driven damping under geometry assumptions. Aujol, Dossal, Hoàng, L., Rondepierre, '22, accepted in AMOP.

What happens if F satisfies $\mathcal{G}_\mu^2$?

Improved integrability of the gradient

- **Theorem ([3])**: if F is convex and C^2, satisfies $\mathcal{G}_\mu^2$ and has a unique minimizer. Then, for $\alpha \geqslant 3$ and $\beta > 0$:

$$\int_{t_0}^{+\infty} t^{\alpha-\varepsilon}\|\nabla F(x(t))\|^2 dt < +\infty, \forall \varepsilon \in (0,1).$$

$$\downarrow$$

$$\int_{t_0}^{+\infty} t^{\alpha-\varepsilon}(F(x(t)) - F^*) dt < +\infty, \forall \varepsilon \in (0,1).$$

[3] Fast convergence of inertial dynamics with Hessian-driven damping under geometry assumptions. Aujol, Dossal, Hoàng, L., Rondepierre, '22, accepted in AMOP.

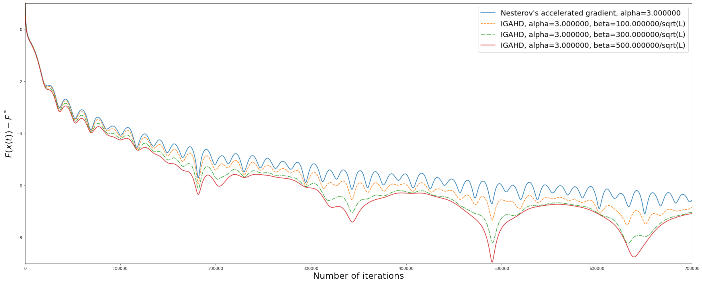
Attenuating oscillations via Hessian-driven damping

Derivating a numerical scheme: IGAHD (Attouch, Chbani, Fadili and Riahi,'20)

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta H_F(x(t))\dot{x}(t) + \left(1 + \frac{\beta}{t}\right)\nabla F(x(t)) = 0.$$

$$\downarrow$$

$$\begin{cases} x_k = y_{k-1} - s\nabla F(y_{k-1}), \\ y_k = x_k + \dfrac{k-1}{k+\alpha-1}(x_k - x_{k-1}) - \beta\sqrt{s}(\nabla F(x_k) - \nabla F(x_{k-1})) - \dfrac{\beta\sqrt{s}}{k}\nabla F(x_{k-1}), \end{cases}$$



Attenuating oscillations via Hessian-driven damping

Summary

The Hessian-driven damping term is a **physical way** to attenuate oscillations. As this is a relatively recent subject of research, there are some limitations:

- the behavior of the numerical schemes derivated from (DIN-AVD) is not fully understood (current convergence rates hold if β is **small**),
- the dependency in β is not known,
- there is no proof showing that it is faster than classical inertial schemes.

Outline

Inertia without uniqueness of the minimizers

Problem statement

Let F satisfy a growth condition (e.g. $\mathcal{G}_\mu^2$ or $\mathcal{SC}_\mu$).

Most improved convergence results for first-order inertial methods (and corresponding dynamical systems) rely on the assumption that F **has a unique minimizer**:

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}_\mu^2$ and unique minimizer	$\mathcal{G}_\mu^2$
Proximal gradient method	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
Heavy-Ball methods	$\mathcal{O}\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right)$	$\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
FISTA	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-2}\right)$

Is this hypothesis necessary to get fast convergence rates?

Inertia without uniqueness of the minimizers

Problem statement

Let F satisfy a growth condition (e.g. $\mathcal{G}_\mu^2$ or $\mathcal{SC}_\mu$).

Most improved convergence results for first-order inertial methods (and corresponding dynamical systems) rely on the assumption that F **has a unique minimizer**:

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}_\mu^2$ and unique minimizer	$\mathcal{G}_\mu^2$
Proximal gradient method	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
Heavy-Ball methods	$\mathcal{O}\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right)$	$\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
FISTA	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-2}\right)$

Is this hypothesis necessary to get fast convergence rates?

No!

The continuous setting

Consider the **Heavy-Ball friction system**:

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = 0$$

Classical Lyapunov energy for this system:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2$$

Inertia without uniqueness of the minimizers

The continuous setting

Consider the **Heavy-Ball friction system**:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0$$

Classical Lyapunov energy for this system:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*(t)) + \dot{x}(t)\|^2$$

where $x^*(t)$ is the projection of $x(t)$ onto the set of minimizers of F denoted X^*.

Inertia without uniqueness of the minimizers

The continuous setting

Consider the **Heavy-Ball friction system**:

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = 0$$

Classical Lyapunov energy for this system:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*(t)) + \dot{x}(t)\|^2$$

where $x^*(t)$ is the projection of $x(t)$ onto the set of minimizers of F denoted X^*.

$\rightarrow$ The differentiability of $\mathcal{E}$ depends on the regularity of X^*!

If X^* is **sufficiently regular** (e.g. polyhedral), several convergence results can be extended **without the uniqueness assumption** (e.g. Siegel, '19, Aujol, Dossal and Rondepierre, '23).

Inertia without uniqueness of the minimizers

The discrete setting

Consider **V-FISTA** (Beck,'17):

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1})) \\ y_k = x_k + \alpha(x_k - x_{k-1}) \end{cases}$$

Classical discrete Lyapunov energy for this system:

$$\mathcal{E}_k = s(F(x_k) - F^*) + \frac{1}{2}\|\lambda(x_k - x^*) + x_k - x_{k-1}\|^2$$

The discrete setting

Consider **V-FISTA** (Beck,'17):

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1})) \\ y_k = x_k + \alpha(x_k - x_{k-1}) \end{cases}$$

Classical discrete Lyapunov energy for this system:

$$\mathcal{E}_k = s(F(x_k) - F^*) + \frac{1}{2}\|\lambda(x_k - x_k^*) + x_k - x_{k-1}\|^2$$

where x_k^* is the projection of x_k onto the set of minimizers of F denoted X^*.

The discrete setting

Consider **V-FISTA** (Beck,'17):

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1})) \\ y_k = x_k + \alpha(x_k - x_{k-1}) \end{cases}$$

Classical discrete Lyapunov energy for this system:

$$\mathcal{E}_k = s(F(x_k) - F^*) + \frac{1}{2}\|\lambda(x_k - x_k^*) + x_k - x_{k-1}\|^2$$

where x_k^* is the projection of x_k onto the set of minimizers of F denoted X^*.

$\rightarrow$ Trickier calculations
$\rightarrow$ No assumption on X^* required!

Main results: V-FISTA

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1})) \\ y_k = x_k + \alpha(x_k - x_{k-1}) \end{cases}$$

Theorem ([4]): If F satisfies $\mathcal{G}^2_\mu$, $s = \frac{1}{L}$ and $\alpha = 1 - \frac{5}{3\sqrt{3}}\sqrt{\frac{\mu}{L}}$:

$$F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{2}{3\sqrt{3}}\sqrt{\frac{\mu}{L}}k}\right)$$

- Uniqueness of the minimizer is not required.
- Theoretical guarantees for non optimal values of α.

[4] Fast Convergence of Heavy-Ball Dynamics and Derived Scheme Without Uniqueness of the Minimizer. Aujol, Dossal, L., Rondepierre, to be submitted.

Strong convergence of FISTA

Main results: FISTA

$$\forall k > 0, \begin{cases} x_k = \text{prox}_{sh}(y_{k-1} - s\nabla f(y_{k-1})) \\ y_k = x_k + \dfrac{k-1}{k+\alpha-1}(x_k - x_{k-1}) \end{cases}$$

Theorem ([5]): If there exists $\varepsilon > 0$, $K > 0$ and $\gamma > 2$ such that F satisfies the following inequality for any minimizer x^*

$$\forall x \in B(x^*, \varepsilon),\ K d(x, X^*)^\gamma \leqslant F(x) - F^*,$$

then for α sufficiently large:

$$F(x_k) - F^* = \mathcal{O}\left(k^{-\frac{2\gamma}{\gamma-2}}\right) \text{ and } \|x_k - x_{k-1}\| = \mathcal{O}\left(k^{-\frac{\gamma}{\gamma-2}}\right)$$

$\rightarrow$ The sequence $(x_k)_{k\in\mathbb{N}}$ converges **strongly** to a minimizer of F.

[5] Strong Convergence of FISTA under a Weak Growth Condition. Aujol, Dossal, L., Rondepierre, to be submitted.

Outline

Conclusion

Summary:

- Study of algorithmic designed to eliminate the need to know the growth parameter $\rightarrow$ **Restart strategy**
- Analysis of a physical approach aiming at attenuating oscillations $\rightarrow$ **Hessian-driven damping**
- Proof that inertial methods are still efficient for functions with multiple minimizers.

Conclusion

Unanswered questions:

- Is it possible to adapt geometry parameter estimation to Heavy-Ball type methods (restart scheme)?
- Could restarting strategies be combined to Hessian-driven damping? (yes $\rightarrow$ Maulen and Peypouquet, '23)
- How can high-resolution ODEs (see Shi et al., '18) improve convergence analysis?
- Is it possible to use the Performance Estimation Problem approach (Drori and Teboulle, '14, Taylor, Hendrickx and Glineur, '17, Taylor and Drori, '22):
  - to analyse (DIN-AVD)-schemes?
  - for functions satisfying growth conditions (but not strongly convex)?
- How do inertial methods behave in a non-convex setting? (Good luck Julien!)

Thank you for your attention!

Preprints:

- Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. FISTA restart using an automatic estimation of the growth parameter. *Submitted in 2021 to JOTA (minor revision)*. ⟨hal-03153525v4⟩
- Jean-François Aujol, Charles Dossal, Văn Hào Hoàng, Hippolyte Labarrière, Aude Rondepierre. Fast convergence of inertial dynamics with Hessian-driven damping under geometry assumptions. *Submitted in 2022 to AMOP (accepted)*. ⟨hal-03693218v2⟩
- Jean-François Aujol, Luca Calatroni, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Parameter-Free FISTA by Adaptive Restart and Backtracking. *Submitted in 2023 to SIOPT*. ⟨hal-04172497⟩

Forthcoming preprints:

- Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Fast Convergence of Heavy-Ball Dynamics and Derived Scheme Without Uniqueness of the Minimizer.
- Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Strong Convergence of FISTA under a Weak Growth Condition.

Website:

https://www.math.univ-toulouse.fr/~hlabarri/

References I

T. Alamo, D. Limon, and P. Krupa.
Restart FISTA with global linear convergence.
pages 1969–1974, 2019.

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi.
First-order optimization algorithms via inertial systems with hessian driven damping.
Mathematical Programming, pages 1–43, 2020.

H. Attouch, J. Peypouquet, and P. Redont.
Fast convex optimization via inertial dynamics with hessian driven damping.
Journal of Differential Equations, 261(10):5734–5783, 2016.

J.-F. Aujol, C. Dossal, and A. Rondepierre.
Convergence rates of the heavy-ball method under the łojasiewicz property.
Mathematical Programming, pages 1–60, 2022.

J.-F. Aujol, C. Dossal, and A. Rondepierre.
Fista is an automatic geometrically optimized algorithm for strongly convex functions.
Mathematical Programming, Apr 2023.

A. Beck.
First-order methods in optimization.
SIAM, 2017.

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202, 2009.

References II

Y. Drori and M. Teboulle.
Performance of first-order methods for smooth convex minimization: a novel approach.
Mathematical Programming, 145(1):451–482, Jun 2014.

O. Fercoq and Z. Qu.
Adaptive restart of accelerated gradient methods under local quadratic growth condition.
IMA Journal of Numerical Analysis, 39(4):2069–2095, 2019.

G. Garrigos, L. Rosasco, and S. Villa.
Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry.
Mathematical Programming, pages 1–60, 2022.

J. J. Maulén and J. Peypouquet.
A speed restart scheme for a dynamics with hessian-driven damping.
Journal of Optimization Theory and Applications, Sep 2023.

I. Necoara, Y. Nesterov, and F. Glineur.
Linear convergence of first order methods for non-strongly convex optimization.
Mathematical Programming, 175(1):69–107, 2019.

Y. Nesterov.
A method of solving a convex programming problem with convergence rate $o(1/k^2)$.
In *Sov. Math. Dokl*, volume 27, 1983.

B. O'donoghue and E. Candes.
Adaptive restart for accelerated gradient schemes.
Foundations of computational mathematics, 15(3):715–732, 2015.

References III

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su.
Understanding the acceleration phenomenon via high-resolution differential equations.
Mathematical Programming, 195(1):79–148, Sep 2022.

W. Su, S. Boyd, and E. Candes.
A differential equation for modeling nesterov's accelerated gradient method: theory and insights.
Advances in neural information processing systems, 27, 2014.