An overview of accelerated methods in convex optimization

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

Financed by the French Agence Nationale de la Recherche (ANR) under reference ANR-PRC-CE23 MaSDOL

Institut de Mathématiques de Toulouse, INSA Toulouse, Institut de Mathématiques de Bordeaux

Séminaire Image Optimisation et Probabilités, 12th January 2023

Plan

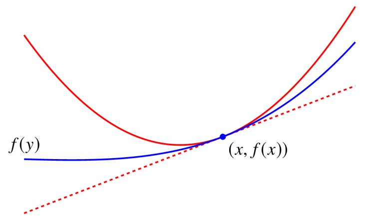
Framework and motivations

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:

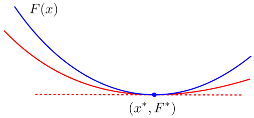
How fast does $F(x_k) - F^*$ decreases?

Framework and motivations

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 ($\mathcal{G}_\mu^2$):**
  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:

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

Framework and motivations

Classical algorithms

Gradient Descent/Forward-Backward:

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

Nesterov's accelerated gradient/FISTA (Beck and Teboulle, 2009):

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

where $\alpha > 0$. In general, $\alpha = 3$.
Heavy-Ball type schemes:

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

where $\alpha \in (0,1)$.

Plan

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

Gradient descent $\rightarrow$ Gradient flow

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

The continuous setting: a guideline for the discrete analysis

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

Gradient descent$\rightarrow$ Gradient flow

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

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

The continuous setting: a guideline for the discrete analysis

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

Gradient descent$\rightarrow$ Gradient flow

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

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

$$\downarrow$$

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

The continuous setting: a guideline for the discrete analysis

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

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

$$\downarrow$$

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

Heavy-Ball schemes→ Heavy-Ball Friction system

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

$$\downarrow$$

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

The continuous setting: a guideline for the discrete analysis

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?

The continuous setting: a guideline for the discrete analysis

Convergence results

Convergence rates of $F(x(t)) - F^*$:

	F convex	F μ-strongly convex
Gradient flow (Gradient descent)	$O\left(t^{-1}\right)$	$O\left(e^{-\mu t}\right)$
Heavy-Ball friction (Heavy-Ball schemes)	$O\left(t^{-1}\right)$	$O\left(e^{-2\sqrt{\mu}t}\right)$ if F is C^2
Asymptotic Vanishing Damping (Nesterov's accelerated gradient)	$O(t^{-2})$	$O\left(t^{-\frac{2\alpha}{3}}\right)$ Actually faster in finite time (see [1])

[1] J-F Aujol, Charles Dossal, Aude Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. 2021. ⟨hal-03491527⟩

Plan

Restart strategies

About inertia

Recall the definition of Nesterov's accelerated gradient/FISTA:

$$\forall k > 0, \begin{cases} x_k = y_{k-1} - s\nabla F(y_{k-1}), \\ y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}) \end{cases}$$

$\rightarrow$ taking in account the previous iterates generates inertia.

Issue

Under growth assumptions such as $\mathcal{SC}_\mu$ or $\mathcal{G}^2_\mu$, inertial methods have to be **parametrized according to the growth parameter** μ to reach fast convergence rates.

$\rightarrow$ μ **is rarely computable!!**

Restart strategies

Restarting FISTA, why?

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

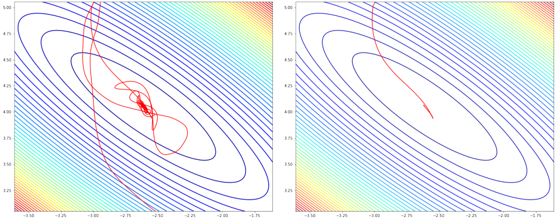


Figure: Trajectory of the iterates of FISTA (left) and FISTA restart (right) for a least-squares problem ($N = 20$).

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
    $k = k+1, i = i+1$
    $x_k = y_{k-1} - s\nabla f(y_{k-1})$
    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

Objective: get a restart condition that

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

Restart strategies

Empiric FISTA restart (O'Donoghue and Candès, 2015, Beck and Teboulle, 2009)

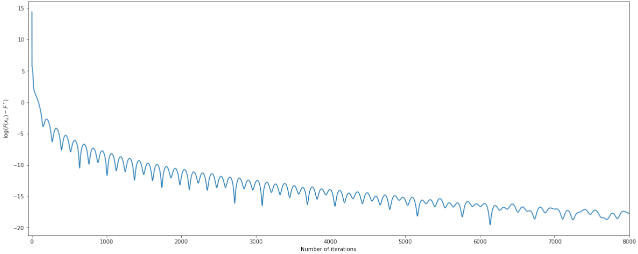
Restart under some exit condition

- 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

Fixed FISTA restart (Necoara et al., 2019)

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

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

Restart strategies

Fixed FISTA restart (Necoara et al., 2019)

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

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

Adaptive FISTA restart (Alamo et al., 2019, Fercoq and Qu, 2019)

Restart according to the geometry of F and previous iterations.

- Adaptive restart by Alamo et al.: $F(x_k) - F^* = O\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$.
- Adaptive restart by Fercoq and Qu: $F(x_k) - F^* = o\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}(2-\sqrt{\frac{\mu}{\mu_0}})}\sqrt{\frac{\mu}{L}}k}\right)$.

Restart strategies

Strategy of our scheme:

- to estimate the growth parameter μ at each restart,
- to adapt the number of iterations of the following restart according to this estimation.
- to stop the algorithm when the exit condition $\|\nabla F(r_j)\| \leqslant \varepsilon$ is satisfied.

Algorithm 2 : Automatic FISTA restart

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

$n_0 = \lfloor 2C \rfloor$

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

$n_1 = \lfloor 2C \rfloor$

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 $\|\nabla F(r_j)\| \leqslant \varepsilon$

Restart strategies

Theorem (Aujol, Dossal, L., Rondepierre, 2021)

If F satisfies the assumptions stated before and $C > 4$, then

$$F(r_j^+) - F^* = O\left(e^{-\frac{\log\left(\frac{C^2}{4}-1\right)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right).$$

Let $C = 6.38$, then

$$F(r_j^+) - F^* = O\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right).$$

Restart strategies

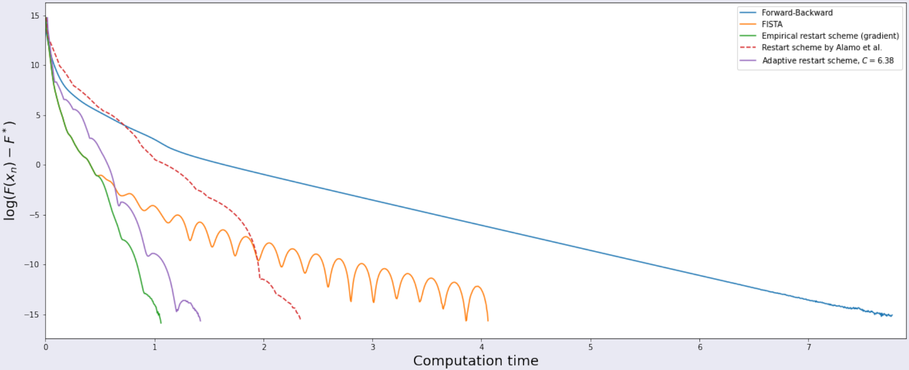
Image inpainting:

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

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

Restart strategies

Image inpainting:



Restart strategies

Summary:

Algorithm	Convergence rate
Forward-Backward	$O\left(e^{-\frac{\mu}{L}k}\right)$
Optimal FISTA restart	$O\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$
Empirical FISTA restart	$O(k^{-2})$
FISTA restart by Fercoq and Qu	$O\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}(2-\sqrt{\frac{\mu}{\mu_0}})}\sqrt{\frac{\mu}{L}}k}\right)$
FISTA restart by Alamo et al.	$O\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$
Automatic FISTA restart	$O\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right)$

Restart strategies

Adding backtracking on L (joint work with Luca Calatroni, to be submitted)

This restart strategy can be extended to functions whose Lipschitz constant L **cannot be estimated** or poorly: this involves a variant of FISTA which estimates L using **backtracking**.

$\rightarrow$ The method is fully automatic while ensuring a fast convergence rate.

Plan

Attenuating oscillations introducing Hessian-driven damping

Hessian-driven damping

(DIN-AVD) system (**Attouch, Peypouquet and Redont, 2016**)

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

- Attenuation of the oscillations through the introduction of a geometry-driven damping term.

Attenuating oscillations introducing Hessian-driven damping

Hessian-driven damping

(DIN-AVD) system (**Attouch, Peypouquet and Redont, 2016**)

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

- Attenuation of the oscillations through the introduction of a geometry-driven damping term.

Integrability properties

- **Attouch, Peypouquet and Redont, 2016**: if F is convex and C^2, $\alpha \geqslant 3$ and $\beta > 0$:

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

- **Aujol, Dossal, Hoàng, L. and Rondepierre, 2022**: if F is convex and C^2, satisfies $\mathcal{G}_\mu^2$ and has a unique minimizer. Then, for $\alpha \geqslant 3$ and $\beta > 0$:

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

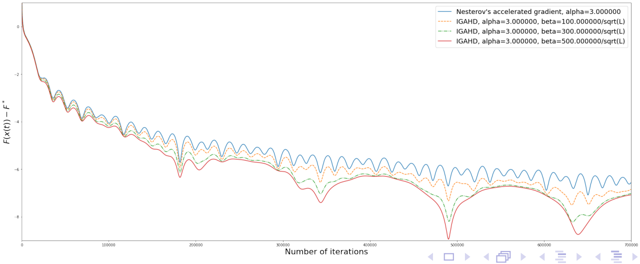
Attenuating oscillations introducing Hessian-driven damping

Derivating a numerical scheme: IGAHD (Attouch, Chbani, Fadili and Riahi, 2020)

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

$$\downarrow$$

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



Attenuating oscillations introducing 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.

Plan

Conclusion

Other comments/questions:

- How can high-resolution ODEs (see [2]) improve convergence analysis?
- Is it possible to adapt geometry parameter estimation to Heavy-Ball type methods?
- Can we combine restart with parallel calculations?
- Are inertial methods still fast without uniqueness of the minimizer? (current work)

[2] Shi, B., Du, S.S., Jordan, M.I. et al. Understanding the acceleration phenomenon via high-resolution differential equations. Math. Program. 195, 79–148 (2022). https://doi.org/10.1007/s10107-021-01681-8

Conclusion

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. 2021. ⟨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. 2022. ⟨hal-03693218v2⟩

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

I am open to post-doc offers!:)

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.
FISTA is an automatic geometrically optimized algorithm for strongly convex functions.
HAL preprint: hal-03491527, 2021.

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.

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.

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.

References II

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.

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.

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.