# Fast algorithmic methods for optimization and learning

## Improving "Fast Iterative Shrinkage-Thresholding Algorithm": Faster, Smarter and Greedier

Presenté par **Cleque Marlain MBOULOU**
Enseignants de l'UE:
**Mr. Samir Adly**

January 2024

## Abstract

The paper provides a comprehensive overview of various improvements to the **Fast Iterative Shrinkage-Thresholding Algorithm** (**FISTA**), a theyll-known optimization algorithm. The authors propose modifications to the original **FISTA** scheme, including a "**lazy-start" strategy** and **adaptive and greedy strategies** to enhance practical performance. They also discuss the convergence of the generated sequence and the optimal convergence rate. The paper includes a comparison of the proposed schemes through numerical experiments in different application areas such as inverse problems, machine learning, and signal/image processing. Key contributions include proving sequence convergence, designing a **lazy-start strategy** for faster performance, and proposing novel **adaptive and greedy strategies** to push the algorithm's limits. The paper is structured to present the problem, the proposed modifications, theoretical analysis, and numerical experiments

1. Introduction
   - Forward–Backward-type splitting schemes
   - Problems
   - Contributions
2. A sequence convergent FISTA scheme
   - Two observations & FISTA-Mod
3. Lazy-start strategy
   - Advantage of lazy-start
4. Adaptive acceleration
   - Strong convexity is available
   - Strong convexity is not available
   - Greedy FISTA
5. Nesterov's accelerated scheme
6. Numerical experiments

The acceleration of first-order optimisation methods is an active research topic of non-smooth optimisation. Inertial techniques proposed by *Polyak* ( **"Heavy-ball"** ) and *Nesterov* improve objective function convergence and the performance of gradient descent. Beck and Teboulle have extended to the non-smooth case and proposed **FISTA** scheme which is the main focus of this paper.

$$\min_{x \in \mathcal{H}} \Phi(x) := F(x) + R(x) \tag{1}$$

where $\mathcal{H}$ is a real Hilbert space. The following assumptions are assumed throughout the paper:

- $R : \mathcal{H} \to ]-\infty, +\infty]$ is proper convex and lotheyr semi-continuous (lsc);
- $F : \mathcal{H} \to ]-\infty, +\infty[$ is convex differentiable, with gradient $\nabla F$ being $L$-Lipschitz continuous forsome $L > 0$
- The set of minimizers is non-empty, i.e., $Argmin(\Phi) \neq \emptyset$

# Forward–Backward splitting and Inertial Forward–Backward

**Forward–Backward splitting (FBS)**: With initial point $x_0 \in \mathcal{H}$ chosen arbitrarily and the step size $\lambda_k \in ]0, \frac{2}{L}[$, the standard **FBS** iteration without relaxation reads as:

$$x_{k+1} = \mathtt{prox}_{\lambda_k R}(x_k - \lambda_k \nabla F(x_k))$$

**Inertial Forward–Backward**:Proposed by Moudafi and Oliny, under the setting of finding zeros of monotone inclusion problem

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \mathtt{prox}_{\lambda_k R}(x_k - \lambda_k \nabla F(x_k)) \end{cases}$$

where $a_k$ is the inertial parameter which controls the momentum $x_k - x_{k-1}$

# Original FISTA

**FISTA** belongs to the class of inertial FBS algorithms. What differentiates **FISTA** from others is the restriction on step-size $\lambda_k$ and a special rule for updating $a_k$. The convergence is guaranteed under choices of $\lambda_k$ and $a_k$.

**Input:** $t_0 = 1$, $\gamma = \frac{1}{L}$, $x_0 \in H$, $x_{-1} = x_0$
**Repeat:**

1. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
2. $a_k = \frac{t_{k-1}-1}{t_k}$
3. $y_k = x_k + a_k(x_k - x_{k-1})$
4. $x_{k+1} = \text{prox}_{\gamma R}(y_k)$

**Algorithm 1:** The original FISTA scheme: FISTA-BT

# FISTA-CD

Though in **FISTA-BT** convergence of the objective function is guaranteed, the convergence of $(x_k)_{k\in\mathbb{N}}$ is a open problem. Chambolle and Dossal proved the convergence of $\{x_k\}_{k\in\mathbb{N}}$ by considering the following rule to update $t_k$: let $d > 2$:

$$t_k = \frac{k+d}{d}, a_k = \frac{t_{k-1}-1}{t_k} = \frac{k-1}{k+d} \quad (7)$$

they refer this improvement by **FISTA-CD**

# Problem

The **FISTA-BT** method, although theoretically achieving optimal convergence rates, faces practical challenges, particularly in terms of oscillatory behaviour. Various modifications have been proposed, such as **monotonic FISTA** and **restarted FISTA**, but questions remain about the convergence of the original **FISTA-BT**. In addition, parameter choices ($\lambda_k$ and $\alpha_k$), in particular the value of d in **FISTA-CD**, lack clear theoretical justification. Estimation of strong convexity in practice is challenging, and low-complexity approaches need to be explored. Finally, while the re-launch of FISTA has improved practical performance, research is underway to further improve the system.

**A sequence convergent FISTA scheme:**The above problems are the main motivations of this paper, and our contributions are summarised below. By studying the $t_k$ updating rule (7) of **FISTA-BT** and its difference with 4, they propose a modified FISTA scheme which applies the following rule,

$$t_k = \frac{p + \sqrt{q + r t_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1} - 1}{t_k},$$

where $p, q \in (0, 1]$ and $r \in (0, 4]$.

**Lazy-start strategy:** a "lazy-start" strategy is proposed for practical acceleration in the proposed scheme (**??**) and **FISTA-CD** . This approach, slowing down the convergence speed of $a_k$ approaching 1, proves significantly faster than standard schemes for certain problems.

To overcome persistent oscillatory behavior in **FISTA** despite the **lazy-start strategy**, two additional acceleration techniques are proposed : **restarting adaptation to local strong convexity** and a **greedy scheme**.

- **The adaptive scheme** effectively addresses oscillations related to strong convexity without the need for direct estimation, achieving top-tier performance.
- **The greedy scheme** explores oscillation mechanisms, demonstrating faster practical performance compared to the restarting **FISTA** .

**Nesterov's accelerated schemes:** The previous result is extended to Nesterov's schemes. Such extension can also significantly improve the performance when compared to the original schemes.

# Observations

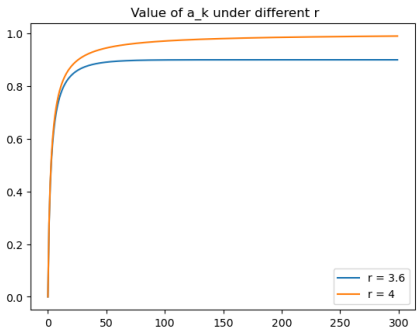
The main problems of the current FISTA schemes are caused by the behaviour of $a_k$, that $a_k$ converges to 1 too fast.In this section, they shall first discuss how to introduce control parameters to FISTA-BT which leads to a modified FISTA scheme, and then present convergence analysis. they will use the rule **??** to modify **FISTA-BT**

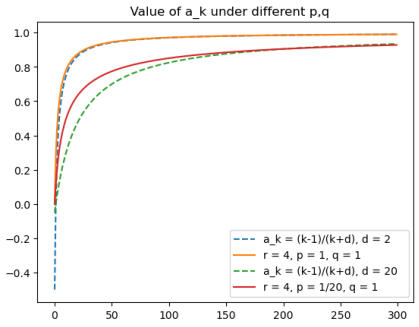
# Two observation

## Observation 1: Replacing 4 with $r > 0$

$$t_k = \frac{1+\sqrt{1+rt_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1}-1}{t_k}$$

$$r \in ]0,4[ : t_k \to \frac{4}{4-r} < +\infty,\ a_k \to \frac{r}{4} < 1,$$
$$r = 4 : t_k \approx \frac{k+1}{2} \to +\infty,\ a_k \to 1,$$
$$r \in ]4,+\infty[ : t_k \propto \left(\frac{\sqrt{r}}{2}\right)^k \to +\infty,\ a_k \to \frac{2}{\sqrt{r}} < 1,$$




## Observation 2: with $p, q > 0$, and restrict $r \in ]0,4]$ :

$$t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1}-1}{t_k}$$

Depending on the choices of $p, q$ and $r$, this time we have

$$r \in ]0,4[ : t_k \to \frac{2p+\Delta}{4-r} < +\infty,\ a_k \to \frac{2p+\Delta-(4-r)}{2p+\Delta} < 1,$$
$$r = 4 : t_k \approx \frac{k+1}{2}p \to +\infty,\ a_k \to 1,$$

where $\Delta \overset{\text{def}}{=} \sqrt{rp^2+(4-r)q}$.

# FISTA-Mod

**Input:** $p, q > 0, r \in ]0, 4], t_0 = 1, \gamma \leq \frac{1}{L}, x_0 \in H, x_{-1} = x_0$
**Repeat:**

1. $t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2}$
2. $a_k = \frac{t_{k-1}-1}{t_k}$
3. $y_k = x_k + a_k(x_k - x_{k-1})$
4. $x_{k+1} = \text{prox}_{\gamma R}(y_k)$

**Algorithm 2:** FISTA-Mod

# Convergence rate of the objective function (Part 1)

In this part, they present the global convergence properties of the FISTA-Mod scheme. they first demonstrate that the optimal convergence rate of FISTA-BT is preserved by FISTA-Mod, and they also demonstrate the convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$.

**Theorem II.1 (Convergence rate of the objective)** For the FISTA-Mod scheme (6), let $r = 4$ and choose $p \in ]0, 1]$, $q \in ]0, (2-p)^2]$. Then

$$\Phi(x_k) - \Phi(x^*) \leq \frac{2L}{p^2(k+1)^2}||x_0 - x^*||^2.$$

**Remark II.1.** Compared to the original convergence rate of FISTA-BT , which is $\Phi(x_k) - \Phi(x^*) \leq \frac{2L}{(k+1)^2}||x_0 - x^*||^2$, parameter $p$ appears in the obtained rate estimation, and $p = 1$ yields the smallest constant in the rate. Although $p < 1$ gives a bigger constant in the rate estimation, as they shall see below, it allows us to prove the $o(1/k^2)$ convergence rate.

# Convergence rate of the objective function (Part 2)

**Theorem II.2 (From $O(1/k^2)$ to $o(1/k^2)$)** For the FISTA-Mod scheme , let $r = 4$ and choose $p \in (0, 1]$, $q > 0$ such that

$$q \leq \frac{(2-p)^2}{2}$$

then there holds

$$\Phi(x_k) - \Phi(x^*) \leq \frac{2L}{p^2(k+1)^2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

If $p^2 \leq q$. Then $\Phi(x_k) - \Phi(x^*) = o(1/k^2)$.

**Remark II.2.** (i) For the original FISTA-BT scheme, they have strictly $0 = t_{k-1} - (t_k^2 - tk)$, hence unable to obtain $o(1/k^2)$ convergence rate. (ii) One byproduct of Theorem II.5 is that one can show that the sequence $\{x_k\}_{k\in\mathbb{N}}$ is bounded.

# Convergence rate of the objective function (Part 3)

**Theorem II.4 (Convergence of the sequence).** For the FISTA-Mod scheme, let $r = 4$ and choose $p \in ]0, 1[$, $q > 0$ such that $p^2 \leq q$. Then

1. There exists an $x^* \in \text{Argmin}(\Phi)$ to which the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by FISTA-Mod converges theyakly;
2. they have $\|x_k - x_{k-1}\| = o(1/k)$.

In FISTA-Mod, parameters $p$, $q$, and $r$ influence the convergence of $\{x_k\}$. The algorithm achieves $O(1/k^2)$ convergence for $\Phi(x_k) - \min_{x\in H} \Phi(x)$ and $O(1/k)$ convergence for $\|x_k - x_{k-1}\|$.

In this section, they further show that such degree of freedom provided by parameters $p$, $q$, and $r$ allows us to design a so-called "*lazy-start strategy*" which can make **FISTA-Mod/FISTA-CD** much faster in practice.

**Proposition I (Lazy-start FISTA).** For FISTA-Mod and FISTA-CD, consider the following choices of $p$, $q$, and $d$ respectively:

- FISTA-Mod: $p \in \left[\frac{1}{80}, \frac{1}{10}\right]$, $q \in [0, 1]$, and $r = 4$;
- FISTA-CD: $d \in [10, 80]$.

This is a **"lazy-start"** strategy because it slows down the speed of $a_k$ converging to 1.
To discuss the advantage of lazy-start, they consider the simple least square problem below:

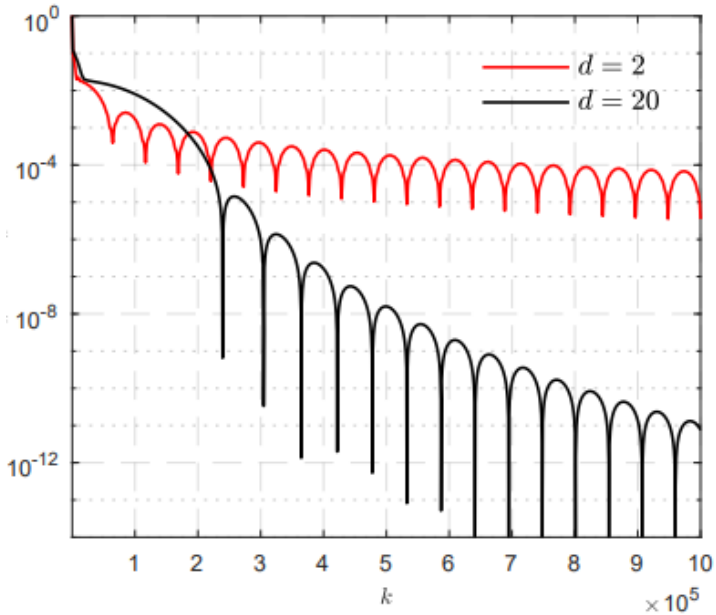
$$\min_{x \in \mathbb{R}^{201}} F(x) := \frac{1}{2}||Ax||_2^2$$

Where $A \in \mathbb{R}^{201\times 201}$ has the following form:

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}_{201\times 201}.$$

This strongly convex function admits a unique minimizer $x^* = 0$

$$y_k = x_k + \frac{k-1}{k+d}(x_k - x_{k-1})$$
$$x_{k+1} = y_k - \frac{1}{L}A^T A y_k = (\mathrm{Id} - \frac{1}{L}A^T A)y_k.$$



According to the theyll-known behavior of FISTA schemes, once $a_k$ is too close to 1, oscillations occur in the trajectories of both $\|x_k - x_{k-1}\|_{k\in\mathbb{N}}$ and $\Phi(x_k) - \Phi(x^*)_{k\in\mathbb{N}}$ when the minimization of strongly convex problem. This oscillation phenomenon decelerates the algorithm's speed, making it slotheyr than the original *Forward–Backward splitting (FBS)* scheme 5. An intuitive explanation for this behavior lies in the interplay of two properties:

1. The considered problems in 1 are locally strongly convex at the minimizer.
2. A relatively larger value of $d$ slows down the convergence of $a_k$ to 1, as illustrated in the previous figure.

The combination of these characteristics makes the algorithm operate more quickly in real-world scenarios.

# Strong convexity is available

Despite the advantages of the lazy-start strategy, this section focuses on resolving the oscillation problem seen in FISTA-BT and FISTA-Mod/FISTA-CD. The adaptive approaches discussed here aim to mitigate this oscillation, particularly in cases where strong convexity information is available or unknown.

Assuming $F$ in problem 1 is $\alpha$-strongly convex and $R$ is only convex, the optimal choices for $p$, $q$, and $r$ are determined. The inertial parameter $a$ is calculated as $a^* = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$. The optimal $r$ is found through solving an equation, resulting in the following expression:

$$\frac{2p+\Delta-(4-r)}{2p+\Delta} = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}},$$

where $\Delta \overset{\text{def}}{=} \sqrt{rp^2+(4-r)q}$. Solve the above equation we get the optimal choice of $r$ which reads

$$\begin{aligned} r = f(\alpha,\gamma;p,q) &= 4(1-p)+4pa^\star+(p^2-q)(1-a^\star)^2 \\ &= 4(1-p)+\frac{4p(1-\sqrt{\gamma\alpha})}{1+\sqrt{\gamma\alpha}}+\frac{4\gamma\alpha(p^2-q)}{(1+\sqrt{\gamma\alpha})^2} \le 4. \end{aligned}$$

Note that we have $f(\alpha,\gamma;p,q) = 4$ for $\alpha = 0$, and $f(\alpha,\gamma;p,q) < 4$ for $\alpha > 0$.

This formulation provides a generalization of FISTA-Mod called $\alpha$-FISTA

**Algorithm 3:** Strongly convex FISTA-Mod ($\alpha$-**FISTA**)

**Initial**: let $p, q > 0$ and $\gamma \leq 1/L$. For $\alpha \geq 0$, determine $r$ as $r = f(\alpha, \gamma; p, q)$. Let $t_0 \geq 1$, and $x_0 \in \mathbb{R}^n, x_{-1} = x_0$.

**repeat**

$$t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2}, \; a_k = \frac{t_{k-1}-1}{t_k},$$
$$y_k = x_k + a_k(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$$

**until** *convergence*;

In the case where $R$ is $\alpha_R$-strongly convex and $F$ is $\alpha_F$-strongly convex, the problem is $(\alpha = \alpha_R + \alpha_F)$-strongly convex, and a generalization of the FISTA scheme for strongly convex problems is proposed: In Algorithm 4, an update rule for $t_k$ and $a_k$ is provided, given by:

$$t_k = \frac{1-qt_{k-1}^2+\sqrt{(1-qt_{k-1}^2)^2+4t_{k-1}^2}}{2} \quad \text{and} \quad a_k = \frac{t_{k-1}-1}{t_k}\frac{1+\gamma\alpha_R-t_k\gamma\alpha}{1-\gamma\alpha_F},$$

# Strong convexity is not available

The strong convexity adaptive $\alpha$-FISTA previously presented performs theyll when the modulus of strong convexity is known; hotheyver, it is acknowledged that strong convexity is often unknown or absent in practice. To address this problem, the algorithm that follows recommends restarting the adaptive scheme.

**Algorithm 4:** Restarting and Adaptive $\alpha$-FISTA (**Rada-FISTA**)

**Initial**: $p,q \in ]0,1], r = 4$ and $\xi < 1, t_0 = 1, \gamma = 1/L$ and $x_0 \in \mathscr{H}, x_{-1} = x_0$.

**repeat**

- Run FISTA-Mod:

$$t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2},\ a_k = \frac{t_{k-1}-1}{t_k},$$
$$y_k = x_k + a_k(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$$

- Restarting: if $(y_k - x_{k+1})^T (x_{k+1} - x_k) \geq 0$,
  - Option I: $r = \xi r$ and $y_k = x_k$;
  - Option II: $r = \xi r, t_k = 1$ and $y_k = x_k$.

**until** *convergence*;

This scheme offers flexibility for scenarios where the modulus $\alpha$ is unknown or where the objective function $\phi$ might not have strong convexity locally or globally. It combines $\alpha$-FISTA with the restarting method from previous research. The method aims to obtain both monotonic convergence and efficient estimation of strong convexity when applicable.

In comparison to Rada-FISTA and restarting FISTA, the section looks at ways to improve the restarting technique's performance and achieve faster results. It identifies the oscillation issue in FISTA schemes due to $a_k$ approaching 1. As a result, the authors suggest Greedy restarting plan below.

**Algorithm 5:** Greedy FISTA

**Initial**: let $\gamma \in [\frac{1}{L}, \frac{2}{L}[$ and $\xi < 1, S > 1$, choose $x_0 \in \mathbb{R}^n, x_{-1} = x_0$.

**repeat**

- Run the iteration:

$$y_k = x_k + (x_k - x_{k-1}),$$
$$x_{k+1} = \text{prox}_{\gamma R}(y_k - \gamma \nabla F(y_k)).$$

- Restarting: if $(y_k - x_{k+1})^T (x_{k+1} - x_k) \geq 0$, then $y_k = x_k$;
- Safeguard: if $\|x_{k+1} - x_k\| \geq S\|x_1 - x_0\|$, then $\gamma = \max\{\xi\gamma, \frac{1}{L}\}$;

**until** *convergence*;

Compared to *Rada-FISTA* and *restarting FISTA*, they find that in practice, $\lambda \in [1/L, 1.3/L]$ offers faster performance.

Several acceleration schemes are introduced by Nesterov; in the sections that follow, they Give the "Constant Step Scheme, III" your full attention. Our algorithm for resolving 1 is as follows:

**Algorithm 6:** Accelerated proximal gradient (APG)

**Initial**: $\tau \in [0,1], \theta_0 = 1, \gamma = 1/L$ and $x_0 \in \mathscr{H}, x_{-1} = x_0$.

**repeat**

Estimate the local strong convexity $\alpha_k$;

$$\theta_k \text{ solves } \theta_k^2 = (1-\theta_k)\theta_{k-1}^2 + \tau\theta_k,$$

$$a_k = \frac{\theta_{k-1}(1-\theta_{k-1})}{\theta_{k-1}^2+\theta_k},$$

$$y_k = x_k + a_k(x_k - x_{k-1}),$$

$$x_{k+1} = \text{prox}_{\gamma R}\big(y_k - \gamma\nabla F(y_k)\big).$$

**until** *convergence*;

When the problem 1 is $\alpha-$strongly convex, then by setting $\tau = \sqrt{\alpha/L} = p$ and $\theta_0 = \tau$ they have:

$$\theta_k \equiv \tau \quad \text{and} \quad a_k \equiv \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}},$$

The definition of $a_k$ implies $\theta_k \in [0,1]$ for all $k \geq 1$. Therefore, the $\theta_k$ in APG is solve by:

$$\theta_k = \frac{-(\sigma\theta_{k-1}^2 - \tau) + \sqrt{(\sigma\theta_{k-1}^2-\tau)^2 + 4\theta_{k-1}^2}}{2}.$$

# A modified APG

**Algorithm 7:** A modified APG scheme(**mAPG**)

**Initial**: Let $\sigma \in [0,1], \gamma = 1/L$ and $\tau = \gamma\alpha\sigma, \theta_0 \in [0,1]$. Set $x_0 \in \mathscr{H}, x_{-1} = x_0$.

**repeat**

$$\theta_k \text{ solves } \theta_k^2 = (1-\sigma\theta_k)\theta_{k-1}^2 + \tau\theta_k,$$

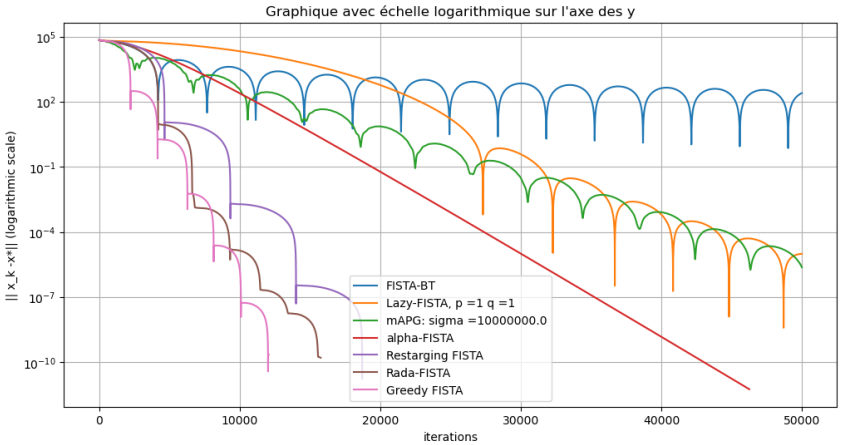
$$a_k = \frac{\theta_{k-1}(1-\theta_{k-1})}{\theta_{k-1}^2+\theta_k},$$

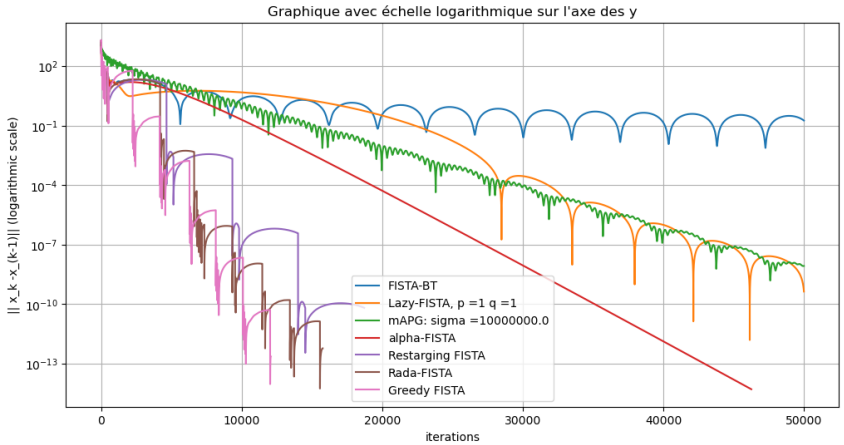
$$y_k = x_k + a_k(x_k - x_{k-1}),$$

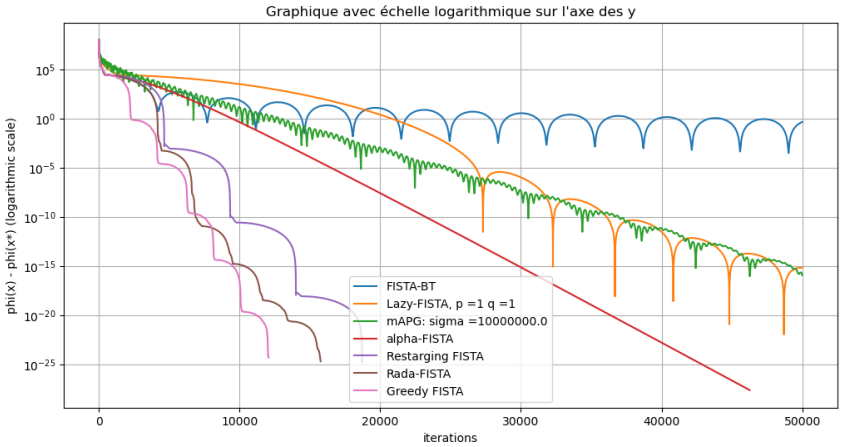
$$x_{k+1} = \text{prox}_{\gamma R}\big(y_k - \gamma\nabla F(y_k)\big).$$

**until** *convergence*;

Now, we present numerical experiments on problem




## Source:

1. Jingwei Liang , Tao Luo & Carola-Bibiane Schönlieb, Improving "Fast Iterative Shrinkage-Thresholding Algorithm": Faster, Smarter and Greedier , 4 Nov 2018

# Fast algorithmic methods for optimization and learning

## Improving "Fast Iterative Shrinkage-Thresholding Algorithm": Faster, Smarter and Greedier

Presenté par **Cleque Marlain MBOULOU**
Enseignants de l'UE:
**Mr. Samir Adly**

January 2024