Un tour d'horizon des résultats récents sur les algorithmes inertiels dans un cadre déterministe

Aude Rondepierre

Joint work with Jean-François Aujol, Charles Dossal

Institut de Mathématiques de Toulouse, INSA de Toulouse

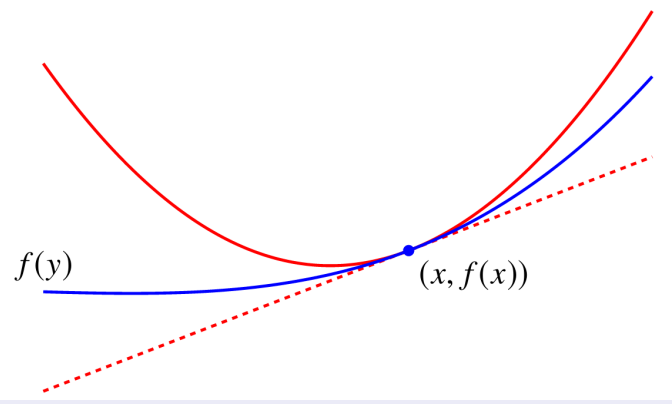
Journées de la fédération OCCIMATH, Montpellier le 28 mai 2026

The setting: Large scale optimization

Let:

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

where $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a convex differentiable function having at least one minimizer x^* and having a L-Lipschitz gradient:



For all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$, we have:

$$f(y) \leqslant \underbrace{f(x) + \langle \nabla f(x), y - x \rangle}_{\text{linear approximation}} + \frac{L}{2}\|y - x\|^2$$

Includes the composite case: $F = f + h$ where f is a convex differentiable function and h is a convex lower semicontinuous (lsc) simple function.

Goal

Design a **first-order algorithm** (only gradient evaluations) that makes $F(x_n) - F(x^*)$ decrease as fast as possible.

Gradient Descent (Cauchy (1857) - Polyak (1964))

Algorithm (GD)

$$x_{n+1} = x_n - s\nabla F(x_n) \text{ with } s \leqslant \frac{1}{L}.$$

Convergence rate for convex functions

The iterates $(x_n)_{n\in\mathbb{N}}$ weakly converge to a $x^* \in \arg\min(F)$ and:

$$\forall n \in \mathbb{N},\ F(x_n) - F(x^*) \leqslant \frac{\|x_0 - x^*\|^2}{2sn}$$

Slow polynomial convergence:

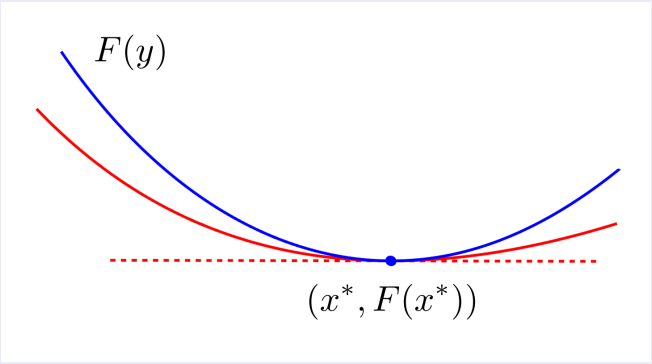
- Number of iterations to reach $F(x_n) - F(x^*) \leqslant \varepsilon$ in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$
- The rate $\frac{1}{n}$ can not be improved assuming only convexity !

Stronger assumptions, better rates (1/2)

The quadratic growth condition

Quadratic growth condition, a relaxation of strong convexity

If F satisfies some quadratic growth condition around its minimizers:



There exists $\mu > 0$ such that:

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

Convergence rate for functions with quadratic growth

$$F(x_n) - F(x^*) = \mathcal{O}\left((1-\kappa)^n\right), \qquad \kappa = \frac{\mu}{L}.$$

Rate achieved for $F(x_1, x_2) = \frac{\mu}{2} x_1^2 + \frac{L}{2} x_2^2$.

Stronger assumptions, better rates (2/2)

The quadratic growth condition

Convergence rate for functions with quadratic growth

$$f(x_n) - f(x^*) = \mathcal{O}(e^{-\kappa n}) = \mathcal{O}(e^{-n\mu/L})$$

Linear convergence but **slow**:

- Number of iterations: $n_\varepsilon^{GD} \sim \frac{L}{\mu} \log(1/\varepsilon) \gg 1$
- Need L/μ iterations to gain factor e in the bound

Practical setting

- High dimension: $N \sim 10^6$ (image processing, machine learning...)
- Conditioning: $\kappa = \mu/L \sim 10^{-6}$ (very small!)
- Problem is highly anisotropic: convergence speed constrained by the flattest direction

Motivating example: LASSO

LASSO problem

$$F(x) = \|Ax - y\|_2^2 + \lambda \|x\|_1$$

- Satisfies a quadratic growth condition with some $\mu > 0$
- But: μ does not depend only on eigenvalues of A - very difficult to estimate in practice
- Non-differentiable: requires proximal algorithms

Key point

All results presented in this talk extend to LASSO and more general composite problems via proximal methods:

$$x_{n+1} = \mathrm{prox}_{sh}(x_n - s\nabla f(x_n)).$$

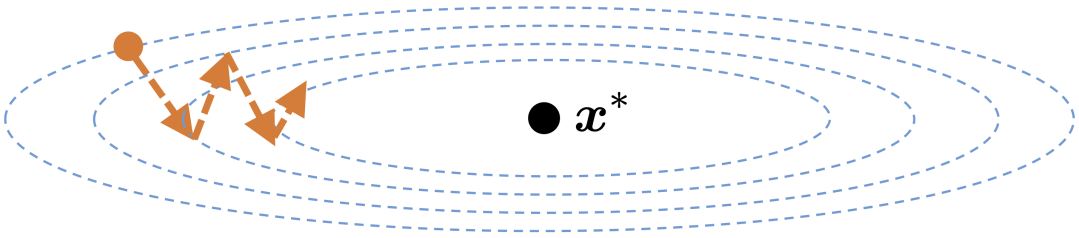
The Heavy Ball method

A first inertial method (Polyak 1964)

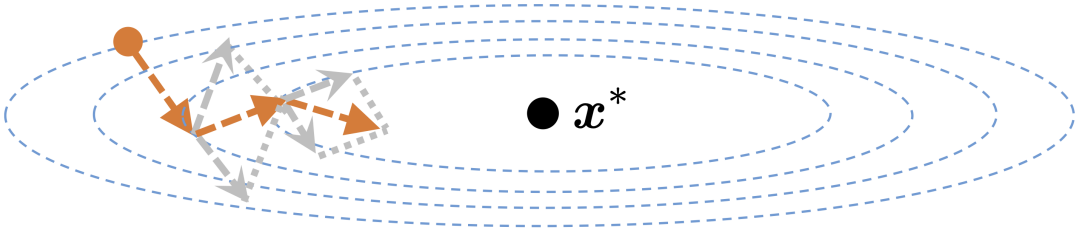
The Heavy ball method

$$\begin{array}{rcl} y_k & = & x_k + a(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla F(x_k) \end{array}, \ \alpha \in [0,1], \ s > 0.$$

where $a \in [0,1]$ is a *fixed* inertial coefficient added to mitigate zigzagging.



gradient descent



heavy-ball method

Link between the continuous ODE and the discrete scheme

The HB algorithm:

$$\begin{array}{rcl} y_k & = & x_k + a(x_k - x_{k-1}) \\ x_{k+1} & = & y_k - s\nabla F(x_k) \end{array}, \ \alpha \in [0,1], \ s > 0.$$

can be seen as a discretization of the second order ODE:

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

with: $s = h^2$ and $a = 1 - \alpha h$ (a: inertia parameter - α: friction parameter).

- Describe the motion of a body in a potential field F with a constant friction.
- High friction $\alpha \leftrightarrow$ low inertia a

The Heavy Ball method

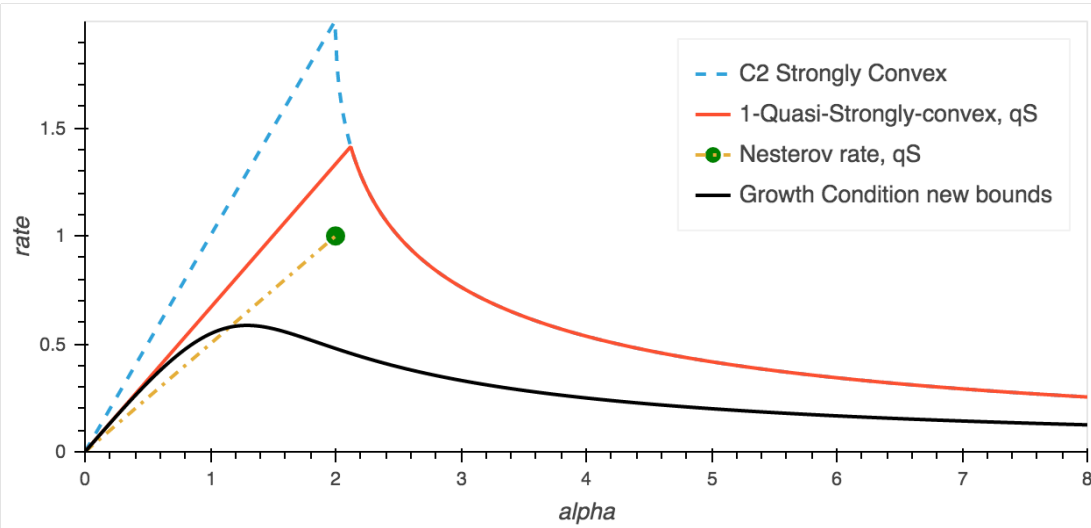
Optimal friction: 1D quadratic analysis

1D quadratic case: $f(x) = \frac{\mu}{2}x^2$

The HB ODE becomes linear:

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

Solution: $x(t) = Ae^{r_1 t} + Be^{r_2 t}$
where r_1, r_2 solve $r^2 + \alpha r + \mu = 0$.



Optimal friction

- Optimal choice: $\alpha = 2\sqrt{\mu}$ (critical damping)
- α too small $\rightarrow$ oscillations ($\alpha = 0$: harmonic oscillator)
- α too large $\rightarrow$ overdamping, loss of inertia benefit
- Requires knowing μ!

The Heavy Ball method

Convergence results and sensitivity to μ

$$\begin{aligned} y_k &= x_k + a(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla F(x_k) \end{aligned}$$

Theorem (Convergence for strongly convex C^2 functions [Polyak 1964])

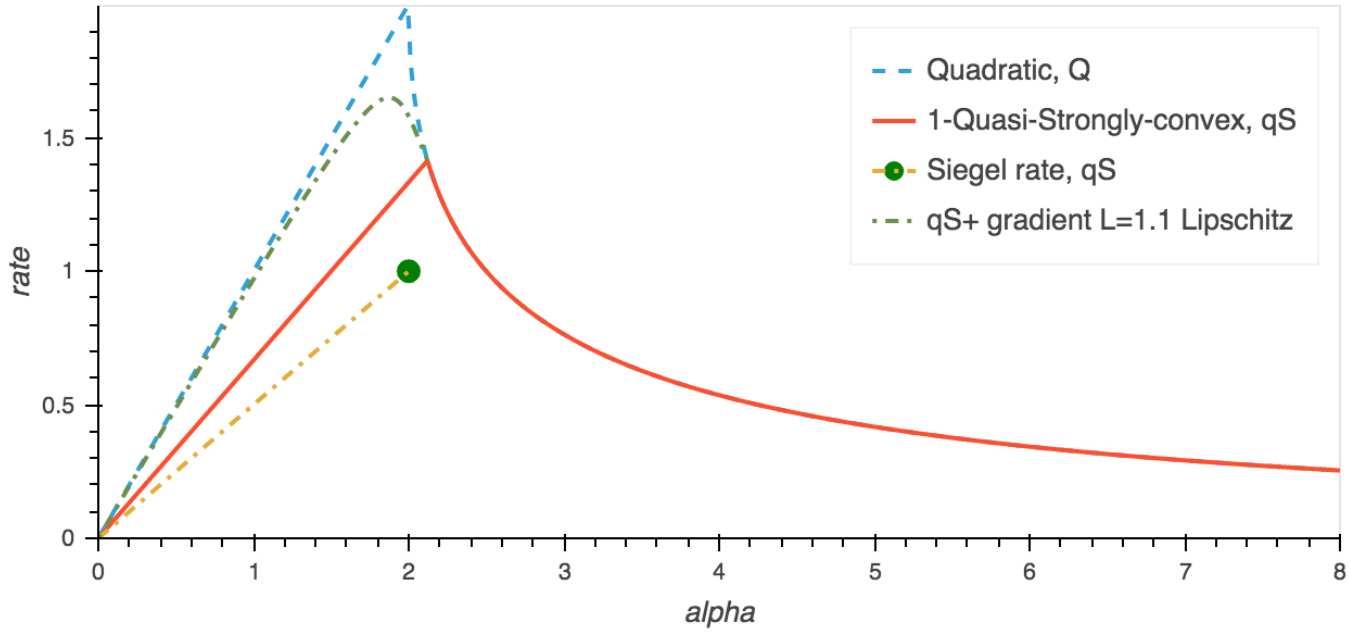
If $a = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$, $\quad s = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2$, *then:*

$$F(x_k) - F^* \leqslant \underbrace{\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k}_{\sim O\left(e^{-2\sqrt{\frac{\mu}{L}}k}\right),\ k\to+\infty} (F(x_0) - F^*).$$

Number of iterations: $n_\varepsilon^{HB} \sim \sqrt{\frac{L}{\mu}}\log(1/\varepsilon)$ $\quad \rightarrow$ much better than GD when $\kappa \ll 1$, but **requires knowing** μ!

The Heavy Ball method

How to choose α to optimize the convergence to a minimizer ?



- For strongly convex functions of class C^2 having a L-Lipschitz gradient, the optimal value of α is: $\alpha = 2\sqrt{\mu}$.
- Changing the step and the inertia, [Ghadimi et al. 2015] prove the geometric cv for C^1 strongly convex functions having a Lipschitz continuous gradient.
- For strongly convex functions of class C^1 having a L-Lipschitz gradient [Siegel 2019]: when $\alpha = 2\sqrt{\mu}$, $F(x(t)) - F^* = \mathcal{O}\left(e^{\sqrt{\mu}t}\right)$.

The Nesterov's accelerated gradient method

Nesterov 1983

$$\begin{aligned} y_k &= x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - s\nabla F(y_k) \end{aligned}$$

where the sequence $(t_k)_{k\in\mathbb{N}}$ is defined by: $t_1 = 1$ and: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

A modified version (Chambolle Dossal 2015)

$$\begin{aligned} y_k &= x_k + \frac{k}{k+\alpha}(x_k - x_{k-1}), \quad \alpha \geqslant 3 \\ x_{k+1} &= y_k - s\nabla F(y_k) \end{aligned}$$

For the class of convex functions, the sequence of iterates satisfies:

$$\forall k \in \mathbb{N},\ F(x_k) - F^* \leqslant \frac{(\alpha+1)\|x_0 - x^*\|^2}{2sk^2}$$

and, for the modified version with $\alpha > 3$, weakly converges to a minimizer of F.

Link between the ODE and the optimization scheme

Discretization of an ODE, Su Boyd and Candès (2015)

$$x_{n+1} = y_n - h\nabla F(y_n) \text{ with } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1})$$

can be seen as a semi-implicit discretization of a solution of

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

With $\dot{x}(t_0) = 0$. Move of a solid in a potential field with a vanishing viscosity $\frac{\alpha}{t}$.

General methodology to analyze optimization algorithms

- Interpreting the optimization algorithm as a discretization of a given ODE.
- Analysis of ODEs using a Lyapunov approach:

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

- Building a sequence of discrete Lyapunov energies adapted to the optimization scheme to get the same decay rates

Convergence analysis of the Nesterov gradient method

Convergence rate in the continuous setting

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function and $x^* \in \arg\min(F) \neq \emptyset$.

- If $\alpha \geqslant 3$,

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^2}\right)$$

[Attouch, Chbani, Peypouquet, Redont 2016]

- If $\alpha > 3$, then $x(t)$ cv to a minimizer of F and:

$$F(x(t)) - F(x^*) = o\left(\frac{1}{t^2}\right)$$

[Su, Boyd, Candes 2016]
[Chambolle, Dossal 2015]
[May 2017]

- If $\alpha < 3$ then no proof of cv of $x(t)$ but:

$$F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$$

[Attouch, Chbani, Riahi 2019]
[Aujol, Dossal 2017]

The Nesterov's accelerated gradient method

For the class of convex functions

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function with $X^* := \arg\min(F) \neq \emptyset$.

$$\left| \begin{array}{rcl} y_n & = & x_n + \dfrac{n}{n+\alpha}(x_n - x_{n-1}) \\ x_{n+1} & = & y_n - h\nabla F\,(y_n) \end{array} \right. , \quad \alpha > 0,\ h < \frac{1}{L}$$

- If $\alpha \geqslant 3$

$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^2}\right)$$

[Nesterov 1984, Su, Boyd, Candes 2016, Chambolle Dossal 2015, Attouch et al. 2018]

- If $\alpha > 3$, then $(x_n)_{n\geqslant 1}$ cv and:

$$F(x_n) - F(x^*) = o\left(\frac{1}{n^2}\right)$$

[Chambolle, Dossal 2015]
[Attouch, Peypouquet 2015]

- If $\alpha \leqslant 3$

$$F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^{\frac{2\alpha}{3}}}\right).$$

[Attouch, Chbani, Riahi 2018]
[Apidopoulos, Aujol, Dossal 2018]

GD vs Nesterov in the strongly convex case

Exponential rate vs Polynomial rate

Assume now that F is additionally μ-strongly convex, or satisfies some quadratic growth condition:

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

Convergence rate for GD

$$\forall n \in \mathbb{N},\ F(x_n) - F^* = \mathcal{O}\left((1-\kappa)^n\right).$$

The number of iterations required to reach an ε-solution is: $n_\varepsilon^{FB} \sim \frac{1}{\kappa} \log\left(\frac{2L}{\varepsilon^2} M_0\right)$.

Convergence rate for Nesterov's accelerated GD [Candès et al 2015], [Attouch Cabot 2017], [ADR 2018].

If F has a unique minimizer,

$$\forall \alpha > 0,\ \forall n \in \mathbb{N},\ F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$$

Nesterov accelerated algorithm for strongly convex functions

$$y_n = x_n + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(x_n - x_{n-1})$$
$$x_{n+1} = y_n - \tfrac{1}{L}\nabla F(y_n)$$

Theorem (Theorem 2.2.3, Nesterov 2013)

Assume that F is μ-strongly convex for some $\mu > 0$. Let $\varepsilon > 0$. Then for $\kappa = \frac{\mu}{L}$ small enough,

$$\forall n \in \mathbb{N},\ F(x_n) - F(x^*) \leqslant 2(1-\sqrt{\kappa})^n \left(F(x_0) - F(x^*)\right),$$

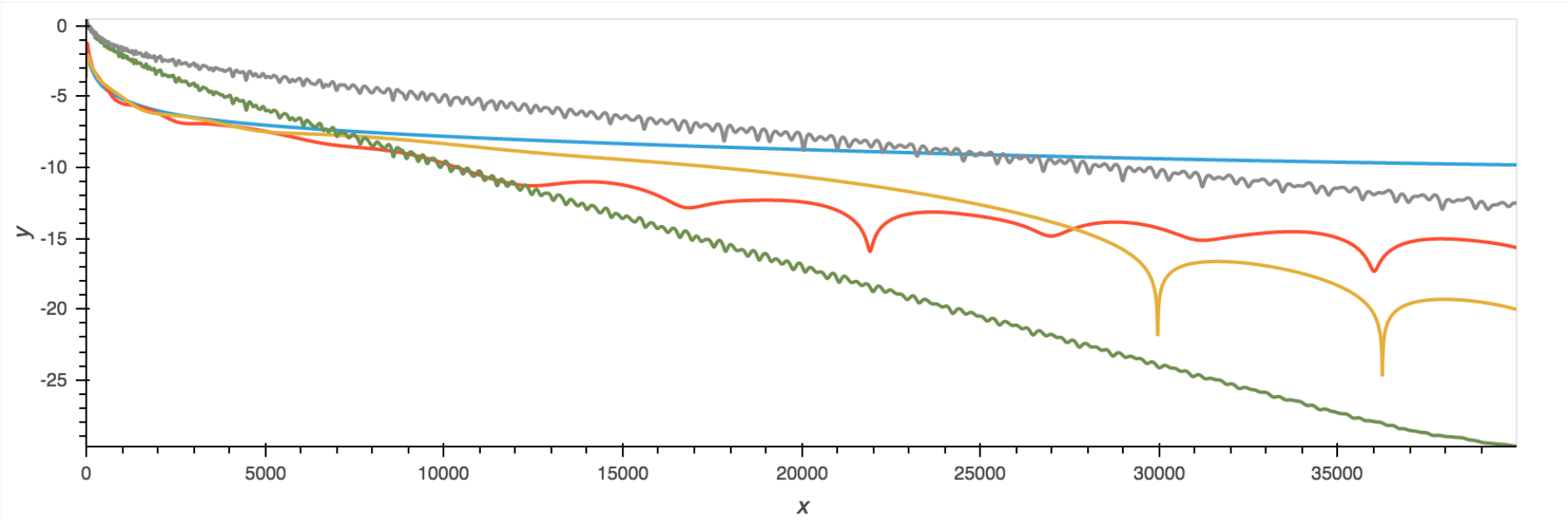
which means that an ε-solution can be obtained in at most:

$$n_\varepsilon^{NSC} = \frac{1}{|\log(1-\sqrt{\kappa})|} \log\left(\frac{4LM_0}{\varepsilon^2}\right). \tag{1}$$

The iterations require an estimation of $\kappa = \frac{\mu}{L}$!

FISTA in the strongly convex case

Numerical comparison



$\log(\|g(x_n)\|)$ along the iterations

FB, FISTA with $\alpha = 8$, FISTA with $\alpha = 30$,

NSC with the true value of μ, NSC with $\widetilde{\mu} = \frac{\mu}{10}$.

FISTA is efficient without knowing μ and its convergence rate does not suffer from any underestimation of μ

FISTA in the strongly convex case

A seemingly disappointing rate...

Convergence rate of FISTA under quadratic growth

For functions satisfying $F(x) - F^* \geqslant \frac{\mu}{2} d(x, X^*)^2$, the sequence generated by FISTA with parameter α satisfies:

$$F(x_n) - F(x^*) \leqslant A_1 \left(\sqrt{\frac{L}{\mu}} \frac{A_2\, \alpha}{n} \right)^{2\alpha/3}$$

with explicit constants A_1, A_2.

First impression

At first sight: worse than GD, far from Heavy Ball!

- Polynomial decay vs exponential
- Seems to contradict excellent practical performance of FISTA

FISTA in the strongly convex case

The surprise: optimal choice of α

Optimization over α for a target precision ε

Minimizing n_ε^{FISTA} over α yields:

$$\alpha_\varepsilon = 3 \log\left(\frac{5\sqrt{L(F(x_0)-F^*)}}{e\varepsilon}\right) \sim \log(1/\varepsilon) \quad \text{(does not require } \mu\text{!)}$$

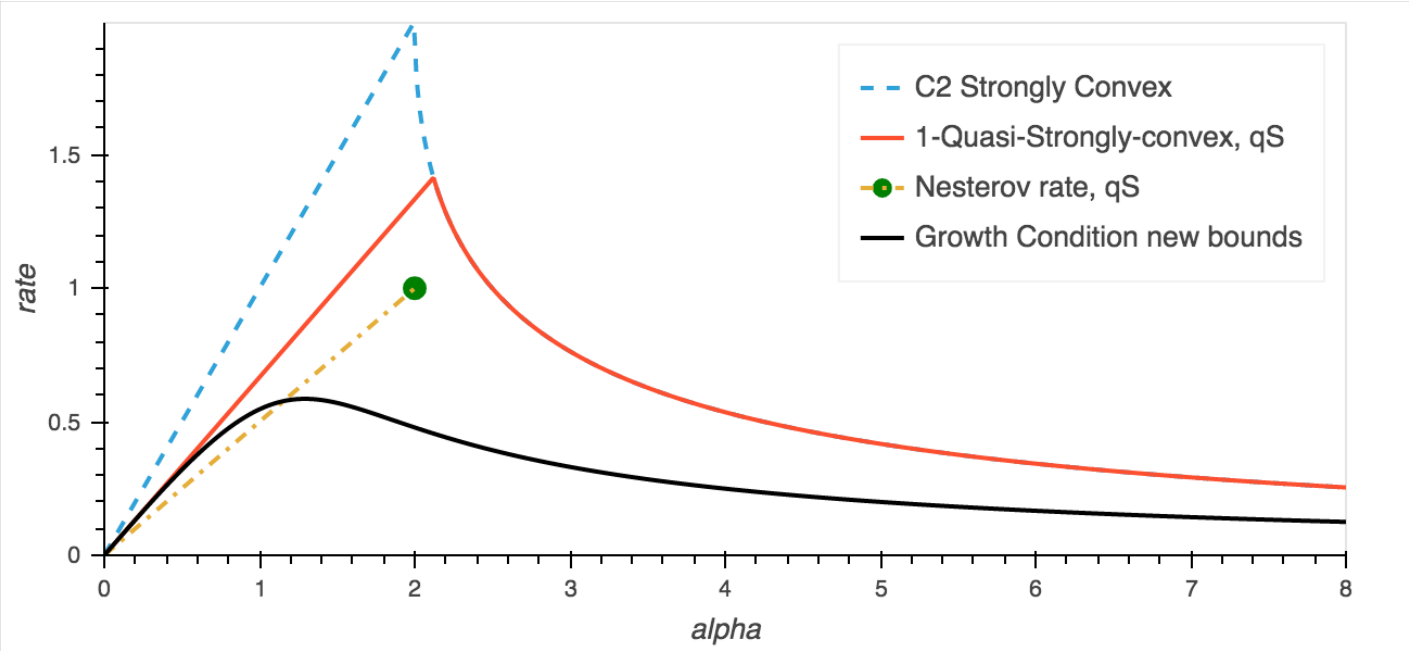
and the corresponding number of iterations:

$$\boxed{n_\varepsilon^{FISTA} := \frac{8e^2}{3\sqrt{\kappa}}\alpha_\varepsilon = \frac{8e^2}{\sqrt{\kappa}} \log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right)}$$

Surprise!

- Comparable to optimal Heavy Ball: $n_\varepsilon^{HB} \sim \sqrt{L/\mu}\,\log(1/\varepsilon)$
- **Without any knowledge of μ!**
- Key: choose $\alpha \sim \log(1/\varepsilon)$, i.e. tune friction to the target precision

Interpretation: surfing the wave



FISTA strategy

Decreases α so that the algorithm stops near the top of the wave

Trade-off

- **HB**: optimal at any precision (if μ known)
- **FISTA**: quasi-optimal at target precision ε (without μ)

Key insight: Automatic geometric adaptation of friction !

Explains excellent practical performance of FISTA.

Convergence rate analysis under some quadratic growth condition

Theorem (Aujol Dossal R. 2023, Aujol Dossal Labarrière R. 2024)

Let $\varepsilon > 0$ *and*

$$\alpha_\varepsilon := 3 \log \left(\frac{5\sqrt{L(F(x_0) - F^*)}}{e\varepsilon} \right)$$ *does not depend on any estimation of* μ.

Let $(x_n)_{n \in \mathbb{R}^N}$ *be a sequence of iterates generated by the Nesterov's accelerated GD with parameter* α_ε. *Then for* $\kappa = \frac{\mu}{L}$ *small enough, an* ε*-solution is reached in at most:*

$$n_\varepsilon^{FISTA} := \frac{8e^2}{3\sqrt{\kappa}} \alpha_\varepsilon = \frac{8e^2}{\sqrt{\kappa}} \log \left(\frac{5\sqrt{LM_0}}{e\varepsilon} \right)$$

iterations.

Theorem (Aujol, Dossal, Labarrière, R. 2024)

If F *satisfies some local quadratic growth condition then, for* α *large enough, the sequence* $(x_k)_{k \in \mathbb{N}}$ *generated by Nesterov GD/FISTA* **strongly** *converges to a minimizer of* F *and the trajectory of iterates has finite length.*

Conclusion

Main results

1. FISTA achieves quasi-optimal rate $O(\sqrt{L/\mu}\log(1/\varepsilon))$
2. **Without knowledge of** μ (automatic adaptation)
3. By choosing $\alpha \sim \log(1/\varepsilon)$
4. Extends to non-differentiable problems (LASSO, etc.)

Key insight

FISTA "surfs the wave":

- Adapts friction geometrically to target precision
- Balances polynomial decay vs optimal friction zone
- Simple, efficient, robust

Reference: J-F. Aujol, Ch. Dossal, A. Rondepierre, *Math. Programming, Series A*, 2023

Conclusion (2/2)

- Restarting FISTA can improve the convergence rate
  - If F is $\mu-$strongly convex, restarting FISTA each $e\sqrt{\frac{L}{\mu}}$ ensures an exponential decay... but μ may be unknown.
  - Estimation of μ: Alamo et al 2020, Fercoq et al. 2023, Aujol Calatroni Dossal R. Labarrière 2024...
- High resolution ODEs enables a more accurate description of the trajectories of the optimization algorithm.
  - Since 2016 Attouch and co-authors combine a Hessian-driven damping term to an asymptotic vanishing damping term resulting in

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

  - The HB scheme

$$\begin{cases} y_n &= x_n + \alpha(x_{n-1} - x_n) \\ x_{n+1} &= y_n - s\nabla F(x_n) \end{cases} \tag{2}$$

  is associated to the following High Resolution ODE (Shi et al 2018)

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + (1 + \sqrt{\mu s})\nabla F(x(t)) = 0. \tag{3}$$