FISTA is an automatic geometrically optimized algorithm for strongly convex functions

Aude Rondepierre

Joint work with Jean-François Aujol, Charles Dossal

and Hippolyte Labarrière

Institut de Mathématiques de Toulouse, INSA de Toulouse

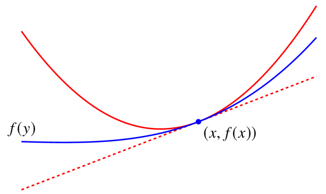
Workshop DIPOpt - Deep learning, image analysis, inverse problems, and optimization, 2023

The setting: composite optimization

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

where:

- f is a convex differentiable function with a L-Lipschitz gradient:



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

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

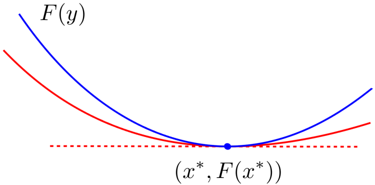
- h is a convex lower semicontinuous (lsc) *simple* function.

$\hookrightarrow$ Application to least square problems, LASSO ($\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - b\|^2 + \|x\|_1$).

$\hookrightarrow$ Applications in Image and Signal processing, machine learning, deep learning, AI,...

The setting: local geometry of convex functions

In this talk we assume that the composite convex function $F = f + h$ satisfies a quadratic growth condition around its set of minimizers:



Quadratic growth condition

Let $X^* = \arg\min F$ and $F^* = \min F$. There exists $\mu > 0$ such that:

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

Strong convexity

F is μ-strongly convex iff $F - \frac{\mu}{2}\|\cdot\|^2$ is convex. In the differentiable case:

$$\forall (x,y) \in \mathbb{R}^N \times \mathbb{R}^N,\ F(y) \geqslant F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2.$$

Quadratic growth is a relaxation of strong convexity

LASSO problem with A invertible

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_1$$

Then there exists $\mu > 0$ such that F is μ-strongly convex.

LASSO problem with A non injective

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_1$$

Then there exists $\mu > 0$ such that F satisfies $\mathcal{G}^2_\mu$, but F is not μ-strongly convex. [Bolte et al 2013]

The setting: Large scale optimization

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

where:

- f is a convex differentiable function with a L-Lipschitz gradient.
- h is a convex l.s.c. function.
- F satisfies some quadratic growth condition $\mathcal{G}^2_\mu$ where μ is not perfectly known.

Goal

- First order optimization methods i.e. methods that can only use the values of the function F and/or the values of its gradient (or subgradient).
- Assume that F has at least one minimizer x^*.
  - Speed in term of decrease of $F(x_k) - F(x^*)$
  - How to define a tractable stopping criterium ?

Outline

1. **Analyzing optimization algorithms for a given accuracy ε**
   - Notion of ε-solution
   - A tractable stopping criterion
2. **The Forward-Backward and FISTA algorithms**
   - The Forward-Backward algorithm
   - FISTA a fast proximal gradient method
   - FB vs FISTA in the strongly convex case
3. **FISTA is an automatic geometrically optimized algorithm for strongly convex functions**
   - The dynamical system intuition
   - Convergence rates under some quadratic growth condition
   - Comparisons
4. **Strong convergence of FISTA**

Analyzing optimization algorithms for a given accuracy ε

Notion of ε-solution

The minimizers of $F = f + h$ are characterized: $0 \in \partial F(x)$, or equivalently for any $\gamma > 0$,

$$x = prox_{\gamma h}(x - \gamma \nabla f(x))$$

where:

$$\text{prox}_{\gamma h}(x) = \arg\min_{y \in \mathbb{R}^N} \gamma h(y) + \frac{1}{2}\|y - x\|^2.$$

Definition (ε-solution)

Let

$$g(x) := L\left(x - \text{prox}_{\gamma h}(x - \frac{1}{L}\nabla f(x))\right)$$

be the composite gradient mapping associated to F, and $\varepsilon > 0$. An iterate x_n is said to be an ε-solution of $\min_{x \in \mathbb{R}^N} F(x)$ if:

$$\|g(x_n)\| \leqslant \varepsilon.$$

NB: in the differentiable case ($h = 0$) we have: $g(x) = \nabla f(x)$.

A tractable stopping criterion

$$\|g(x_n)\| \leqslant \varepsilon$$

Two useful properties:

1. $\forall x \in \mathbb{R}^N,\; F(x^+) - F^* \leqslant \frac{2}{\mu}\|g(x)\|^2$ [Aujol Dossal Labarrière R. 2021]
2. $\forall x \in \mathbb{R}^N,\; \frac{1}{2L}\|g(x)\|^2 \leqslant F(x) - F^*$ [Nesterov 2007]

A sufficient condition

If:

$$F(x_n) - F^* \leqslant \frac{1}{2L}\varepsilon^2,$$

then x_n is an ε-solution of $\min_{x\in\mathbb{R}^N} F(x)$.

General methodology

1. Getting bounds in finite time on $F(x_n) - F^*$.
2. Interpretation in terms of ε-solution: compute the number n of iterations required to reach an ε-solution of $\min_{x \in \mathbb{R}^N} F(x)$ i.e. such that:

$$F(x_n) - F^* \leqslant \frac{1}{2L}\varepsilon^2.$$

	Convergence rate $F(x_n) - F^*$	Nb n of iterations to reach a ε-solution prop. to
Polynomial decrease	$\frac{1}{n^\beta}$	$n \geqslant \left(\frac{2L}{\varepsilon^2}\right)^{\frac{1}{\beta}}$
Exponential decrease	$(1-\kappa)^n$	$n \geqslant \frac{2}{\lvert\log(1-\kappa)\rvert} \log\left(\frac{\sqrt{2L}}{\varepsilon}\right)$

Forward-Backward algorithm

A fixed point algorithm

Let $\gamma > 0$.The minimizers of the composite convex function $F = f + h$ are exactly characterized by:

$$x = prox_{\gamma h}\left(x - \gamma \nabla f(x)\right)$$

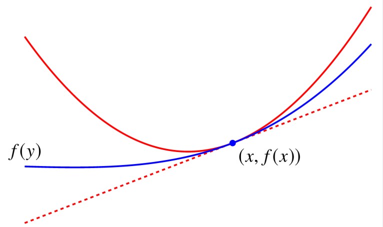
Forward-Backward algorithm

$$\begin{aligned} &x_0 \in \mathbb{R}^N \\ &x_{n+1} = prox_{\gamma h}(x_n - \gamma \nabla f(x_n)), \quad \gamma > 0. \end{aligned}$$

Interpretation

Instead of minimizing $F = f + g$, minimize at each iteration n its quadratic upper bound:

$$x \mapsto f(x_n) + \langle \nabla f(x_n), x - x_n \rangle + \frac{L}{2}\|x - x_n\|^2 + h(x)$$



- Gradient method ($h = 0$, unconstrained optimization). Then:

$$prox_{\gamma h}(x) = \arg\min_{y \in \mathbb{R}^N} \left(0 + \frac{1}{2}\|y - x\|^2 \right) = x$$

Hence: $x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n)$.

- Gradient method ($h = 0$, unconstrained optimization). Then:

$$prox_{\gamma h}(x) = \arg\min_{y \in \mathbb{R}^N} \left(0 + \frac{1}{2}\|y - x\|^2 \right) = x$$

Hence: $x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n)$.

- Gradient projection method ($h = i_C$, constrained convex optimization).

$$prox_{\gamma h}(x) = \arg\min_{y \in \mathbb{R}^N} \left(i_C(y) + \frac{1}{2}\|y - x\|^2 \right) = P_C^{\perp}(x).$$

Hence: $x_{n+1} = p_C^{\perp}(x_n - \frac{1}{L}\nabla f(x_n))$.

- Gradient method ($h = 0$, unconstrained optimization). Then:

$$prox_{\gamma h}(x) = \arg\min_{y\in\mathbb{R}^N}\left(0 + \frac{1}{2}\|y - x\|^2\right) = x$$

Hence: $x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n)$.

- Gradient projection method ($h = i_C$, constrained convex optimization).

$$prox_{\gamma h}(x) = \arg\min_{y\in\mathbb{R}^N}\left(i_C(y) + \frac{1}{2}\|y - x\|^2\right) = P_C^{\perp}(x).$$

Hence: $x_{n+1} = p_C^{\perp}(x_n - \frac{1}{L}\nabla f(x_n))$.

- Iterative Soft-Thresholding Algorithm (ISTA) ($h = \|\cdot\|_1$):

$$prox_{\gamma h}(x) = sign(x)\max(0, |x| - \gamma).$$

and: $x_{n+1} = prox_{\frac{1}{L}h}(x_n - \frac{1}{L}\nabla f(x_n))$.

$$(FB) \quad x_{n+1} = prox_{\gamma h}(x_n - \gamma \nabla f(x_n)), \quad \gamma > 0.$$

Convergence rates in the convex case

1. If $\gamma < \frac{2}{L}$ then (FB) is a descent algorithm and the iterates $(x_n)_{n\in\mathbb{N}}$ cv to a minimizer of F.
2. Let $\gamma = \frac{1}{L}$.
$$\forall n \geqslant 1,\ F(x_n) - F^* \leqslant \frac{2L\|x_0 - x^*\|^2}{n}$$

$$(FB) \quad x_{n+1} = prox_{\gamma h}(x_n - \gamma \nabla f(x_n)), \quad \gamma > 0.$$

Convergence rates in the convex case

1. If $\gamma < \frac{2}{L}$ then (FB) is a descent algorithm and the iterates $(x_n)_{n\in\mathbb{N}}$ cv to a minimizer of F.
2. Let $\gamma = \frac{1}{L}$.

$$\forall n \geqslant 1,\ F(x_n) - F^* \leqslant \frac{2L\|x_0 - x^*\|^2}{n} \leqslant \frac{1}{2L}\varepsilon^2$$

The number of iterations required by FB to reach an ε-solution is at most:

$$n_\varepsilon \geqslant \frac{4L^2}{\varepsilon^2}\|x_0 - x^*\|^2 = \mathcal{O}\left(\frac{L^2}{\varepsilon^2}\right).$$

FISTA an accelerated proximal gradient method

FISTA - Beck Teboulle 2009, Nesterov 1984

$$\begin{aligned} y_n &= x_n + \frac{t_n - 1}{t_{n+1}}(x_n - x_{n-1}) \\ x_{n+1} &= prox_{\frac{1}{L}h}\left(y_n - \frac{1}{L}\nabla f(y_n)\right). \end{aligned}$$

where $t_1 = 1$ and the sequence $(t_n)_{n\in\mathbb{N}}$ is determined as the positive root of:

$$t_{n+1}^2 - t_{n+1} = t_n^2.$$

For the class of convex functions, they prove:

$$F(x_n) - F^* \leqslant \frac{2L\|x_0 - x^*\|^2}{(n+1)^2}$$

[Nesterov 1984] The $\mathcal{O}\left(\frac{1}{n^2}\right)$ rate is optimal for first order methods in the class of convex functions.

FISTA a fast proximal gradient method

FISTA - Chambolle Dossal 2015, Su Boyd Candès 2016

$$\begin{aligned} y_n &= x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}) \qquad \alpha \geqslant 3 \\ x_{n+1} &= prox_{\frac{1}{L}h}\left(y_n - \frac{1}{L}\nabla f(y_n)\right). \end{aligned}$$

- Initially Nesterov (1984) proposed a choice equivalent to $\alpha = 3$. Convergence of iterates for $\alpha > 3$ [Chambolle-Dossal 2015].
- For the class of composite convex functions:

$$\forall n \geqslant 1,\ F(x_n) - F^* \leqslant \frac{L(\alpha-1)^2\|x_0 - x^*\|^2}{2(n+\alpha-2)^2}$$

The number of iterations required for FISTA to reach an ε-solution is in $\mathcal{O}\left(\frac{L^2}{\varepsilon}\right)$ which is better than FB.

FB vs FISTA in the strongly convex case
Exponential rate vs Polynomial rate (1/3)

Assume now that F additionally satisfies some quadratic growth condition:

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

Let $\kappa = \frac{\mu}{L}$ be the inverse of the conditioning.

Convergence rate for FB [Garrigos, Rosasco, Villa 2017]

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant (1-\kappa)^n (F(x_0) - F^*).$$

The number of iterations required to reach an ε-solution is:

$$n_\varepsilon^{FB} = \frac{1}{|\log(1-\kappa)|} \log\left(\frac{2L}{\varepsilon^2}(F(x_0) - F^*)\right) \sim \frac{1}{\kappa} \log\left(\frac{2L}{\varepsilon^2} M_0\right).$$

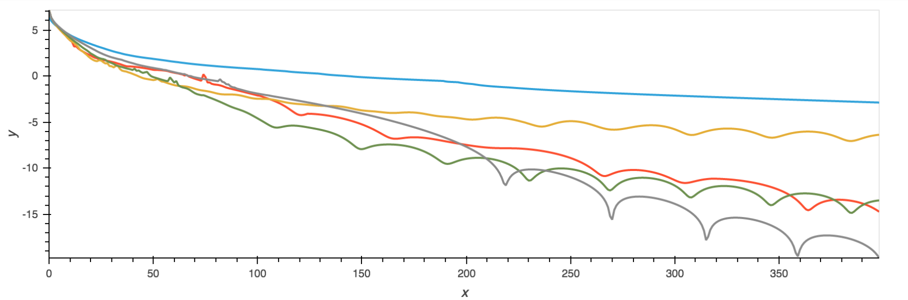
Convergence rate for FISTA [Candès et al 2015], [Attouch Cabot 2017], [ADR 2018].

Assume additionally that F has a unique minimizer.

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

FB vs FISTA in the strongly convex case

Exponential rate vs Polynomial rate (2/3)



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

FB, FISTA-restart, FISTA with $\alpha = 3$, FISTA with $\alpha = 12$, FISTA with $\alpha = 30$.

Motivation to provide a non-asymptotic analysis of FISTA and to compare rates in finite time !

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,

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

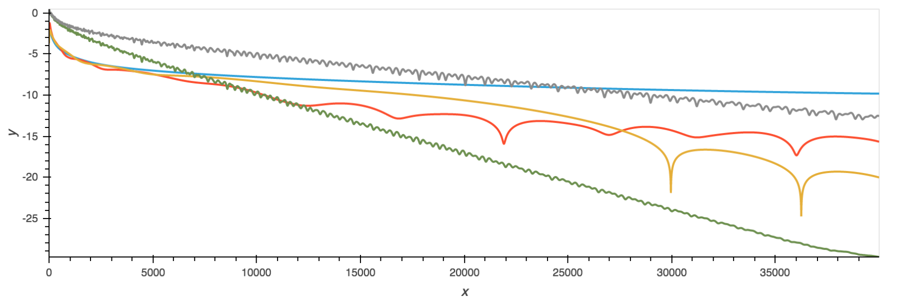
which means that an ε-solution can be obtained in at most:

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

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

FISTA in the strongly convex case

Differentiable case



$\log(\|g(x_n)\|)$ 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

General methodology to analyze optimization algorithms

- Interpreting the optimization algorithm as a discretization of a given ODE:

$$\text{Gradient descent iteration:} \quad \frac{x_{n+1} - x_n}{h} + \nabla F(x_n) = 0$$

$$\text{Associated ODE:} \quad \dot{x}(t) + \nabla F(x(t)) = 0.$$

- Analysis of ODEs using a Lyapunov approach:

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

  - $\mathcal{E}$ is decreasing along the trajectory, and thus $F(x(t)) - F^* = \mathcal{O}\left(\frac{1}{t}\right)$.
- Building a sequence of discrete Lyapunov energies adapted to the optimization scheme to get the same decay rates

Discretization of an ODE, Su Boyd and Candès (15)

The scheme defined by

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

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

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

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

(Discretization step: $h = \sqrt{s}$ and $x_n \simeq x(n\sqrt{s})$)

Convergence rate analysis for FISTA in finite time

Sketch of proof

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2, \quad \lambda = \frac{2\alpha}{3}.$$

Assume that F has a quadratic growth and a unique minimizer x^*.

1. Prove some differential inequation:

$$\forall t \geqslant t_0,\ \mathcal{E}'(t) + \frac{\lambda - 2}{t}\mathcal{E}(t) \leqslant \varphi(t)\mathcal{E}(t).$$

2. Integrate it between any t_1 and t:

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \leqslant \mathcal{E}(t_1)\left(\frac{t_1}{t}\right)^{\lambda-2} e^{\phi(t_1)}.$$

3. Choose t_1 such that the previous bound is as tight as possible:

$$\forall t \geqslant t_1,\ F(x(t)) - F^* \leqslant C_1 e^{\frac{2}{3}C_2(\alpha-3)}\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}}.$$

Convergence rate analysis for FISTA in finite time

How to tune α to get a fast exponential decay

Let ε be a given accuracy. Let us make some rough calculations:

- For any $\alpha > 3$, we have:

$$\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}} \leqslant \varepsilon \iff t \geqslant \frac{\alpha}{\sqrt{\mu}}\left(\frac{1}{\varepsilon}\right)^{\frac{3}{2\alpha}}$$

$\hookrightarrow$ Polynomial decay.

Let ε be a given accuracy. Let us make some rough calculations:

- For any $\alpha > 3$, we have:

$$\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}} \leqslant \varepsilon \iff t \geqslant \frac{\alpha}{\sqrt{\mu}}\left(\frac{1}{\varepsilon}\right)^{\frac{3}{2\alpha}}$$

  $\hookrightarrow$ Polynomial decay.

- Choose now:

$$\alpha = C\log\left(\frac{1}{\varepsilon}\right).$$

  Then

$$\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}} \leqslant \varepsilon \iff t \geqslant \frac{Ce^{\frac{3}{2C}}}{\sqrt{\mu}}\log\left(\frac{1}{\varepsilon}\right)$$

  $\hookrightarrow$ Exponential decay !

Convergence rate analysis in finite time [ADR 2021]

FISTA for composite optimization with a quadratic growth condition

Theorem

Let $\varepsilon > 0$ *and*

$$\alpha_\varepsilon := 3\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right) \qquad \textit{where:} \qquad M_0 = F(x_0) - F^*.$$

Let $(x_n)_{n\in\mathbb{R}^N}$ *be a sequence of iterates generated by the FISTA algorithm with parameter* $\alpha_{1,\varepsilon}$. *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.

- α_ε does not depend on μ or any estimation of μ.
- n_ε^{FISTA} depends on the real value of μ.
- Fast exponential decay.

Comparisons with Forward-Backward and Nesterov SC

Let $\varepsilon > 0$ and $\alpha = 3\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right)$.

Comparison with Forward-Backward algorithm

For $\kappa = \frac{\mu}{L}$ small enough,

$$n_\varepsilon^{FISTA} = \frac{4e^2}{\sqrt{\kappa}}\log\left(\frac{5LM_0}{e^2\varepsilon^2}\right) \leqslant n_\varepsilon^{FB} = \frac{1}{|\log(1-\kappa)|}\log\left(\frac{2LM_0}{\varepsilon^2}\right).$$

Comparison with Nesterov for strongly convex functions

Let $\varepsilon > 0$. If μ is known, for $\kappa = \frac{\mu}{L}$ small enough, NSC is faster than FISTA. But if μ is not perfectly known and for $\tilde{\mu} \leqslant \mu$

$$n_\varepsilon^{NSC} = \frac{1}{\left|\log(1-\sqrt{\frac{\tilde{\mu}}{L}})\right|}\log\left(\frac{4LM_0}{\varepsilon^2}\right) \geqslant \frac{1}{|\log(1-\sqrt{\kappa})|}\log\left(\frac{4LM_0}{\varepsilon^2}\right) \qquad (2)$$

In practice, FISTA may outperform NSC even for smaller underestimations of μ.

A first conclusion

	Geometry of F	References	Convergence rate for $F(x_n) - F^*$	Number of iterations to reach an ε solution
FB	Convex	N84, BT09	$\frac{2L\Vert x_0 - x^*\Vert^2}{n}$	$\frac{4L^2}{\varepsilon^2}\Vert x_0 - x^*\Vert^2$
FISTA with $\alpha = 3$	Convex	N84, BT09	$\frac{2L\Vert x_0 - x^*\Vert^2}{(n+1)^2}$	$\frac{2L}{\varepsilon}\Vert x_0 - x^*\Vert$
FB	Convex and $\mathcal{G}^2_\mu$	Garrigos 17	$(1+\kappa)^{-n}(F(x_0) - F^*)$	$\mathcal{O}\left(\frac{1}{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$
NSC	Strongly convex Requires estimate of μ	Nesterov 13	$2(1-\sqrt{\kappa})^n(F(x_0) - F^*)$	$\mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\log\left(\frac{1}{\varepsilon}\right)\right)$
FISTA $\alpha \geqslant 3$	Convex and $\mathcal{G}^2_\mu$ Uniqueness of minimizer	Attouch 18 ADR19	$\mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$	Unknown
FISTA $\alpha = 3\log\left(\frac{5\sqrt{LM_0}}{e\,\varepsilon}\right)$	Convex and $\mathcal{G}^2_\mu$ Uniqueness of minimizer	ADR23	$\mathcal{O}\left(e^{-Cn\sqrt{\kappa}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\log\left(\frac{1}{\varepsilon}\right)\right)$

- No need to estimate the growth parameter μ and the convergence rate does not suffer from an underestimation of μ.

J-F Aujol, Ch. Dossal, A.R. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. Mathematical Programming 2023.

Inertial methods without the uniqueness of the minimizer

All known improved convergence rates for first-order inertial methods rely on the assumption that F **has a unique minimizer**:

Algorithm	Strong convexity	$\mathcal{G}^2_\mu$ and unique minimizer	$\mathcal{G}^2_\mu$
Forward-Backward	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$
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 ($\alpha > 3$)	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right)$	$\mathcal{O}\left(k^{-2}\right)$

Is this hypothesis necessary to get fast convergence rates?

Strong convergence of FISTA

Theorem

If F satisfies some flat growth condition i.e. if there exists $\gamma \geqslant 2$ and $\eta > 0$ such that for any minimizer x^*,

$$\exists \eta > 0,\ \forall x \in B(x^*, \eta),\ Kd(x, X^*)^\gamma \leqslant F(x) - F^*$$

then, for α large enough, the sequence $(x_k)_{k\in\mathbb{N}}$ generated by FISTA converges **strongly** *to a minimizer of F. More precisely:*

1. *If $\gamma = 2$ and $\alpha > 3$, previous results are still valid and:*

$$\|x_n - x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\alpha}{3}}\right).$$

2. *If $\gamma > 2$ and $\alpha > 5 + \frac{8}{\gamma-2}$, we get:*

$$F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\gamma}{\gamma-2}}\right), \qquad \|x_n - x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\gamma}{\gamma-2}}\right).$$

In the continuous setting

$$\mathcal{E}(t) = t^2(F(x(t)) - F(x^*)) + \frac{1}{2}\left\|\lambda(x(t) - x^*(t)) + t\dot{x}(t)\right\|^2, \quad \lambda = \frac{2\alpha}{3}.$$

- Requires some additional properties on the set of minimizers.

In the discrete setting for $\gamma = 2$

$$E_n = \frac{2n^2}{L}(F(x_n) - F^*) + \left\|\lambda(x_{n-1} - x^*_{n-1}) + n(x_n - x_{n-1})\right\|^2$$

- No additional properties required on the set of minimizers !

Conclusion about FISTA and inertial methods

- No need to estimate the growth parameter μ and the convergence rate does not suffer from an underestimation of μ.

  J-F Aujol, Ch. Dossal, A.R. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. Mathematical Programming 2023.

- The iterates generated by FISTA strongly converge to a minimizer for the class composite convex functions F satisfying some local/global growth condition.

  Article in preparation with JF Aujol, C Dossal and H Labarriere.

- Inertial methods are more efficient than the gradient descent without the assumption of uniqueness of the minimizer.
- Next step: removing the convexity assumption.