Inertial methods beyond minimizer uniqueness

Hippolyte Labarrière

Joint work with Jean-François Aujol, Charles Dossal and Aude Rondepierre

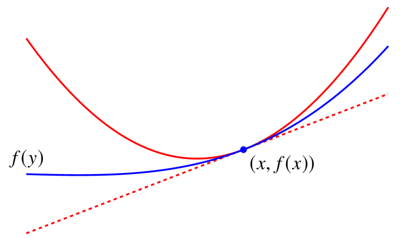
Journées SMAI-MODE 2024
Université de Lyon
March 29, 2024

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^*.

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?

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:

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

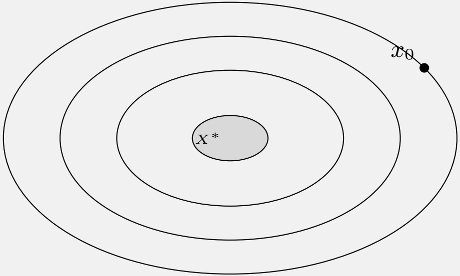
$\rightarrow$ Simple but slow!

Inertia in optimization

A classical algorithm: the proximal gradient method

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

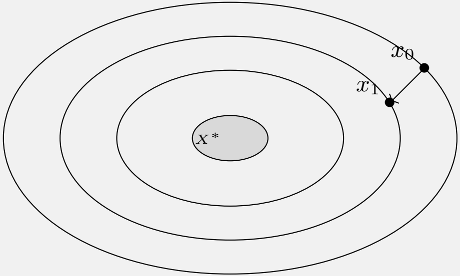
Illustration



A classical algorithm: the proximal gradient method

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

Illustration

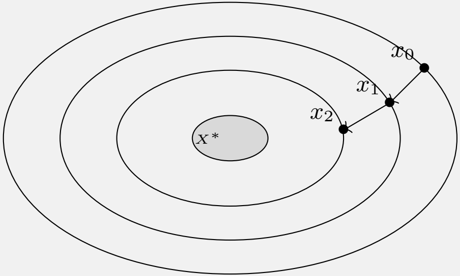


Inertia in optimization

A classical algorithm: the proximal gradient method

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

Illustration

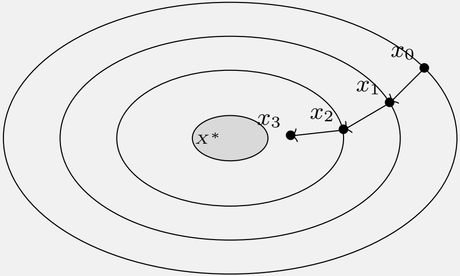


Inertia in optimization

A classical algorithm: the proximal gradient method

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

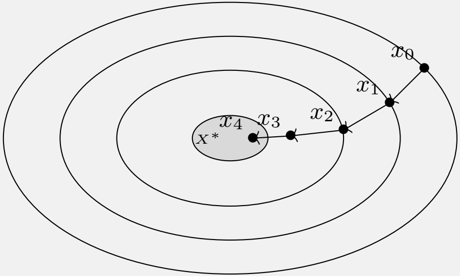
Illustration



A classical algorithm: the proximal gradient method

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

Illustration

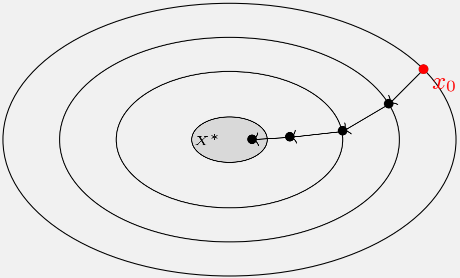


Introducing inertia

→ Apply the same transformation to a shifted point.

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

Illustration



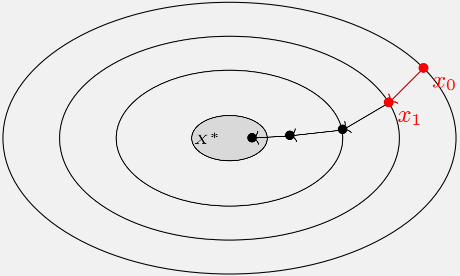
Inertia in optimization

Introducing inertia

→ Apply the same transformation to a shifted point.

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

Illustration



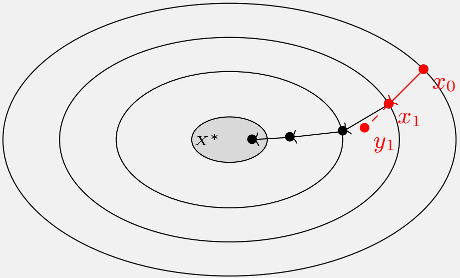
Inertia in optimization

Introducing inertia

$\rightarrow$ Apply the same transformation to a shifted point.

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

Illustration



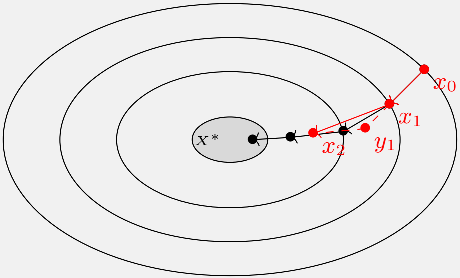
Inertia in optimization

Introducing inertia

$\rightarrow$ Apply the same transformation to a shifted point.

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

Illustration



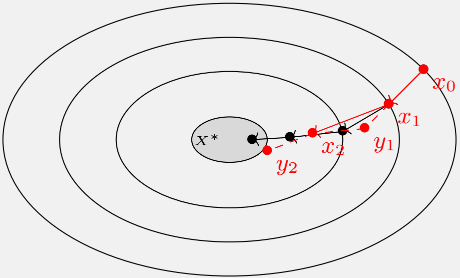
Inertia in optimization

Introducing inertia

→ Apply the same transformation to a shifted point.

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

Illustration

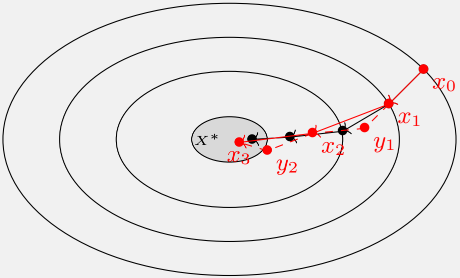


Introducing inertia

$\rightarrow$ Apply the same transformation to a shifted point.

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

Illustration



Introducing inertia

$\rightarrow$ Apply the same transformation to a shifted point.

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

Rising question

How to chose α_k?

Inertia in optimization

Introducing inertia

$\rightarrow$ Apply the same transformation to a shifted point.

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

Rising question

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 $\rightarrow \alpha_k = \frac{k-1}{k+\alpha-1}$.

Optimization and geometry

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.

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 (constant friction)	$\mathcal{O}\left(k^{-1}\right)$	$\mathcal{O}\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right)$
FISTA (vanishing friction)	$\mathcal{O}\left(k^{-2}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$

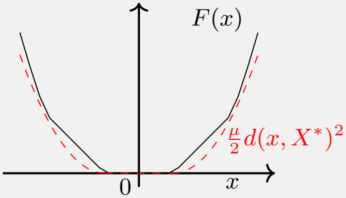
Classical geometry assumptions

- **Quadratic growth condition ($\mathcal{G}^2_\mu$)**:
  F has a quadratic growth around its set of minimizers if

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

Practical example: LASSO problem:

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



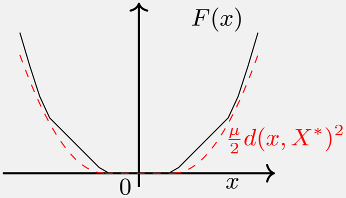
Classical geometry assumptions

- **Quadratic growth condition ($\mathcal{G}^2_\mu$)**:
  F has a quadratic growth around its set of minimizers if

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

Practical example: LASSO problem:

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



- **Hölderian error bound ($\mathcal{H}^\gamma$)**:
  F has a γ-Hölderian growth around its set of minimizers (with $\gamma > 2$) if

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

Optimization and geometry

Problem statement

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

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$ and unique minimizer	$\mathcal{G}^2_\mu$
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	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-2}\right)$

$\rightarrow$ FISTA restart schemes for $\mathcal{G}^2_\mu$: $\mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$ without an uniqueness assumption!

Is this hypothesis necessary to get fast convergence rates?

Inertia without uniqueness of the minimizers

Our strategy

Consider **V-FISTA** (Beck,'17, Nesterov, '03):

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

Classical discrete Lyapunov energy for this system:

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

Our strategy

Consider **V-FISTA** (Beck,'17, Nesterov, '03):

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

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^*.

Inertia without uniqueness of the minimizers

Our strategy

Consider **V-FISTA** (Beck,'17, Nesterov, '03):

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

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^*.

→ Trickier calculations
→ 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 (Aujol, Dossal, L., Rondepierre, '24): 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.
- Better worst-case bound than any FISTA restart scheme: $\mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$.
- α depends on $\frac{\mu}{L}$!

Main results: FISTA for $\mathcal{G}_\mu^2$

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

Theorem (Aujol, Dossal, L., Rondepierre, '24): If F satisfies $\mathcal{G}_\mu^2$, $s = \frac{1}{L}$ and $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$:

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

- Uniqueness of the minimizer is not required.
- Finite time bound available.
- α can be parametrized according to the expected accuracy to get improved performance.

Main results: FISTA for $\mathcal{H}^\gamma$

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

Theorem (Aujol, Dossal, L., Rondepierre, '24): 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 $\alpha > 5 + \frac{8}{\gamma-2}$:

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

Main results: FISTA for $\mathcal{H}^\gamma$

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

Theorem (Aujol, Dossal, L., Rondepierre, '24): If there exists $\varepsilon > 0$, $K > 0$ and $\gamma > 2$ such that F satisfies the following inequality for any minimizer x^*

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

then for $\alpha > 5 + \frac{8}{\gamma-2}$:

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

Corollary: Under the same assumptions, for any $\alpha > 5$, the sequence $(x_k)_{k\in\mathbb{N}}$ converges **strongly** to a minimizer of F.

Conclusion

Take-away message

In the convex setting, inertia is still **relevant** for functions having **multiple minimizers**!

	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$	$\mathcal{H}^\gamma$	Convexity
Best option	HB	HB	FISTA	FISTA

Pending questions:

- Heavy Ball methods require to know the growth parameter of F: could an adaptive strategy be applied to avoid this issue?
- Could the Performance Estimation Problem approach (Drori and Teboulle, '14, Taylor, Hendrickx and Glineur, '17, Taylor and Drori, '22) allow to find tighter bounds?
- How do inertial methods behave in a non-convex setting?

Conclusion

Thank you for your attention!

Preprints:

- Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Heavy Ball Momentum for Non-Strongly Convex Optimization, 2024, arXiv preprint arXiv:2403.06930.
- Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Strong Convergence of FISTA under a Weak Growth Condition, currently in writing.

My thesis manuscript (in french!):

- Hippolyte Labarrière, 2023, Étude de méthodes inertielles en optimisation et leur comportement sous conditions de géométrie.

Website:

https://hippolytelbrrr.github.io/

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.

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.

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.

References II

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. 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.

A key tool: the continuous setting

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

Gradient descent→ Gradient flow

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

A key tool: the continuous setting

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

Gradient descent$\rightarrow$ Gradient flow

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

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

A key tool: the continuous setting

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

Gradient descent$\rightarrow$ Gradient flow

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

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

$$\downarrow$$

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

A key tool: 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

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

$$\downarrow$$

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

A key tool: 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?

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:

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

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).

An ugly bound

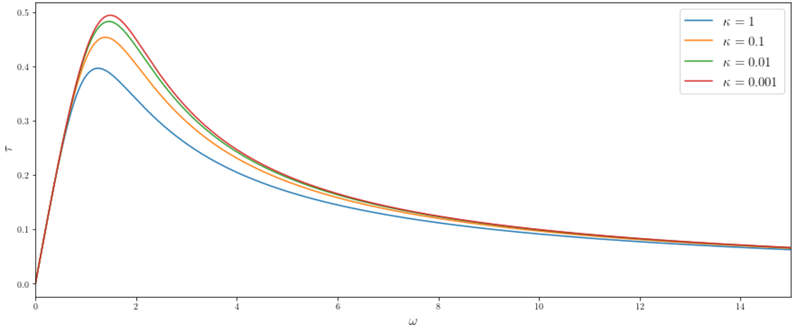
Main results: V-FISTA

If F satisfies $\mathcal{G}^2_\mu$, $s = \frac{1}{L}$ $\alpha = 1 - \omega\sqrt{\kappa}$ where $\kappa = \frac{\mu}{L}$, $\omega \in \left(0, \frac{1}{\sqrt{\kappa}}\right)$. Then, for any $k \in \mathbb{N}$:

$$F(x_k) - F^* \leqslant \left(1 + (\omega - \tau)^2 + (\omega - \tau)\omega\tau\sqrt{\kappa}\right)\left(1 - \tau\sqrt{\kappa} + \tau^2\kappa\right)^k (F(x_0) - F^*),$$

if

$$\left(1 - \omega\sqrt{\kappa}\right)\tau^3 - \omega\left(2 - \omega\sqrt{\kappa}\right)\tau^2 + (\omega^2 + 2)\tau - \omega \leqslant 0.$$



An other ugly bound

Main results: FISTA

If F satisfies $\mathcal{G}_\mu^2$, $s = \frac{1}{L}$, $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$, then

$$\forall k \geqslant \frac{3\alpha}{\sqrt{\kappa}}, \; F(x_k) - F^* \leqslant \frac{9}{4} e^{-2} M_0 \left(\frac{8e}{3\sqrt{\kappa}} \alpha \right)^{\frac{2\alpha}{3}} k^{-\frac{2\alpha}{3}},$$

where $M_0 = F(x_0) - F^*$ denotes the potential energy of the system at initial time.