Inertia in Optimization: Acceleration and Adaptivity

Hippolyte Labarrière

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

DAO Team seminar
Laboratoire Jean Kuntzmann
18 November 2024

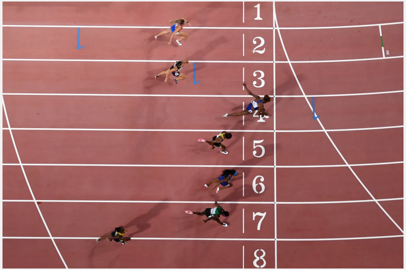
MaLGa MACHINE LEARNING GENOA CENTER

Università di Genova

Optimization, what is this?

$\rightarrow$ Find a set of parameters that minimizes a quantity.

Find the route that minimizes journey time.



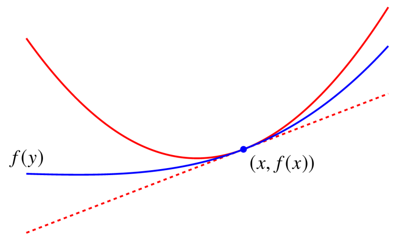
Find the training that leads to the best 100-meter time.

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^*.

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)

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

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

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

Inertia

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:

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

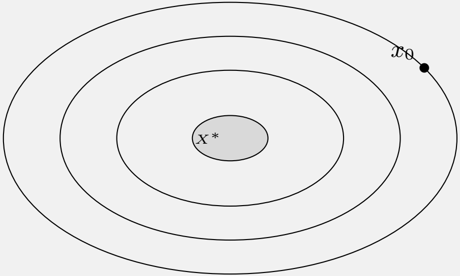
$\rightarrow$ Simple but slow!

Inertia in optimization

A classical algorithm: the proximal gradient method

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

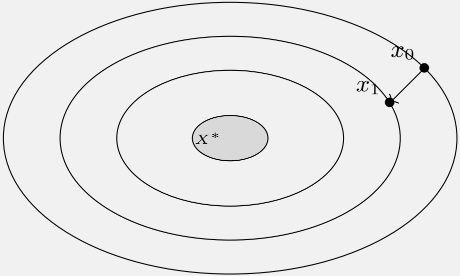
Illustration



A classical algorithm: the proximal gradient method

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

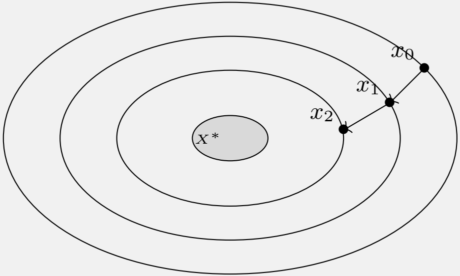
Illustration



A classical algorithm: the proximal gradient method

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

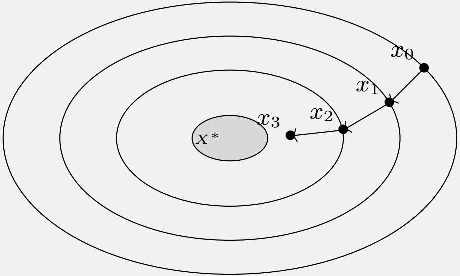
Illustration



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

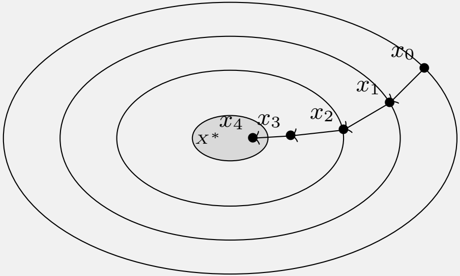


Inertia in optimization

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

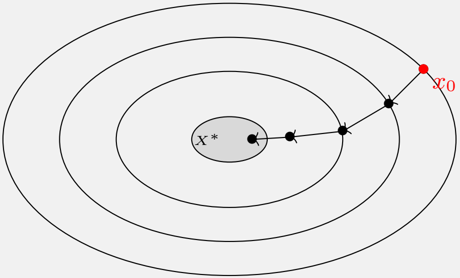


Introducing inertia

→ 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

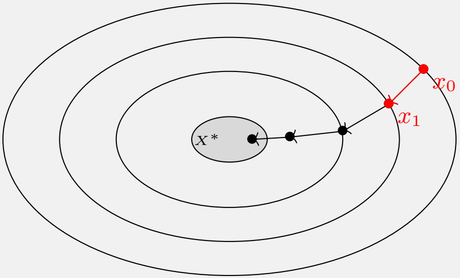


Introducing inertia

$\rightarrow$ 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

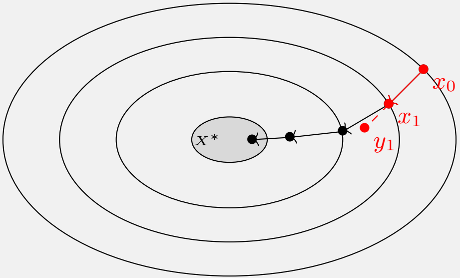


Introducing inertia

→ 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



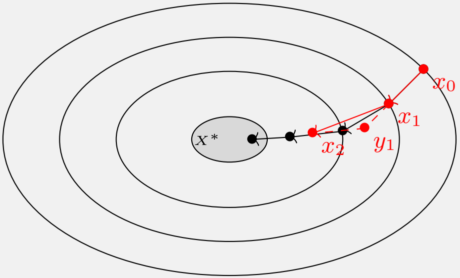
Inertia

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

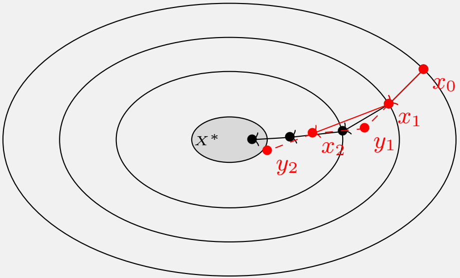


Introducing inertia

→ 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



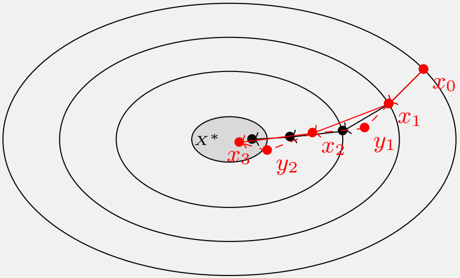
Inertia

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

→ Apply the same transformation to a shifted point.

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

Rising question

How to chose α_k?

Inertia

Introducing inertia

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

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

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

Geometry of convex functions

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

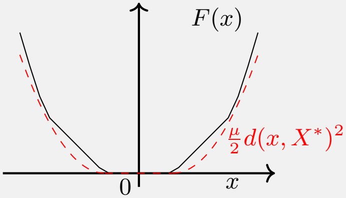
Classical geometry assumptions

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

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

Practical example: LASSO problem:

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



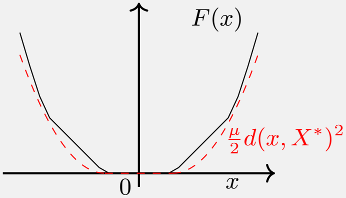
Classical geometry assumptions

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

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

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

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

Outline

Inertia between convexity and strong convexity

Framework

$$\min_{x\in\mathbb{R}^N} F(x) \text{ for } F \text{ satisfying some geometry assumption.}$$

What did we know?

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$	$\mathcal{H}^\gamma$	Convexity
PGD	$e^{-\frac{\mu}{L}k}$			k^{-1}
Heavy-Ball	$e^{-2\sqrt{\frac{\mu}{L}}k}$			k^{-1}
FISTA	$k^{-\frac{2\alpha}{3}}$			k^{-2}

Inertia between convexity and strong convexity

Framework

$$\min_{x\in\mathbb{R}^N} F(x) \text{ for } F \text{ satisfying some geometry assumption.}$$

What did we know?

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$	$\mathcal{H}^\gamma$	Convexity
PGD	$e^{-\frac{\mu}{L}k}$	$e^{-\frac{\mu}{L}k}$	$k^{-\frac{\gamma}{\gamma-2}}$	k^{-1}
Heavy-Ball	$e^{-2\sqrt{\frac{\mu}{L}}k}$	$e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}$	$k^{-\frac{\gamma}{\gamma-2}}$ *	k^{-1}
FISTA	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\gamma}{\gamma-2}}$	k^{-2}

If F has a unique minimizer!!

*in the continuous setting (Begout et al., '15).

Inertia between convexity and strong convexity

Framework

$$\min_{x\in\mathbb{R}^N} F(x) \text{ for } F \text{ satisfying some geometry assumption.}$$

What did we know?

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$	$\mathcal{H}^\gamma$	Convexity
PGD	$e^{-\frac{\mu}{L}k}$	$e^{-\frac{\mu}{L}k}$	$k^{-\frac{\gamma}{\gamma-2}}$	k^{-1}
Heavy-Ball	$e^{-2\sqrt{\frac{\mu}{L}}k}$	$e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}$	$k^{-\frac{\gamma}{\gamma-2}}$ *	k^{-1}
FISTA	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\gamma}{\gamma-2}}$	k^{-2}

If F has a unique minimizer!!

Is it really necessary?

*in the continuous setting (Begout et al., '15).

How to avoid the uniqueness assumption?

Our strategy

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

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

where $F = f + h$ is such that $\frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*$ for any $x \in \mathbb{R}^N$.

Inertia between convexity and strong convexity

How to avoid the uniqueness assumption?

Our strategy

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

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

where $F = f + h$ is such that $\frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*$ for any $x \in \mathbb{R}^N$.
Classical discrete Lyapunov energy for this system:

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

How to avoid the uniqueness assumption?

Our strategy

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

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

where $F = f + h$ is such that $\frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*$ for any $x \in \mathbb{R}^N$.
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^*.

How to avoid the uniqueness assumption?

Our strategy

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

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

where $F = f + h$ is such that $\frac{\mu}{2}d(x, X^*)^2 \leqslant F(x) - F^*$ for any $x \in \mathbb{R}^N$.
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!

Inertia between convexity and strong convexity

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 $\mathcal{G}^2_\mu$

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

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

Inertia between convexity and strong convexity

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 F is coercive and 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 $\alpha > 5 + \frac{8}{\gamma-2}$:

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

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 F is coercive and 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.

Inertia between convexity and strong convexity

What do we know now?

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$	$\mathcal{H}^\gamma$	Convexity
PGD	$e^{-\frac{\mu}{L}k}$	$e^{-\frac{\mu}{L}k}$	$k^{-\frac{\gamma}{\gamma-2}}$	k^{-1}
Heavy-Ball	$e^{-2\sqrt{\frac{\mu}{L}}k}$	$e^{-\frac{2}{3\sqrt{3}}\sqrt{\frac{\mu}{L}}k}$	$k^{-\frac{\gamma}{\gamma-2}}$	k^{-1}
FISTA	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\gamma}{\gamma-2}}$	k^{-2}

Take-away message

Inertia is **not impacted** by the non uniqueness of the minimizers.

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

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 Iterates under Hölderian and Quadratic Growth Conditions, 2024, arxiv:2407.17063.

Outline

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):

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

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

Restarting FISTA, why?

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

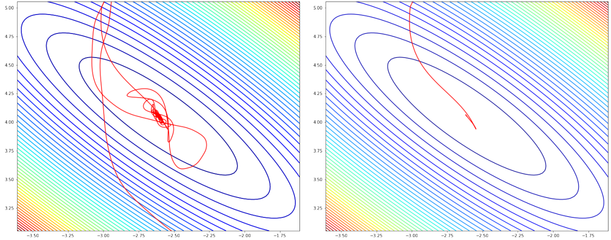


Figure: Projection of the trajectory of the iterates of FISTA (left) and FISTA restart (right) for a least-squares problem ($N = 20$).

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

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

→ Restarting when detecting an oscillation

- via F:

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

- via ∇F:

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

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

→ Restarting when detecting an oscillation

- via F:

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

- via ∇F:

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

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.

Restart strategies

Adaptive FISTA restart

Restart according to the geometry of F and previous iterations.

- Fercoq and Qu, '19: $F(x_k) - F^* = o\left(\exp\left(-\frac{\sqrt{2}-1}{2\sqrt{e}\left(2-\sqrt{\frac{\mu}{\mu_0}}\right)}\sqrt{\frac{\mu}{L}}k\right)\right)$.
- 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: $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{2\sqrt{2}}\sqrt{\frac{\mu}{L}}k}\right)$.

Restart strategies

Introduction of an automatic restart scheme (Aujol, Dossal, L., Rondepierre,'21)

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.

Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. FISTA restart using an automatic estimation of the growth parameter, 2021, ⟨hal-03153525v4⟩.

Restart strategies

Algorithm 2 : Automatic FISTA restart

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

$n_0 = \lfloor 2C \rfloor$

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

$n_1 = \lfloor 2C \rfloor$

repeat

$j = j + 1$

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

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

Summary:

Algorithm	Convergence rate
Forward-Backward	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
V-FISTA	$\mathcal{O}\left(e^{-\frac{9}{20}\sqrt{\frac{\mu}{L}}k}\right)$
Optimal FISTA restart	$\mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$
Heuristic FISTA restart	$\mathcal{O}(k^{-2})$
Fercoq and Qu '19	$\mathcal{O}\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}(2-\sqrt{\frac{\mu}{\mu_0}})}\sqrt{\frac{\mu}{L}}k}\right)$
Alamo et al. '19	$\mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$
Alamo et al. '22	$\mathcal{O}\left(e^{-\frac{\ln(15)}{4e}\sqrt{\frac{\mu}{L}}k}\right)$
Renegar and Grimmer '22	$\mathcal{O}\left(e^{-\frac{1}{2\sqrt{2}}\sqrt{\frac{\mu}{L}}k}\right)$
Automatic FISTA restart	$\mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right)$

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.

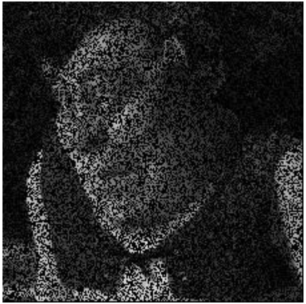
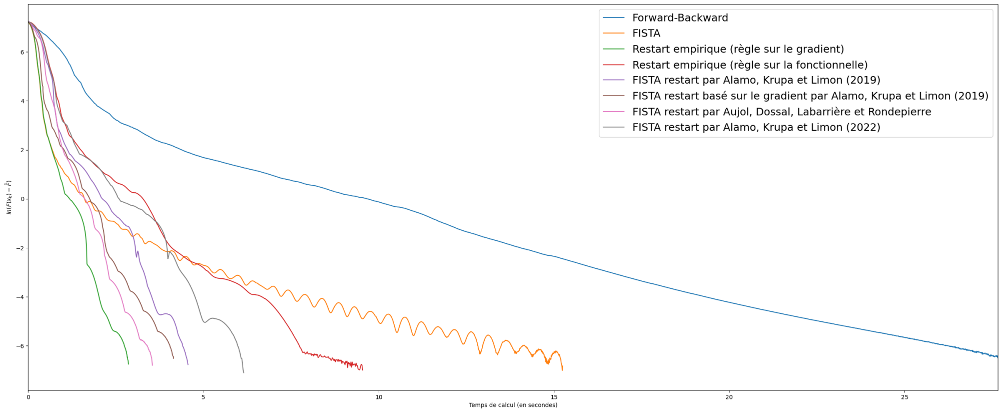


Image inpainting:



What if the Lipschitz constant L is not known?

Combining backtracking and restarting: Free-FISTA (Aujol, Calatroni, Dossal, L., Rondepierre, '24)

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

Jean-François Aujol, Luca Calatroni, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Parameter-Free FISTA by Adaptive Restart and Backtracking, 2024, *SIAM Journal on Optimization*.

An other approach

FISTA is far from optimal for functions satisfying strong growth conditions!

Recall

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}_\mu^2$
FISTA	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\alpha}{3}}$
Optimal FISTA restart	$e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}$	$e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}$

An other approach

FISTA is far from optimal for functions satisfying strong growth conditions!

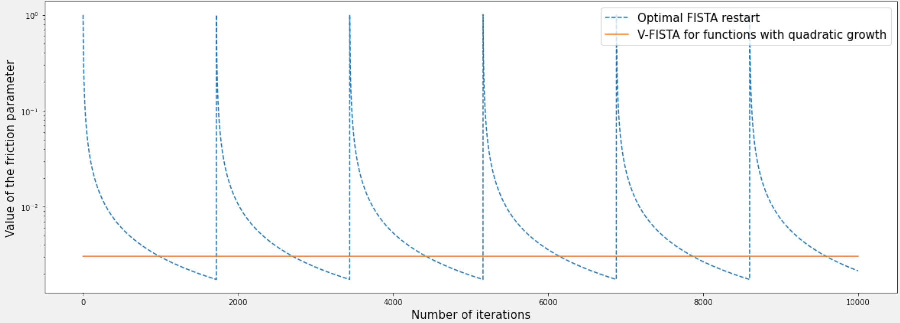
Recall

Algorithm	$\mathcal{SC}_\mu$	$\mathcal{G}^2_\mu$
FISTA	$k^{-\frac{2\alpha}{3}}$	$k^{-\frac{2\alpha}{3}}$
Optimal FISTA restart	$e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}$	$e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}$
V-FISTA (HB)	$e^{-\sqrt{\frac{\mu}{L}}k}$	$e^{-\frac{2}{3\sqrt{3}}\sqrt{\frac{\mu}{L}}k}$

Behavior of the friction parameter

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

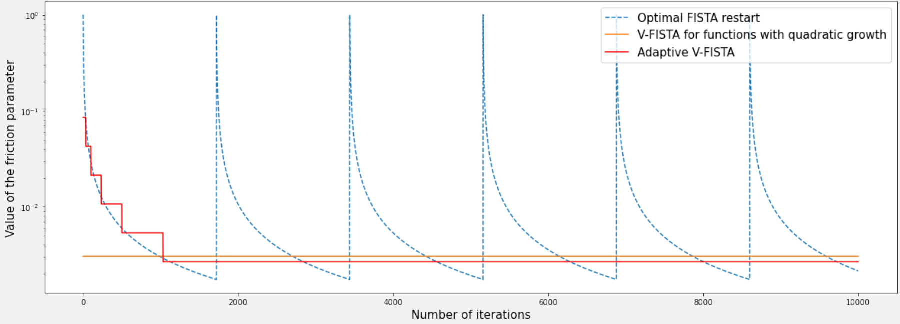
$\rightarrow$**Friction parameter:** $1 - \alpha_k$



Behavior of the friction parameter

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

$\rightarrow$**Friction parameter:** $1 - \alpha_k$



Keep piecewise constant friction to be faster!

An other approach

An adaptive procedure for fast methods (L., 2024)

Consider a method $\mathcal{A}$ generating $(x_k)_{k\in\mathbb{N}}$ such that

$$F(x_k) - F^* \leqslant Ae^{-B\sqrt{\frac{\mu}{L}}k}(F(x_0) - F^*)$$

for some $A, B > 0$ **if** $\frac{\mu}{L}$ **is available.**
$\rightarrow$ An adaptive scheme:

- that allows to apply $\mathcal{A}$ when $\frac{\mu}{L}$ is not known with **theoretical guarantees**.
- that can be combined with heuristic techniques (O'Donoghue and Candès, '15) for improved performance.
- which can be extended for methods involving backtracking on L (losing the theoretical guarantees).

Hippolyte Labarrière. Adaptive techniques for linearly fast methods with unknown condition number, currently in writing.

Conclusion

Take-away messages

· Inertia is **not impacted** by the non uniqueness of the minimizers.

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

· If the condition number is not known $\rightarrow$ FISTA restart... or Adaptive V-FISTA!

Pending questions:

- Could the Performance Estimation Problem (PEP) approach (Drori and Teboulle,'14, Taylor, Hendrickx and Glineur,'17, Taylor and Drori,'22 ...) allow to find tighter bounds?
- Then, could it help to build faster adaptive schemes?
- Can we obtain better convergence guarantees for adaptive step-size methods (Malitsky and Mishchenko,'20,'24, Barzilai-Borwein stepsize) under growth conditions?

Conclusion

Thank you for your attention!

Publications and 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, Luca Calatroni, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre. Parameter-Free FISTA by Adaptive Restart and Backtracking, 2024, *SIAM Journal on Optimization*.
- 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 Iterates under Hölderian and Quadratic Growth Conditions, 2024, arxiv:2407.17063.

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/

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.

The continuous setting

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

Gradient descent $\rightarrow$ Gradient flow

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

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

The continuous setting

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

Gradient descent $\rightarrow$ Gradient flow

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

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

$$\downarrow$$

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

The continuous setting

Nesterov's accelerated gradient→Asymptotic vanishing damping system (Su, Boyd and Candès, '14)

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

$$\downarrow$$

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

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?

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

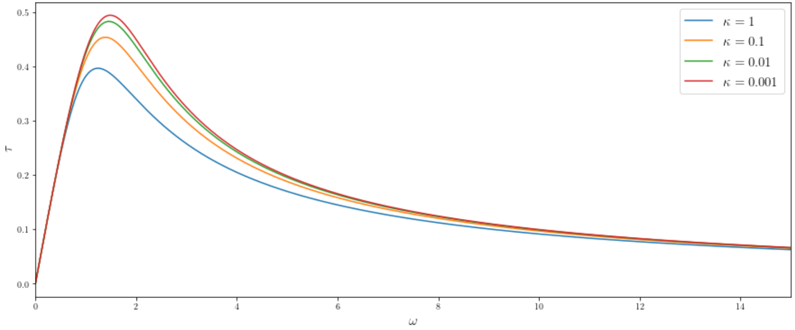
Main results: V-FISTA

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