Improving FISTA: Faster, Smarter and Greedier

Jingwei Liang[*] Carola-Bibiane Schönlieb[†]

Abstract. The "fast iterative shrinkage-thresholding algorithm", a.k.a. FISTA, is one of the most well-known first-order optimisation scheme in the literature, as it achieves the worst-case $O(1/k^2)$ optimal convergence rate in terms of objective function value. However, despite the optimal theoretical rate, in practice the (local) oscillatory behaviour of FISTA often damps its efficiency. Over the past years, various efforts are made in the literature to improve the practical performance of FISTA, such as monotone FISTA, restarting FISTA and backtracking strategies. In this paper, we propose a simple yet effective modification to FISTA which has two advantages: it allows us to 1) prove the convergence of generated sequence; 2) design a so-called "lazy-start" strategy which can up to an order faster than the original scheme. Moreover, by exploring the properties of FISTA scheme, we propose novel adaptive and greedy strategies which probes the limit of the algorithm. The advantages of the proposed schemes are tested through problems arising from inverse problem, machine learning and signal/image processing.

Key words. FISTA, inertial Forward–Backward, lazy-start strategy, adaptive and greedy acceleration

AMS subject classifications. 65K05, 65K10, 90C25, 90C31.

1 Introduction

The acceleration of first-order optimisation methods is an active research topic of non-smooth optimisation. Over the past decades, various acceleration techniques are proposed in the literature. Among them, one most widely used is the "inertia technique" which owns to [27] by Polyak where he proposed the so called heavy-ball method which dramatically speed-up the performance of gradient descent. Under a similar spirit, in [22] Nesterov proposed another accelerated gradient scheme which improves the $O(1/k)$ objective function convergence rate of gradient descent to $O(1/k^2)$. The extension to non-smooth optimisation was due to Beck and Teboulle, who proposed the FISTA scheme [6] which is the main focus of this paper. A different approach, in terms of parameter update, but with same convergence rate can be also found in [24].

In this paper, we are interested in the following structured non-smooth optimization problem, which is the sum of two convex functionals,

$$\min_{x\in\mathcal{H}} \ \Phi(x) \overset{\text{def}}{=} F(x) + R(x), \tag{$\mathcal{P}$}$$

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

(**H.1**) $R : \mathcal{H} \to]-\infty, +\infty]$ is proper convex and lower semi-continuous (lsc);

(**H.2**) $F : \mathcal{H} \to]-\infty, +\infty[$ is convex differentiable, with gradient ∇F being L-Lipschitz continuous for some $L > 0$;

(**H.3**) The set of minimizers is non-empty, *i.e.* $\mathrm{Argmin}(\Phi) \neq \emptyset$.

Problem ($\mathcal{P}$) covers many interesting problems arising from inverse problems, signal/image processing, computer vision and machine learning, to name few. We refer to Section 7 the numerical experiment section for more concrete examples.

[*] DAMTP, University of Cambridge, Cambridge, UK. E-mail: jl993@cam.ac.uk.

[†] DAMTP, University of Cambridge, Cambridge, UK. E-mail: cbs31@cam.ac.uk.

1.1 Forward–Backward-type splitting schemes

In the literature, one widely used approach for solving ($\mathcal{P}$) is the Forward–Backward splitting (FBS) method [17]. Over the past decades, numerous variants of FBS are proposed under different purpose. Below, we present a brief overview of Forward–Backward-type schemes and mainly focus on the ones using inertia technique.

1.1.1 Forward–Backward splitting and inertial schemes

Forward–Backward splitting With initial point $x_0 \in \mathcal{H}$ chosen arbitrarily, the standard FBS iteration without relaxation reads as

$$x_{k+1} \overset{\text{def}}{=} \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big),\ \gamma_k \in]0, 2/L], \tag{1.1}$$

where γ_k is the step-size, and $\mathrm{prox}_{\gamma R}$ is called the *proximity operator* of R which is defined by

$$\mathrm{prox}_{\gamma R}(\cdot) \overset{\text{def}}{=} \min_{x\in\mathcal{H}} \gamma R(x) + \tfrac{1}{2}\|x - \cdot\|^2. \tag{1.2}$$

FBS recovers gradient descent when $R = 0$ and the proximal point algorithm [29] when $F = 0$.

Similar to gradient descent, FBS is a descent method, *i.e.* objective function value is non-increasing under proper step-size. The convergence properties of FBS are well established in the literature:

- The convergence of the generated sequence $\{x_k\}_{k\in\mathbb{N}}$ and the objective function value $\Phi(x_k)$ are guaranteed [12] as long as γ_k is chosen such that $0 < \underline{\gamma} \le \gamma_k \le \bar{\gamma} < \frac{2}{L}$;
- Convergence rate: we have $\Phi(x_k) - \min_{x\in\mathcal{H}} \Phi(x) = o(1/k)$ for the objective function value [19] and $\|x_k - x_{k-1}\| = o(1/\sqrt{k})$ for the sequence $\{x_k\}_{k\in\mathbb{N}}$ [15]. Moreover, linear convergence rate can be obtained for instance under strong convexity.

Inertial Forward–Backward The first inertial Forward–Backward was proposed by Moudafi and Oliny in [20], under the setting of finding the zeros of monotone inclusion problem. Specify the algorithm to solve ($\mathcal{P}$), one obtains the following iteration: let $\gamma_k \in]0, 2/L[$ and

$$\begin{aligned} y_k &= x_k + a_k(x_k - x_{k-1}), \\ x_{k+1} &= \mathrm{prox}_{\gamma_k R}\big(y_k - \gamma_k \nabla F(x_k)\big), \end{aligned} \tag{1.3}$$

where a_k is the *inertial parameter* which controls the momentum $x_k - x_{k-1}$. The above scheme recovers the heavy-ball method when $R = 0$, and becomes the scheme proposed in [18] if we replace $\nabla F(x_k)$ with $\nabla F(y_k)$. We refer to [16] for a more general discussion of inertial Forward–Backward splitting schemes.

The convergence of (1.3) can be guaranteed under proper choices of γ_k and a_k. Under the same step-size choice, (1.3) could be significantly faster than FBS in practice. However, except for special cases (*e.g.* quadratic problem as in [28]), in general there is no convergence rate established for (1.3).

1.1.2 The FISTA schemes

By form, FISTA belongs to the class of inertial FBS schemes. What differentiates FISTA from the others is the restriction on step-size γ_k and special rule for updating a_k. Moreover, FISTA schemes have convergence rate guarantee on the objective function value, which is the consequence of the updating rule of a_k.

The original FISTA The FISTA scheme is first proposed in [6], which is described below in Algorithm 1.

Algorithm 1: The original FISTA scheme

```
Initial: $t_0 = 1, \gamma = 1/L$ and $x_0 \in \mathcal{H}, x_{-1} = x_0$.
repeat
    $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2},\ a_k = \frac{t_{k-1}-1}{t_k},$
    $y_k = x_k + a_k(x_k - x_{k-1}),$                                  (1.4)
    $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$
    $k = k + 1;$
until convergence;
```

As we observe, FISTA first computes t_k and then updates a_k using t_k and t_{k-1}. Due to such way of parameter choice, FISTA achieves $O(1/k^2)$ convergence rate for $\Phi(x_k) - \min_{x\in\mathcal{H}} \Phi(x)$ which is optimal [21]. For the rest of the paper, to distinguish the original FISTA from the one in [10] and the proposed modified FISTA scheme, we use "FISTA-BT" to refer Algorithm 1.

Sequence convergent FISTA Though achieving optimal convergence rate for objective function value, the convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 has been an open problem. This question was answered in [10], where Chambolle and Dossal proved the convergence of $\{x_k\}_{k\in\mathbb{N}}$ by considering the following rule to update t_k:

$$t_k = \frac{k+d}{d}, \quad a_k = \frac{t_{k-1}-1}{t_k} = \frac{k-1}{k+d}. \tag{1.5}$$

Such a rule maintains the $O(1/k^2)$ objective convergence rate, moreover allows the authors proving the convergence of $\{x_k\}_{k\in\mathbb{N}}$. Later on in [3], (1.5) was studied under the continuous time dynamical system setting, and the convergence rate of objective function was proved to be $o(1/k^2)$ [2]. For the rest of the paper, we shall use "FISTA-CD" to refer (1.5).

1.2 Problems

It has been almost a decade since FISTA-BT was proposed, various variants are proposed in the literature, for instance the monotone FISTA [5] and restarting FISTA [25] which aim to solve the oscillatory behaviour of FISTA schemes, the FISTA-CD [10] for the convergence of iterates, and the backtracking strategy for adapting Lipschitz constant [8]. However, there are still important questions to deal with:

- Though [10] proves the convergence of the FISTA-CD scheme under (1.5), the convergence of the original FISTA-BT still is unclear;
- The performances of FISTA-CD is almost identical to FISTA-BT if d in (1.5) is chosen close to 2. However, when relatively large value of d is chosen, then significant practical acceleration can be obtained. For instance, it is reported in [16] that for $d = 50$, the performance can be several times faster than $d = 2$. However, there is no proper guidelines on how to choose the value of d in practice.
- When the problem ($\mathcal{P}$) is strongly convex, there exists an optimal choice for a_k [23]. However, in practice, very often the problem is only locally strongly convex with strong convexity unknown. Knowing strong convexity allows us to apply the optimal scheme [23], however estimating it can be time consuming. Moreover, is optimal scheme the fastest in practice, or do we really need the strong convexity?
- The restarting FISTA successfully suppresses the oscillatory behaviour of FISTA schemes, hence achieving much faster practical performance. Then, can we further improve this scheme?

1.3 Contributions

The above problems are the main motivations of this paper, and our contributions are summarised below.

A sequence convergent FISTA scheme By studying the t_k updating rule (1.4) of FISTA-BT and its difference with (1.5), we propose a modified FISTA scheme which applies the following rule,

$$t_k = \frac{p + \sqrt{q + rt_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1}-1}{t_k}, \tag{1.6}$$

where $p, q \in]0, 1]$ and $r \in]0, 4]$, see also Algorithm 2. Such modification has two advantages

- It maintains the $O(1/k^2)$ (actually $o(1/k^2)$) convergence rate of the original FISTA-BT (Theorem 3.3);
- It allows to prove the convergence of $\{x_k\}_{k\in\mathbb{N}}$ (Theorem 3.5);

It also allows us to show that the original FISTA-BT is also optimal in terms of the constant appears in the $O(1/k^2)$ rate, see Eq. (3.7) of Theorem 3.3.

Lazy-start strategy For the proposed scheme and FISTA-CD, owing to the free parameters, we propose in Section 4 a so-called "lazy-start" strategy for practical acceleration. The idea of such strategy is to slow down the speed of a_k approaching 1, which can lead to a much faster practical performance. For certain problems, such strategy can be an order faster than the original schemes, see Section 7 for illustration. Moreover, we provide simple practical guidelines on how to choose these parameters in practice.

Adaptive and greedy acceleration Though lazy-start strategy can significantly speed up the performance of FISTA, it still suffers the oscillatory behaviour of FISTA schemes since the inertial parameter a_k eventually converges to 1. By combining with the restarting technique of [25], in Section 5 we propose two different acceleration strategies: restarting adaptation to (local) strong convexity and greedy scheme.

The oscillatory behaviour of FISTA schemes is often related to the strong convexity of the problem. When the strong convexity α is non-zero, there exists an optimal choice [23], *i.e.* $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$, where γ is the step-size and α is the strong convexity. Moreover, under such choice the iteration will no longer oscillate. As many problems in practice are only locally strongly convex, plus the fact that the estimating the strong convexity in general can be quite time consuming. In Section 5, we propose an adaptive scheme (Algorithm 4) which combines the restarting technique and the modified parameter update rule (1.6). Such adaptive scheme avoids the direct estimation of strong convexity and achieve state-of-the-art performance.

Though closely related, strongly convexity is only a sufficient condition for the oscillatory behaviour of FISTA schemes. We investigate the mechanism of oscillation and the restarting technique, and propose a greedy scheme (Algorithm 5) which probes the limit of the oscillation and restarting technique. By doing so, the greedy scheme can achieve a faster practical than the restarting FISTA of [25].

Nesterov's accelerated schemes Given the close relation between FISTA and the Nesterov's accelerated schemes [23], we also extend the above result, particularly the modified FISTA scheme to Nesterov's schemes. Such extension can also significantly improve the performance when compared to the original schemes.

1.4 Paper organisation

The rest of the paper is organised as following. Some notations and preliminary result are collected in Section 2. The proposed sequence convergent FISTA scheme is presented in Section 3. The lazy-start strategy and the adaptive/greedy acceleration schemes are presented in Section 4 and Section 5 respectively. In Section 6, we extend the result to Nesterov's accelerated schemes. Numerical experiments are presented in Section 7.

2 Preliminaries

Throughout the paper, $\mathcal{H}$ is a Hilbert space equipped with scalar product $\langle\cdot,\,\cdot\rangle$ and norm $\|\cdot\|$. Id denotes the identity operator on $\mathcal{H}$. $\mathbb{N}$ is the set of non-negative integers and $k \in \mathbb{N}$ is the index. $x^\star \in \mathrm{Argmin}(\Phi)$ denotes a global minimiser of $(\mathcal{P})$.

The sub-differential of a function $R \in \Gamma_0(\mathcal{H})$ is the set-valued mapping defined by

$$\partial R : \mathcal{H} \rightrightarrows \mathcal{H},\, x \mapsto \big\{ g \in \mathcal{H} \,|\, R(x') \geq R(x) + \langle g,\, x'-x\rangle,\ \forall x' \in \mathcal{H} \big\}. \tag{2.1}$$

Definition 2.1 (Monotone operator). A set-valued mapping $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is said to be monotone if,

$$\langle x_1 - x_2,\, v_1 - v_2\rangle \geq 0,\ \forall (x_1, v_1) \in \mathrm{gph}\,(A) \quad \text{and} \quad (x_2, v_2) \in \mathrm{gph}\,(A). \tag{2.2}$$

It is maximal monotone if $\mathrm{gph}\,(A)$ can not be contained in the graph of any other monotone operators.

It is well-known that for $R \in \Gamma_0(\mathcal{H})$, ∂R is maximal monotone [30], and that $\mathrm{prox}_R = (\mathrm{Id} + \partial R)^{-1}$.

Definition 2.2 (Cocoercive operator). Let $\beta \in]0, +\infty[$, $B : \mathcal{H} \to \mathcal{H}$, then B is β-cocoercive if

$$\langle B(x_1) - B(x_2),\, x_1 - x_2\rangle \geq \beta \|B(x_1) - B(x_2)\|^2,\ \forall x_1, x_2 \in \mathcal{H}. \tag{2.3}$$

The L-Lipschitz continuous gradient ∇F of function $F \in C^{1,1}(\mathcal{H})$ is $\frac{1}{L}$-cocoercive [4].

Lemma 2.3 (Descent lemma [7]). *Suppose that $F : \mathcal{H} \to \mathbb{R}$ is convex continuously differentiable and ∇F is L-Lipschitz continuous. Then, given any $x, y \in \mathcal{H}$,*

$$F(x) \leq F(y) + \langle \nabla F(y),\, x - y\rangle + \frac{L}{2}\|x - y\|^2.$$

Given any $x, y \in \mathcal{H}$, define $E_\gamma(x, y)$ which contains the majorization of F,

$$E_\gamma(x, y) \overset{\text{def}}{=} R(x) + F(y) + \langle x - y,\, \nabla F(y)\rangle + \tfrac{1}{2\gamma}\|x - y\|^2.$$

It is obvious that $E_\gamma(x, y)$ is strongly convex with respect to x, hence denote the unique minimiser as

$$\begin{aligned} e_\gamma(y) &\overset{\text{def}}{=} \operatorname{argmin}\big\{E_\gamma(x, y) : x \in \mathbb{R}^n\big\} \\ &= \operatorname{argmin}_x\big\{\gamma R(x) + \tfrac{1}{2}\|x - (y - \gamma\nabla F(y))\|^2\big\} \\ &= \operatorname{prox}_{\gamma R}\big(y - \gamma\nabla F(y)\big). \end{aligned} \tag{2.4}$$

We have the following two basic lemmas from [6].

Lemma 2.4 (Optimality condition of $e_\gamma(y)$). *Given $y \in \mathcal{H}$, let $y^+ = e_\gamma(y)$, then*

$$0 \in \gamma\partial R(y^+) + \big(y^+ - (y - \gamma\nabla F(y))\big) = \gamma\partial R(y^+) + (y^+ - y) + \gamma\nabla F(y).$$

Lemma 2.5 ([6, Lemma 2.3]). *Let $y \in \mathcal{H}$ and $\gamma \in]0, 2/L[$ such that*

$$\Phi(e_\gamma(y)) \le E_\gamma(e_\gamma(y), y),$$

then for any $x \in \mathcal{H}$, we have

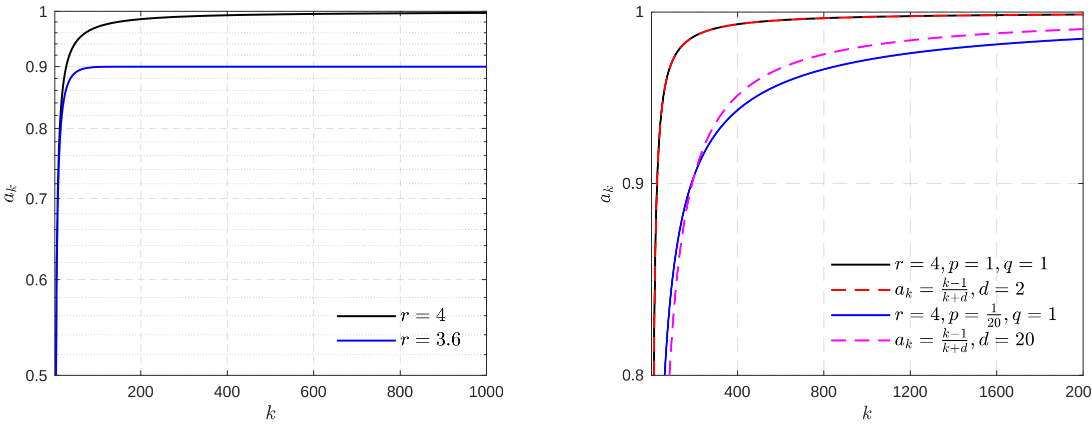
$$\Phi(x) - \Phi(e_\gamma(y)) \ge \tfrac{1}{2\gamma}\|e_\gamma(y) - y\|^2 + \tfrac{1}{\gamma}\langle y - x,\, e_\gamma(y) - y\rangle.$$

Lemma 2.6 ([10, Lemma 3.1]). *Given $y \in \mathcal{H}$ and $\gamma \in]0, 1/L]$, let $y^+ = e_\gamma(y)$, then for any $x \in \mathcal{H}$, we have*

$$\Phi(y^+) + \tfrac{1}{2\gamma}\|y^+ - x\|^2 \le \Phi(y) + \tfrac{1}{2\gamma}\|y - x\|^2.$$

3 A sequence convergent FISTA scheme

We first discuss two observations obtained from the t_k update rule in FISTA-BT which lead to a modified FISTA scheme, then present convergence analysis.



(a) Value of a_k under different r

(b) Value of a_k under different p, q

Figure 1: Different effects of p, q and d. (a) r controls the limiting value of a_k; (b) p, q control the speed of a_k approaching its limit.

3.1 Two observations & FISTA-Mod

Recall the t_k update rule in the original FISTA-BT [6], that reads

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},\ \ a_k = \frac{t_{k-1} - 1}{t_k}.$$

For the following discussions, we replace the constants $1, 1$ and 4 in the update of t_k with three parameters p, q and r and study how will they affect the behaviour of t_k and consequently of a_k.

3.1.1 Observation I

Consider first replacing 4 with a non-negative r, we get

$$t_k = \frac{1+\sqrt{1+rt_{k-1}^2}}{2}, \ a_k = \frac{t_{k-1}-1}{t_k}. \tag{3.1}$$

With simple calculation, we obtain that depending on the value of r, the limiting value of t_k consists of three different cases:

$$\begin{aligned} r \in]0,4[: t_k &\to \frac{4}{4-r} < +\infty, \ a_k \to \frac{r}{4} < 1, \\ r = 4 : t_k &\approx \frac{k+1}{2} \to +\infty, \ a_k \to 1, \\ r \in]4,+\infty[: t_k &\propto \Big(\frac{\sqrt{r}}{2}\Big)^k \to +\infty, \ a_k \to \frac{2}{\sqrt{r}} < 1. \end{aligned} \tag{3.2}$$

Eq. (3.2) implies that r controls $\lim_{k\to+\infty} t_k$, hence $\lim_{k\to+\infty} a_k$. In Figure 1 (a), we show graphically two choices of r: $r = 4$ and $r = 3.6$. It can be observed that, a_k indeed converges to two different values.

3.1.2 Observation II

Now further replace the two 1's in (3.1) with $p, q > 0$, and restrict $r \in]0,4]$:

$$t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2}, \ a_k = \frac{t_{k-1}-1}{t_k}. \tag{3.3}$$

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

$$\begin{aligned} r \in]0,4[: t_k &\to \frac{2p+\Delta}{4-r} < +\infty, \ a_k \to \frac{2p+\Delta-(4-r)}{2p+\Delta} < 1, \\ r = 4 : t_k &\approx \frac{k+1}{2}p \to +\infty, \ a_k \to 1, \end{aligned} \tag{3.4}$$

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

Eq. (3.4) is quite similar to (3.2), in the sense that a_k converges to 1 for $r = 4$ and to some value smaller than 1 when $r < 4$. Moreover, for $r = 4$, the growth of t_k is controlled by p, indicating that we can control the speed of a_k approaching 1 via p, which is illustrated graphically in Figure 1 (b). Under $r = 4$, two different choices of p, q are considered, $(p, q) = (1, 1)$ and $(p, q) = (\frac{1}{20}, 1)$. Clearly, a_k approaches 1 much slower for the second choice of p, q. In comparison, we also add a case for (1.5) of FISTA-CD, for which larger value of d leads to slower speed of a_k converging to 1.

Remark 3.1. Let $r < 4$, and denote $\tilde{t} \overset{\text{def}}{=} \frac{2p+\Delta}{4-r}, \tilde{a} = \frac{2p+\Delta-(4-r)}{2p+\Delta}$ the limiting value of t_k, a_k, respectively. Depending on the initial value of t_0, we have

$$\begin{cases} t_0 < \tilde{t} : t_k \nearrow \tilde{t}, \ a_k \nearrow \tilde{a}; \\ t_0 = \tilde{t} : t_k \equiv \tilde{t}, \ a_k \equiv \tilde{a}; \\ t_0 > \tilde{t} : t_k \searrow \tilde{t}, \ a_k \searrow \tilde{a}. \end{cases}$$

3.1.3 The modified FISTA scheme

Based on the above two observations of t_k, we propose a modified FISTA scheme (Algorithm 2), which we call "FISTA-Mod" for short.

Algorithm 2: A modified FISTA scheme

```
Initial: $p, q > 0$ and $r \in ]0,4]$, $t_0 = 1$, $\gamma \le 1/L$ and $x_0 \in \mathbb{R}^n, x_{-1} = x_0$.
repeat
    $t_k = \frac{p+\sqrt{q+rt_{k-1}^2}}{2}, \ a_k = \frac{t_{k-1}-1}{t_k},$
    $y_k = x_k + a_k(x_k - x_{k-1}),$                                    (3.5)
    $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma\nabla F(y_k)\big).$
until convergence;
```

Remark 3.2. When r is strictly smaller than 4, *i.e.* $r < 4$, then Algorithm 2 is simply a variant of the inertial Forward–Backward, and we refer to [16] for more details on its convergence properties.

3.2 Convergence properties of FISTA-Mod

The parameters p, q and r in FISTA-Mod allow us to control the behaviour of t_k and a_k, hence providing possibilities to prove the convergence of the iterates $\{x_k\}_{k\in\mathbb{N}}$. In this part, we provide two convergence results for FISTA-Mod: $o(1/k^2)$ convergence rate for $\Phi(x_k) - \Phi(x^\star)$ and convergence of $\{x_k\}_{k\in\mathbb{N}}$ together with $o(1/k)$ rate for $\|x_k - x_{k-1}\|$. The proofs of these results are inspired by the work of [10, 2], for the sake of self-consistent we present the details of the proofs.

3.2.1 Main result

We present below first the main convergence result, and then provide the corresponding proofs.

Theorem 3.3 (Convergence of objective). *For the FISTA-Mod scheme* (3.5)*, let* $r = 4$ *and choose* $p \in]0,1], q > 0$ *such that*

$$q \le (2-p)^2, \tag{3.6}$$

then there holds

$$\Phi(x_k) - \Phi(x^\star) \le \frac{2L}{p^2(k+1)^2}\|x_0 - x^\star\|^2. \tag{3.7}$$

If moreover $p \in]0,1[$ *and* $q \in [p^2, (2-p)^2]$*, then* $\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$*.*

Remark 3.4. The $O(1/k^2)$ convergence rate (3.7) recovers the result of FISTA-BT [6] for $p = 1$. Since p appears in the denominator, this indicates that FISTA-BT has the *smallest* constant in the $O(1/k^2)$ rate.

Theorem 3.5 (Convergence of sequence). *For the FISTA-Mod scheme* (3.5)*, let* $r = 4, p \in]0,1[$ *and* $q \in [p^2, (2-p)^2]$*, then the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by FISTA-Mod converges weakly to a global minimizer* $x^\star$ *of* Φ*. Moreover, there holds* $\|x_k - x_{k-1}\| = o(1/k)$*.*

3.2.2 Proofs of Theorem 3.3

Before presenting the proof of Theorem 3.3, we recall first the keys of establishing $O(1/k^2)$ convergence for FISTA-BT [6] and $o(1/k^2)$ convergence rate [10, 2].

The pillars for establishing $O(1/k^2)$ convergence rate for FISTA-BT in [6] can be summarised as

- t_k grows to $+\infty$ at a proper speed, *e.g.* $t_k \approx \frac{k+1}{2}$ as pointed out in [6];
- the sequence $\{t_k\}_{k\in\mathbb{N}}$ satisfies

$$t_k^2 - t_k \le t_{k-1}^2. \tag{3.8}$$

In particular, for $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$, one has $t_k^2 - t_k = t_{k-1}^2$.

To further improve the $O(1/k^2)$ convergence rate to $o(1/k^2)$, the key is that the difference $t_{k-1}^2 - (t_k^2 - t_k)$ should also grow to $+\infty$ [10, 2]. For instance, for the FISTA-CD update rule (1.5), one has

$$t_{k-1}^2 - (t_k^2 - t_k) = \tfrac{1}{d^2}\big((d-2)k + d^2 - 3d + 3\big), \tag{3.9}$$

which goes to $+\infty$ as long as $d > 2$ [10, Eq. (13)]. It is worth noting that $t_{k-1}^2 - (t_k^2 - t_k) \to +\infty$ is also the key for proving the convergence of the iterates $\{x_k\}_{k\in\mathbb{N}}$.

We start with the following supporting lemmas. Recall in (3.4) that $t_k \approx \frac{k+1}{2}p$, we show in the lemma below that $\frac{k+1}{2}p$ actually is a lower bound of t_k.

Lemma 3.6 (Lower bound of t_k). *For the t_k update rule* (3.3)*, let* $r = 4$ *and* $p \in]0,1], q > 0$*. Let* $t_0 = 1$*, then for all* $k \in \mathbb{N}$*, there holds*

$$t_k \ge \frac{(k+1)p}{2}. \tag{3.10}$$

Remark 3.7. When $p = 1$, then we have $t_k \ge \frac{k+1}{2}$ which recovers [6, Lemma 4.3].

Proof. Since $p \in]0,1]$, we have $t_0 = 1 \geq \frac{p}{2}$, and $t_1 = \frac{p+\sqrt{q+4}}{2} \geq \frac{p+2}{2} \geq p$. Now suppose (3.10) holds for a given $k \in \mathbb{N}$, *i.e.* $t_k \geq \frac{(k+1)p}{2}$. Then for $k+1$, we have

$$t_{k+1} - \frac{p}{2} = \frac{p + \sqrt{q + 4t_k^2}}{2} - \frac{p}{2} > \frac{p + 2t_k}{2} - \frac{p}{2} = t_k,$$

which concludes the proof. □

Lemma 3.8 (Lower bound of $t_{k-1}^2 - (t_k^2 - t_k)$). *For the t_k update rule (3.3), let $r = 4$ and $p \in [0,1], p^2 - q \leq 0$. Then there holds*

$$\frac{p(1-p)(k+1)}{2} \leq t_{k-1}^2 - (t_k^2 - t_k) \tag{3.11}$$

Remark 3.9. The inequality (3.11) implies that, if we choose $p < 1$, then $t_{k-1}^2 - (t_k^2 - t_k) \to +\infty$.

Proof. For (3.3), when $r = 4$, we have $t_k = \frac{p+\sqrt{q+4t_{k-1}^2}}{2} \Leftrightarrow t_k^2 - pt_k + \frac{1}{4}(p^2 - q) = t_{k-1}^2$. Since $p^2 \leq q$, then

$$\begin{aligned} t_k^2 - pt_k + \tfrac{1}{4}(p^2 - q) = t_{k-1}^2 \implies & \ t_k^2 - pt_k \leq t_{k-1}^2 \\ \iff & \ t_k^2 - t_k + (1-p)t_k \leq t_{k-1}^2 \\ \implies & \ (1-p)t_k \leq t_{k-1}^2 - (t_k^2 - t_k) \\ \text{(Lemma 3.6)} \implies & \ \frac{p(1-p)(k+1)}{2} \leq (1-p)t_k \leq t_{k-1}^2 - (t_k^2 - t_k), \end{aligned} \tag{3.12}$$

which concludes the proof. □

Remark 3.10. The first line of (3.12) implies that $t_k^2 - t_{k-1}^2 \leq pt_k$. Recently it is shown in [1] that $p < 1$ is the key for proving the convergence of the iterates $\{x_1\}_{k\in\mathbb{N}}$; see [1, Theorem 2.1].

Proofs of Theorem 3.3. For (3.3), when $r = 4$, we have t_k is monotonically increasing as $t_k - t_{k-1} \geq \frac{p}{2} > 0$. Moreover, there holds

$$\begin{aligned} t_k^2 - pt_k + \tfrac{1}{4}(p^2 - q) = t_{k-1}^2 \iff & \ t_k^2 - t_k + (1-p)t_k + \tfrac{1}{4}(p^2 - q) = t_{k-1}^2 \\ \implies & \ t_k^2 - t_k + (1-p)t_0 + \tfrac{1}{4}(p^2 - q) \leq t_{k-1}^2 \\ (t_0 = 1) \iff & \ t_k^2 - t_k + \tfrac{1}{4}((2-p)^2 - q) \leq t_{k-1}^2 \\ \text{(owing to (3.6))} \implies & \ t_k^2 - t_k \leq t_{k-1}^2. \end{aligned}$$

Define $v_k = \Phi(x_k) - \Phi(x^\star)$. Applying Lemma 2.5 at the points $(x = x_k, y = y_k)$ and at $(x = x^\star, y = y_k)$ leads to

$$\begin{aligned} \frac{2}{L}(v_k - v_{k+1}) &\geq \|x_{k+1} - y_k\|^2 + 2\langle x_{k+1} - y_k,\, y_k - x_k\rangle \\ -\frac{2}{L}v_{k+1} &\geq \|x_{k+1} - y_k\|^2 + 2\langle x_{k+1} - y_k,\, y_k - x^\star\rangle, \end{aligned}$$

where $x_{k+1} = e_\gamma(y_k)$ (2.4) is used. Multiplying $t_k - 1$ to the first inequality and then adding to the second one yield,

$$\frac{2}{L}\big((t_k - 1)v_k - t_k v_{k+1}\big) \geq t_k\|x_{k+1} - y_k\|^2 + 2\langle x_{k+1} - y_k,\, t_k y_k - (t_k - 1)x_k - x^\star\rangle.$$

Multiply t_k to both sides of the above inequality and use the result $t_k^2 - t_k \leq t_{k-1}^2$, we get

$$\frac{2}{L}\big(t_{k-1}^2 v_k - t_k^2 v_{k+1}\big) \geq t_k^2\|x_{k+1} - y_k\|^2 + 2t_k\langle x_{k+1} - y_k,\, t_k y_k - (t_k - 1)x_k - x^\star\rangle.$$

Apply the Pythagoras relation $2\langle b - a,\, a - c\rangle = \|b - c\|^2 - \|a - b\|^2 - \|a - c\|^2$ to the last inner product of the above inequality we get

$$\begin{aligned} \frac{2}{L}\big(t_{k-1}^2 v_k - t_k^2 v_{k+1}\big) &\geq \|t_k x_{k+1} - (t_k - 1)x_k - x^\star\|^2 - \|t_k y_k - (t_k - 1)x_k - x^\star\|^2 \\ &= \|t_k x_{k+1} - (t_k - 1)x_k - x^\star\|^2 - \|t_{k-1} x_k - (t_{k-1} - 1)x_{k-1} - x^\star\|^2. \end{aligned} \tag{3.13}$$

If $a_k - a_{k+1} \geq b_{k+1} - b_k$ and $a_1 + b_1 < c$, then $a_k < c$ for all $k \geq 1$ [6, Lemma 4.2]. Hence, (3.13) yields,

$$\tfrac{2}{L} t_k^2 v_k \leq \|x_0 - x^\star\|.$$

Apply Lemma 3.6, we get

$$\Phi(x_k) - \Phi(x^\star) \leq \tfrac{2L}{p^2(k+1)^2} \|x_0 - x^\star\|^2,$$

which concludes the proof for the first claim (3.7).

Let $u_k = x_k + t_k(x_{k+1} - x_k)$. Applying Lemma 2.6 with $y = y_k, y^+ = x_{k+1}$ and $x = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k}x^\star$ yields

$$\Phi(x_{k+1}) + \tfrac{1}{2\gamma}\|\tfrac{1}{t_k}u_k - \tfrac{1}{t_k}x^\star\|^2 \leq \Phi\big((1 - \tfrac{1}{t_k})x_k + \tfrac{1}{t_k}x^\star\big) + \tfrac{1}{2\gamma}\|\tfrac{1}{t_k}u_{k-1} - \tfrac{1}{t_k}x^\star\|^2.$$

Applying the convexity of Φ, we further get

$$\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) - (1 - \tfrac{1}{t_k})\big(\Phi(x_k) - \Phi(x^\star)\big) \leq \tfrac{1}{2\gamma t_k^2}\big(\|u_{k-1} - x^\star\|^2 - \|u_k - x^\star\|^2\big).$$

Multiply t_k^2 to both sides of the above inequality,

$$t_k^2\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) - (t_k^2 - t_k)\big(\Phi(x_k) - \Phi(x^\star)\big) \leq \tfrac{1}{2\gamma}\big(\|u_{k-1} - x^\star\|^2 - \|u_k - x^\star\|^2\big).$$

From Lemma 3.8, we have $\frac{p(1-p)(k+1)}{2} - t_{k-1}^2 \leq -(t_k^2 - t_k)$, hence

$$\begin{aligned} &t_k^2\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) - t_{k-1}^2\big(\Phi(x_k) - \Phi(x^\star)\big) + \tfrac{p(1-p)(k+1)}{2}\big(\Phi(x_k) - \Phi(x^\star)\big) \\ &\leq \tfrac{1}{2\gamma}\big(\|u_{k-1} - x^\star\|^2 - \|u_k - x^\star\|^2\big). \end{aligned}$$

Summing the inequality from $k = 1$ to K, we get

$$t_K^2\big(\Phi(x_{K+1}) - \Phi(x^\star)\big) + \tfrac{p(1-p)}{2}\textstyle\sum_{j=1}^K j\big(\Phi(x_j) - \Phi(x^\star)\big) \leq \tfrac{1}{2\gamma}\big(\|v_0 - x^\star\|^2 - \|v_K - x^\star\|^2\big),$$

which means that $\sum_{j=1}^{+\infty} j\big(\Phi(x_j) - \Phi(x^\star)\big) < +\infty$, that is $\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$. □

3.2.3 Proofs of Theorem 3.5

We now turn to the convergence proof of $\{x_k\}_{k\in\mathbb{N}}$, which is inspired by [10]. The key to prove the convergence of $\{x_k\}_{k\in\mathbb{N}}$ is obtaining the summability

$$\textstyle\sum_{k\in\mathbb{N}} k\|x_k - x_{k-1}\|^2 < +\infty.$$

As previously pointed out, the major difference between the t_k update of FISTA-BT (1.4) and FISTA-CD (1.5) is that $t_{k-1}^2 - (t_k^2 - t_k) \to +\infty$ for FISTA-CD. For the proposed FISTA-Mod schemes, as $\frac{p(1-p)k}{2} \leq t_{k-1}^2 - (t_k^2 - t_k)$ also goes to $+\infty$ as long as p is strictly smaller than 1, this allows us to adapt the proof of [10] to FISTA-Mod, hence proving the convergence of $\{x_k\}_{k\in\mathbb{N}}$.

We need two supporting lemmas before presenting the proof of Theorem 3.5. Given $\ell \in \mathbb{N}_+$, define the truncated sum $S_\ell \overset{\text{def}}{=} \frac{q}{4p}\sum_{i=0}^{\ell}\frac{1}{1+i}$ and a new sequence $\bar{t}_k$ by

$$\bar{t}_k \overset{\text{def}}{=} 1 + S_\ell + \big(\tfrac{p}{2} + \tfrac{q}{4p(\ell+1)}\big)k.$$

We have the following lemma showing that $\bar{t}_k$ serves an upper bound of t_k.

Lemma 3.11 (Upper bound of t_k). *For the t_k update rule (3.3), let $r = 4$ and $p, q \in [0, 1]$. For all $k \in \mathbb{N}$, there holds*

$$t_k \leq \bar{t}_k. \tag{3.14}$$

The purpose of bounding t_k from above by a linear function of k is so that we can eventually bound a_k from above, which is needed by the following lemma.

Proof. Given t_k, t_{k+1}, we have

$$t_{k+1} - t_k = \frac{p + \sqrt{q + 4t_k^2}}{2} - t_k = \frac{p}{2} + \frac{\sqrt{q + 4t_k^2} - 2t_k}{2} \leq \frac{p}{2} + \frac{\sqrt{(2t_k + q/(4t_k))^2} - 2t_k}{2} = \frac{p}{2} + \frac{q}{8t_k},$$

which leads to

$$t_{k+1} \le t_k + \tfrac{p}{2} + \tfrac{q}{8t_k} \le 1 + \tfrac{p}{2}k + \sum\nolimits_{i=0}^{k} \tfrac{q}{8t_i} \le 1 + \tfrac{p}{2}k + \tfrac{q}{4p}\sum\nolimits_{i=0}^{k} \frac{1}{i+1}$$
$$\le 1 + \tfrac{p}{2}k + \tfrac{q}{4p}\sum\nolimits_{i=0}^{\ell} \frac{1}{i+1} + \tfrac{q}{4p}\sum\nolimits_{i=\ell+1}^{k} \frac{1}{\ell+1}$$
$$\le 1 + \tfrac{p}{2}k + \tfrac{q}{4p}\sum\nolimits_{i=0}^{\ell} \frac{1}{i+1} + \tfrac{q}{4p}k\frac{1}{\ell+1}$$
$$= 1 + \tfrac{p}{2}k + S_\ell + \tfrac{q}{4p}k\frac{1}{\ell+1} = \bar{t}_k,$$

and we conclude the proof. □

Denote $\lceil x \rceil$ the largest integer that is smaller than x, and define the following two constants

$$b \overset{\text{def}}{=} \left\lceil \frac{p+2}{p+q/(2p(\ell+1))} \right\rceil \quad \text{and} \quad c \overset{\text{def}}{=} \frac{p+2+2S_\ell}{p+q/(2p(\ell+1))}.$$

Lemma 3.12. *For all $j \ge 1$, define*

$$\beta_{j,k} \overset{\text{def}}{=} \prod\nolimits_{\ell=j}^{k} a_\ell,$$

for all j, k, and $\beta_{j,k} = 1$ for all $k < j$. Let $\ell \ge \lceil \frac{q}{p(2-p)} \rceil$, then for all j,

$$\sum\nolimits_{k=j}^{\infty} \beta_{j,k} \le j + c + 2b. \tag{3.15}$$

Proof. We first show that a_k is bounded from above. From the definition of a_k we have

$$a_k = \frac{t_{k-1}-1}{t_k} = \frac{2t_{k-1}-2}{p+\sqrt{q+4t_{k-1}^2}} \le \frac{p+2t_{k-1}-2-p}{p+2t_{k-1}} = 1 - \frac{2+p}{p+2t_{k-1}}$$
$$\text{(Lemma 3.11)} \le 1 - \frac{2+p}{p+2+2S_\ell+(p+\frac{q}{2p(\ell+1)})k} = 1 - \frac{b}{k+c}. \tag{3.16}$$

From (3.16) we have that

$$\beta_{j,k} = \prod\nolimits_{\ell=j}^{k} a_\ell \le \prod\nolimits_{\ell=j}^{k} \frac{\ell+c-b}{\ell+c}.$$

For $k = j, ..., j+2b-1$, we have $\beta_{j,k} < 1$. Then for $k - j \ge 2b$,

$$\beta_{j,k} \le \prod\nolimits_{\ell=j}^{k} \frac{\ell+c-b}{\ell+c} = \frac{j+c-b}{j+c}\,\frac{j+1+c-b}{j+1+c}\cdots\frac{j+c}{j+b+c}\,\frac{j+1+c}{j+b+1+c}\cdots\frac{k+c-b}{k+c}$$
$$= \frac{(j+c-b)\cdots(j+c-1)}{(k+c-b+1)\cdots(k+c)} \le \frac{(j+c-1)^b}{(k+c-b+1)^b}.$$

Therefore,

$$\sum\nolimits_{k=j}^{\infty}\beta_{j,k} \le 2b + \sum\nolimits_{k=j+2b}^{\infty}\beta_{j,k} \le 2b + (j+c-1)^b \sum\nolimits_{k=j+2b}^{\infty} \frac{1}{(k+c-b+1)^b}$$
$$\le 2b + (j+c-1)^b \int_{x=j+2b}^{\infty} \frac{1}{(x+c-b+1)^b}\mathrm{d}x$$
$$\le 2b + (j+c-1)^b \frac{1}{b-1}\frac{1}{(j+b+c+1)^{b-1}}$$
$$\le 2b + \frac{1}{b-1}(j+c-1) \le j+c+2b.$$

The last inequality uses the fact that $b \ge 2$ for $\ell \ge \lceil \frac{q}{p(2-p)} \rceil$. □

Proofs of Theorem 3.5. Applying Lemma 2.6 with $y = y_k$ and $x = x_k$, we get

$$\Phi(x_{k+1}) + \frac{\|x_k - x_{k+1}\|^2}{2\gamma} \le \Phi(x_k) + a_k^2\frac{\|x_{k-1} - x_k\|^2}{2\gamma},$$

which means, defining $\Delta_k \overset{\text{def}}{=} \frac{1}{2}\|x_k - x_{k-1}\|^2$,

$$\Delta_{k+1} - a_k^2\Delta_k \le \gamma(w_k - w_{k+1}).$$

Denote the upper bound of a_k in (3.16) as $\bar{a}_k \overset{\text{def}}{=} 1 - \frac{b}{k+c}, \forall k \geq 2$, and let $\bar{a}_1 = 0$ since $a_1 = 0$. It is then straightforward that

$$\Delta_{k+1} - \bar{a}_k^2 \Delta_k \leq \Delta_{k+1} - a_k^2 \Delta_k \leq \gamma(w_k - w_{k+1}).$$

Multiplying the above inequality with $(k+c)^2$ and summing from $k = 1$ to K lead to

$$\textstyle\sum_{k=1}^K (k+c)^2 (\Delta_{k+1} - \bar{a}_k^2 \Delta_k) \leq \gamma \sum_{k=1}^K (k+c)^2 (w_k - w_{k+1}).$$

Since $\bar{a}_1 = 0$, we derive from above that

$$\begin{aligned}
\textstyle\sum_{k=1}^K (k+c)^2 (\Delta_{k+1} - \bar{a}_k^2 \Delta_k) &= (K+c)^2 \Delta_{K+1} + \textstyle\sum_{k=2}^K \big((k+c-1)^2 - (k+c)^2 \bar{a}_k^2\big) \Delta_k \\
&= (K+c)^2 \Delta_{K+1} + \textstyle\sum_{k=2}^K \big((k+c-1)^2 - (k+c-b)^2\big) \Delta_k \\
&\leq (K+c)^2 \Delta_{K+1} + \textstyle\sum_{k=2}^K 2(b-1)(k+c) \Delta_k \\
&\leq \gamma\big((c+1)^2 w_1 - (c+K)^2 w_{K+1}\big) + \gamma \textstyle\sum_{k=2}^K \big((k+c)^2 - (k+c-1)^2\big) w_k \\
&\leq \gamma\big((c+1)^2 w_1 - (c+K)^2 w_{K+1}\big) + 2\gamma \textstyle\sum_{k=2}^K (k+c) w_k.
\end{aligned}$$

From the proof of Theorem 3.3, we have that $\sum_{k \in \mathbb{N}} k w_k < +\infty$, which in turn implies that $\{k\Delta_k\}_{k \in \mathbb{N}}$ is *summable* and that sequence $\{k^2 \Delta_k\}_{k \in \mathbb{N}}$ is bounded, which also indicates $\|x_k - x_{k-1}\| = o(1/k)$.

Now define

$$\psi_k \overset{\text{def}}{=} \tfrac{1}{2}\|x_k - x^\star\|^2 \quad \text{and} \quad \phi_k \overset{\text{def}}{=} \tfrac{1}{2}\|y_k - x_{k+1}\|^2.$$

By applying the definition of y_k, we have

$$\begin{aligned}
\psi_k - \psi_{k+1} &= \tfrac{1}{2}\langle x_k - x^\star + x_{k+1} - x^\star,\, x_k - x_{k+1}\rangle \\
&= \Delta_{k+1} + \langle y_{a,k} - x_{k+1},\, x_{k+1} - x^\star\rangle - a_k \langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle \\
&\geq \Delta_{k+1} + \gamma\langle \nabla F(y_k) - \nabla F(x^\star),\, x_{k+1} - x^\star\rangle - a_k \langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle.
\end{aligned} \tag{3.17}$$

As ∇F is $\frac{1}{L}$-cocoercive (Definition 2.2), applying Young's inequality yields

$$\begin{aligned}
&\langle \nabla F(y_k) - \nabla F(x^\star),\, x_{k+1} - x^\star\rangle \\
&\geq \tfrac{1}{L}\|\nabla F(y_k) - \nabla F(x^\star)\|^2 + \langle \nabla F(y_k) - \nabla F(x^\star),\, x_{k+1} - y_k\rangle \\
&\geq \tfrac{1}{L}\|\nabla F(y_k) - \nabla F(x^\star)\|^2 - \tfrac{1}{L}\|\nabla F(y_k) - \nabla F(x^\star)\|^2 - \tfrac{L}{2}\phi_k = -\tfrac{L}{2}\phi_k.
\end{aligned} \tag{3.18}$$

Back to (3.17), we get

$$\psi_k - \psi_{k+1} \geq \Delta_{k+1} - \tfrac{\gamma L}{2}\phi_k - a_k\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle. \tag{3.19}$$

For $\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle$, we have

$$\begin{aligned}
\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle &= \langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle + \langle x_k - x_{k-1},\, x_k - x^\star\rangle \\
&= \langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle + (\Delta_k + \psi_k - \psi_{k-1}),
\end{aligned} \tag{3.20}$$

where we applied the usual Pythagoras relation to $\langle x_k - x_{k-1},\, x_k - x^\star\rangle$. Putting (3.20) back into (3.19) and rearranging terms yield

$$\begin{aligned}
\psi_{k+1} - \psi_k - a_k(\psi_k - \psi_{k-1}) &\leq -\Delta_{k+1} + \tfrac{\gamma L}{2}\phi_k + a_k\langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle + a_k\Delta_k \\
&= -\Delta_{k+1} + \tfrac{\gamma L}{2}\phi_k + \langle y_k - x_k,\, x_{k+1} - x_k\rangle + a_k\Delta_k \\
&= -\Delta_{k+1} + \tfrac{\gamma L}{2}\phi_k + \big(a_k^2\Delta_k + \Delta_{k+1} - \tfrac{1}{2}\|y_k - x_{k+1}\|^2\big) + a_k\Delta_k \\
&= \frac{\gamma L - 1}{2}\phi_k + (a_k + a_k^2)\Delta_k,
\end{aligned} \tag{3.21}$$

where the Pythagoras relation is applied again to $\langle y_k - x_k,\, x_{k+1} - x_k\rangle$. Since $\gamma \in]0, 1/L]$ and $a_k \leq 1$, we get from above that

$$\psi_{k+1} - \psi_k - a_k(\psi_k - \psi_{k-1}) \leq 2a_k\Delta_k.$$

Define $\xi_k = \max\{0, \psi_k - \psi_{k-1}\}$, then

$$\xi_{k+1} \leq a_k(\xi_k + 2\Delta_k) \leq 2\sum_{j=2}^{k}\left(\prod_{l=j}^{k} a_l\right)\Delta_j = 2\sum_{j=2}^{k}\beta_{j,k}\Delta_j,$$

Applying Lemma 3.12 and the summability of $\{k\Delta_k\}_{k\in\mathbb{N}}$ lead to

$$\sum_{k=2}^{+\infty}\xi_k \leq 2\sum_{k=1}^{+\infty}\sum_{j=2}^{k}\beta_{j,k}\Delta_j = 2\sum_{j=2}^{k}\Delta_j\sum_{k=1}^{+\infty}\beta_{j,k} \leq 2\sum_{j=2}^{k}(j+c+2b)\Delta_j < +\infty.$$

Then we have

$$\Phi_{k+1} - \sum_{j=1}^{k+1}[\theta_j]_+ \leq \Phi_{k+1} - \theta_{k+1} - \sum_{j=1}^{k}[\theta_j]_+ = \Phi_k - \sum_{j=1}^{k}[\theta_j]_+,$$

which means that $\{\Phi_k - \sum_{j=1}^{k}[\theta_j]_+\}_{k\in\mathbb{N}}$ is monotone non-increasing, hence convergent. It is immediate that sequence $\{\Phi_k\}_{k\in\mathbb{N}}$ is also convergent, meaning that $\lim_{k\to+\infty}\|x_k - x^\star\|$ exists for any $x^\star \in \mathrm{zer}(A+B)$.

Let $\bar{x}$ be a weak cluster point of $\{x_k\}_{k\in\mathbb{N}}$, and let us fix a subsequence, say $x_{k_j} \rightharpoonup \bar{x}$. Applying Lemma 2.4 with $y = y_{k_j}$, we get

$$u_{k_j} \overset{\text{def}}{=} \frac{y_{k_j} - x_{k_j+1}}{\gamma} - \nabla F(y_{k_j}) \in \partial R(x_{k_j+1}).$$

Since ∇F is cocoercive and $y_{k_j} = x_{k_j} + a_{k_j}(x_{k_j} - x_{k_j-1}) \rightharpoonup \bar{x}$, we have $\nabla F(y_{k_j}) \to \nabla F(\bar{x})$. In turn, $u_{k_j} \to -\nabla F(\bar{x})$ since $\gamma > 0$. Since $(x_{k_j+1}, u_{k_j}) \in \mathrm{gph}\,(\partial R)$, and the graph of the maximal monotone operator ∂R is sequentially weakly-strongly closed in $\mathcal{H}\times\mathcal{H}$, we get that $-\nabla F(\bar{x}) \in \partial R(\bar{x})$, *i.e.* $\bar{x}$ is a solution of ($\mathcal{P}$). Opial's Theorem [26] then concludes the proof. □

4 Lazy-start strategy

From the previous section, it can be concluded that a crucial difference between FISTA-Mod (also FISTA-CD) and FISTA-BT is that the former can control the behaviour of t_k via p, q, r (d for FISTA-CD). In this section, we show that such degree of freedom provided by these parameters allows us to design strategies which can make FISTA schemes much faster in practice.

The strategy developed in this section is called "lazy-start", whose main idea is choosing properly the values of p, q for FISTA-Mod and d for FISTA-CD, such that they can slow down the speed of a_k approaching 1.

Proposition 4.1 (Lazy-start FISTA). *For FISTA-Mod and FISTA-CD, consider the following choices of p, q and d respectively:*

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

We consider a least square problem [23] to explain how "lazy-start" can significantly improve the practical performance of FISTA schemes:

$$\min_{x\in\mathbb{R}^n}\left\{F(x) \overset{\text{def}}{=} \tfrac{1}{2}\|Ax-b\|^2\right\}, \tag{4.1}$$

where $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is of the form

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \cdots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}_n.$$

The FISTA-CD scheme is considered, and specialise to the case of solving (4.1), we get

$$\begin{aligned} y_k &= x_k + \frac{k-1}{k+d}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \gamma\nabla F(y_k). \end{aligned} \tag{4.2}$$

Two different values of d are compared:

- Normal FISTA-CD with $d = d_1 = 2$;

- Lazy-start FISTA-CD with $d = d_2 = 20$.

In the numerical test, we set $n = 201$. The convergence profiles of $\|x_k - x^\star\|$ for the above two choices of d are plotted in Figure 2, where the *red line* represents $d_1 = 2$ while the *black line* stands for $d_2 = 20$. The starting points x_0 of two cases are the same and chosen such that $\|x_0 - x^\star\| = 1$. It can be observed that the lazy-start one is significantly faster than the normal choice after $k = 2 \times 10^5$.

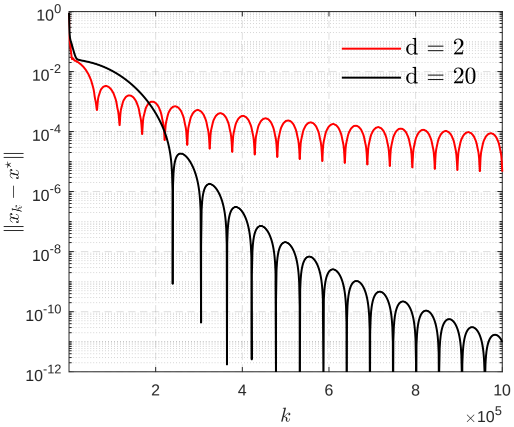


Figure 2: Convergence profiles of $\|x_k - x^\star\|$. The *red solid line* stands for $d_1 = 2$ and the *black solid line* stands for $d_2 = 20$.

The above difference appears not only for (4.1), but rather an observation from many problems; see Section 7 for more examples. To explain the such a difference, we need several intermediate steps:

(1) Let $x^\star$ be the unique solution of (4.1). As the problem is quadratic, (4.2) can be written in to a fixed-point iteration $z_{k+1} = M_k z_k$ where $z_k = (x_k - x^\star; x_{k-1} - x^\star)$ and see (4.3) for the definition of M_k. Define $\widetilde{M}_k \overset{\text{def}}{=} \prod_{i=1}^{k-1} M_{k-i}$, then we have $z_k = \widetilde{M}_k z_1$.

(2) Let ρ_k be the leading eigenvalue of M_k, then for any $a_k \in [0, 1]$, we have $|\rho_k| < 1$.

(3) Let $\tilde{\rho}_k$ be the leading eigenvalue if $\widetilde{M}_k$, though can not be proved, we can show numerically that $|\tilde{\rho}_k| \le \mathcal{C}_1 \prod_{i=1}^{k-1} |\rho_i|$ where $\mathcal{C}_1$ is a constant. The spectral radius theorem also gives $\|\widetilde{M}_k\| \le \mathcal{C}_2 |\tilde{\rho}_k|$ where $\mathcal{C}_2$ is also a constant. All together imply that we can bound $\|\widetilde{M}_k\|$ by $\prod_{i=1}^{k-1} |\rho_i|$.

(4) The discussion then boils down to compare the value of $\prod_{i=1}^{k-1} |\rho_i|$ under different choices of d, and it can be shown that for $d = 20$ the value of $\prod_{i=1}^{k-1} |\rho_i|$ can be several order smaller than that of $d = 2$ for large enough k, hence showing the advantage of lazy-start strategy.

For the rest of the section, we discuss the above steps in details.

4.1 Fixed-point formulation and spectral properties

Since the problem is strongly convex, it admits a unique solution which is denoted as $x^\star$. Moreover, owing to the quadratic form of F, its gradient reads $\nabla F(x) = A^T(Ax - b)$, and it is easy to obtain from (4.2) that,

$$x_{k+1} - x^\star = G(y_k - x^\star) = (1 + a_k)G(x_k - x^\star) - G(x_k - x^\star),$$

where $G \overset{\text{def}}{=} \mathrm{Id} - \frac{1}{L}A^T A$. Now define

$$z_k \overset{\text{def}}{=} \begin{pmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{pmatrix} \quad \text{and} \quad M_k \overset{\text{def}}{=} \begin{bmatrix} (1 + a_k)G & -a_k G \\ \mathrm{Id} & 0 \end{bmatrix}. \tag{4.3}$$

Then it is immediate that

$$z_{k+1} = M_k z_k. \tag{4.4}$$

Recursively apply the above relation, we get

$$z_k = \left(\prod_{i=1}^{k-1} M_{k-i}\right) z_1,$$

and for the sake of simplicity we denote $\widetilde{M}_k \overset{\text{def}}{=} \prod_{i=1}^{k-1} M_{k-i}$.

4.1.1 Spectral property of M_k

We first present the spectral property of M_k by invoking existing result from [16]. Denote $\alpha > 0, \eta < 1$ the smallest and largest eigenvalues $A^T A$ and G, respectively. We then have $\eta = 1 - \gamma\alpha$. Given M_k, denote ρ_k its leading eigenvalue, then ρ_k can be expressed by η and a_k, and their relation can be described by the following lemma taken from Proposition 4.6 and Section 4.4 in [16].

Lemma 4.2 ([16]). *Suppose $(v_1; v_2)$ is an eigenvector of M_k corresponding to eigenvalue ρ_k, then it must satisfy $v_1 = \rho_k v_2$. Moreover, v_2 is an eigenvector of G associated to the eigenvalue η, and*

- *The expression of ρ_k reads*

$$\rho_k = \frac{(1+a_k)\eta + \sqrt{(1+a_k)^2\eta^2 - 4a_k\eta}}{2} \tag{4.5}$$

- *The magnitude of ρ_k is*

$$|\rho_k| = \begin{cases} \frac{(1+a_k)\eta + \sqrt{(1+a_k)^2\eta^2 - 4a_k\eta}}{2} < 1 : (1+a_k)^2\eta \geq 4a_k, \\ \sqrt{a_k\eta} < 1 : (1+a_k)^2\eta \leq 4a_k. \end{cases} \tag{4.6}$$

Moreover, $|\rho_k|$ attains the minimal value $\rho^\star = 1 - \sqrt{\gamma\alpha}$ when a_k equals to $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$.

Remark 4.3. We refer to [16, 14] for more details about the dependence of ρ_k on η and a_k. Below, we specify several situations of Lemma 4.2 and moreover its connection with Nesterov's optimal scheme [23].

- From (4.5) and (4.6), simple calculation yields $(1+a^\star)^2\eta = 4a^\star$ and

$$\rho_k \begin{cases} \text{real} : (1+a_k)^2\eta \geq 4a_k, \\ \text{complex} : (1+a_k)^2\eta \leq 4a_k. \end{cases}$$

  According to [23, Constant Step Scheme III], $a^\star$ is the optimal inertial parameter when the problem is strongly convex, and $\rho^\star$ is the optimal convergence rate can be achieve by (4.2).

  The complex eigenvalue ρ_k is also the reason why FISTA oscillates. More precisely, as long as one has $a_k \in]a^\star, 1]$, ρ_k will be complex and the iteration (4.2) will oscillate.

- Eq. (4.6) indicates that $|\rho_k| = \sqrt{\eta} > \eta$ for $a_k = 1$. For FISTA schemes, as $\lim_{k\to+\infty} a_k = 1$, this means for strongly convex problems, FISTA schemes eventually is slower than the vanilla gradient descent/Forward–Backward. We refer to [16] for more discussions.

4.1.2 Spectral property of $\widetilde{M}_k$

Now we turn to the spectral property of $\widetilde{M}_k$, unfortunately, unlike the case of M_k, this time we can only discuss through numerical illustration.

Let $\tilde{\rho}_k$ be the leading eigenvalue of $\widetilde{M}_k$, in general there is no clear corresponding between $\tilde{\rho}_k$ and $\rho_i, i = 1, ..., k-1$. For instance, there is no $\tilde{\rho}_k = \prod_{i=1}^{k-1} \rho_{k-i}$, nor $|\tilde{\rho}_k| = \prod_{i=1}^{k-1} |\rho_{k-i}|$. However, $|\tilde{\rho}_k|$ can be bounded from above by $\prod_{i=1}^{k-1} |\rho_{k-i}|$. Owing to the spectral theorem, we can bound $\|\widetilde{M}_k\|$ from above by $|\tilde{\rho}_k|$. All these together mean we can bound $\|\widetilde{M}_k\|$ from above by $\prod_{i=1}^{k-1} |\rho_{k-i}|$ which is the content of the next proposition.

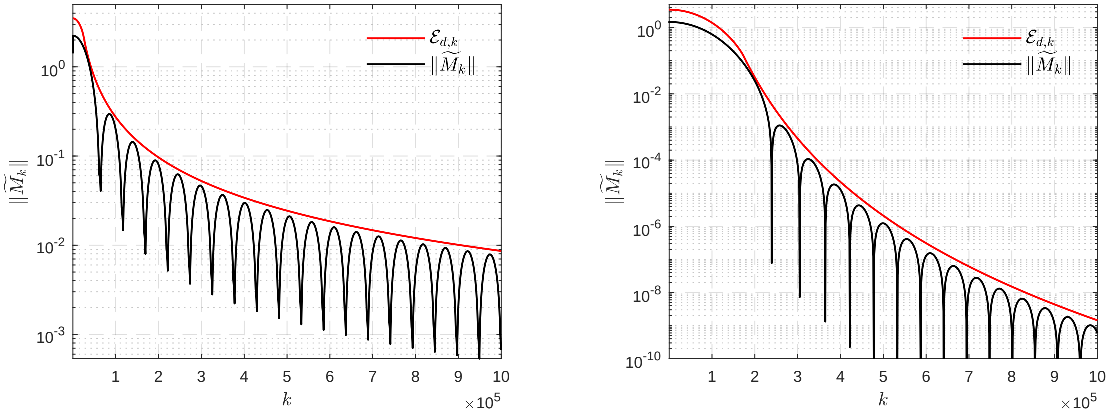
Proposition 4.4 (Envelope of $\|\widetilde{M}_k\|$). *For the matrix $\widetilde{M}_k = \prod_{i=1}^{k-1} M_{k-i}$, let ρ_i be the eigenvalue of M_i for $i = 1, ..., k-1$, then there exists $\mathcal{T} > 0$ such that*

$$\|\widetilde{M}_k\| \leq \mathcal{T} \prod_{i=1}^{k-1} |\rho_{k-i}| \tag{4.7}$$

holds for all $k \geq 1$. In particular, let $\widetilde{\mathcal{T}}$ be the minimal value such that (4.7) holds, then

$$\mathcal{E}_{d,k} \overset{\text{def}}{=} \Big\{ \widetilde{\mathcal{T}} \prod_{i=1}^{k-1} |\rho_{k-i}| \Big\}_{k\in\mathbb{N}}$$

is called the envelope of $\|\widetilde{M}_k\|$.



(a) $d = 2$

(b) $d = 20$

Figure 3: The value of $\|\widetilde{M}_k\|$ and the corresponding envelope $\mathcal{E}_{d,k}$: (a) $d = 2$, (b) $d = 20$.

For problem (4.1) with $n = 201$, we show graphically in Figure 3 the value of $\|\widetilde{M}_k\|$ and the corresponding envelope $\mathcal{E}_{d,k}$. The plots of Figure 3 (a) correspond to $d = 2$, with the red line standing for $\mathcal{E}_{d,k}$ and the black line being the $\|\widetilde{M}_k\|$, the value of $\widetilde{\mathcal{T}}$ for this case is 3.5. The plots of Figure 3 (b) are for $d = 20$, for which we also have $\widetilde{\mathcal{T}} = 3.5$. Again, note that $\|\widetilde{M}_k\|$ can reach much smaller value for $d = 20$ than that of $d = 2$.

4.2 The advantage of lazy-start

For d_1, d_2, we note $a_{d_1,k}, a_{d_2,k}$ the corresponding inertial parameter, $M_{d_1,k}, M_{d_2,k}$ the matrix of (4.3), then the matrices $\widetilde{M}_{d_1,k}, \widetilde{M}_{d_2,k}$ and corresponding envelopes $\mathcal{E}_{d_1,k}$ and $\mathcal{E}_{d_2,k}$.

4.2.1 Properties of $|\rho_k|$

Since $\mathcal{E}_{d,k}$ is determined by the product of $|\rho_k|$, let us first check the profile of $|\rho_k|$ under $d_1 = 2$ and $d_2 = 20$. Denote $\rho_{d_1,k}, \rho_{d_2,k}$ the leading eigenvalues of $M_{d_1,k}, M_{d_2,k}$, respectively. The modulus of them are shown in Figure 4 (a), where the red line is $|\rho_{d_1,k}|$ and the black line stands for $|\rho_{d_2,k}|$. We can observe that

- For both cases, the values of $|\rho_{d_1,k}|, |\rho_{d_2,k}|$ decrease first, until reaching $\rho^\star = 1 - \sqrt{\gamma\alpha}$ (see Lemma 4.2), and then start to increase until reaching $\sqrt{\eta}$;
- As d_2 slows down the speed of a_k growing (see Figure 1), so does the speed $|\rho_{d_2,k}|$ reaching $\rho^\star$. Such a mismatch of approaching $\rho^\star$ is the key of lazy-start strategy being faster.

Denote K_{eq} the point $|\rho_{d_2,k}|$ equals to $\rho^\star$, then we have

$$K_{\mathrm{eq}} = \left\lceil \frac{1 + a^\star d_2}{1 - a^\star} \right\rceil + 1, \tag{4.8}$$

where $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$ is the optimal value mentioned in Lemma 4.2.

4.2.2 Comparison of $\mathcal{E}_{d,k}$

Next we compare $\mathcal{E}_{d_1,k}, \mathcal{E}_{d_2,k}$, whose values are plotted in Figure 4 (b), where the red and black lines are corresponding to $\mathcal{E}_{d_1,k}$ and $\mathcal{E}_{d_2,k}$ respectively. Observe that, $\mathcal{E}_{d_1,k}$ and $\mathcal{E}_{d_2,k}$ intersect for certain k which turns out very close to K_{eq}[1]. For $k \geq K_{\mathrm{eq}}$, the difference between $\mathcal{E}_{d_1,k}$ and $\mathcal{E}_{d_2,k}$ becomes increasingly larger.

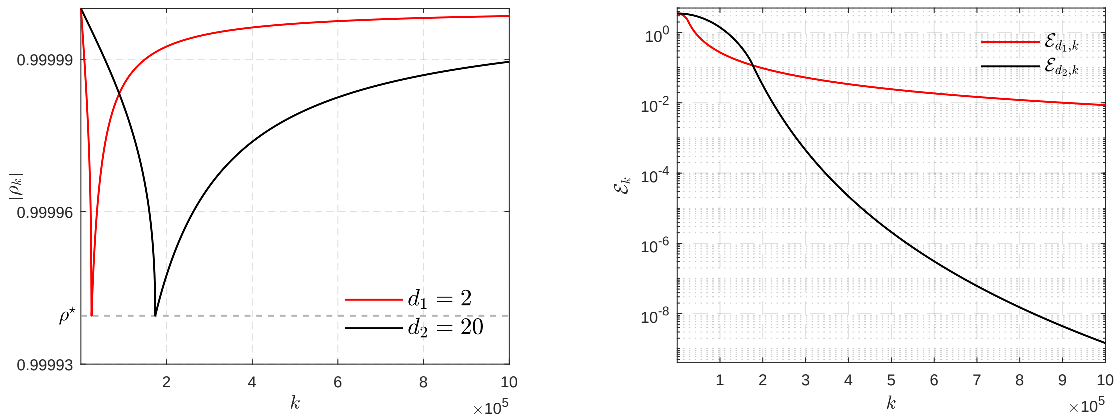
Denote $a_{d_1,k}, a_{d_2,k}$ the corresponding a_k of d_1 and d_2 respectively, then from (4.6) we have that for $k \geq K_{\mathrm{eq}}$,

$$|\rho_{d_1,k}| = \sqrt{a_{d_1,k}\eta} \quad \text{and} \quad |\rho_{d_2,k}| = \sqrt{a_{d_2,k}\eta}$$

and $|\rho_{d_1,k}| \geq |\rho_{d_2,k}|$ since $a_{d_1,k} \geq a_{d_2,k}$. Define $\mathcal{R}_k$ by

$$\mathcal{R}_k \stackrel{\text{def}}{=} \prod\nolimits_{i=K_{\mathrm{eq}}}^{k} \frac{|\rho_{d_1,i}|}{|\rho_{d_2,i}|},$$

[1] The real value of such k is approximately $1.018K_{\mathrm{eq}}$.



(a) Value of ρ_k

(b) Value of $\mathcal{E}_{d,k}$

Figure 4: The value of ρ_k and $\mathcal{E}_{d,k}$ under d_1, d_2.

which is the accumulation of $\frac{|\rho_{d_1,i}|}{|\rho_{d_2,i}|}$. Let $k \geq K_{\text{eq}} + 2(d_2 - d_1)$, then we get

$$\begin{aligned}\mathcal{R}_k &= \prod\nolimits_{i=K_{\text{eq}}}^{k} \frac{|\rho_{d_1,i}|}{|\rho_{d_2,i}|} = \prod\nolimits_{i=K_{\text{eq}}}^{k} \frac{\sqrt{a_{d_1,i}}}{\sqrt{a_{d_2,i}}} = \prod\nolimits_{i=K_{\text{eq}}}^{k} \sqrt{\frac{i+d_2}{i+d_1}} \\ &= \prod\nolimits_{i=K_{\text{eq}}}^{k} \left(\frac{K_{\text{eq}}+d_2}{K_{\text{eq}}+d_1} \frac{K_{\text{eq}}+1+d_2}{K_{\text{eq}}+1+d_1} \cdots \frac{K_{\text{eq}}+d_2-d_1+d_2}{K_{\text{eq}}+d_2-d_1+d_1} \cdots \frac{k-2+d_2}{k-2+d_1} \frac{k-1+d_2}{k-1+d_1} \frac{k+d_2}{k+d_1} \right)^{1/2} \\ &= \prod\nolimits_{j=0}^{d_2-d_1-1} \left(\frac{k+d_1+1+j}{K_{\text{eq}}+d_1+j} \right)^{1/2} \approx \left(\frac{k+d_2}{K_{\text{eq}}+d_2-1} \right)^{(d_2-d_1)/2}.\end{aligned} \tag{4.9}$$

Since $\gamma = 1/L$, define $\mathcal{C} \overset{\text{def}}{=} L/\alpha$ the condition number. Recall the definition of $K_{\text{eq}} = \lceil \frac{1+a^\star d_2}{1-a^\star} \rceil + 1$ and that $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$, we have from(4.9) that

$$\begin{aligned}\mathcal{R}_k \approx \left(\frac{k+d_2}{\frac{1+a^\star d_2}{1-a^\star} + 1 + d_2 - 1} \right)^{(d_2-d_1)/2} &= \left(\frac{(1-a^\star)(k+d_2)}{1+d_2} \right)^{(d_2-d_1)/2} \\ &= \left(1 - \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}} \right)^{(d_2-d_1)/2} \left(\frac{k+d_2}{1+d_2} \right)^{(d_2-d_1)/2} \\ &= \left(\frac{2}{\sqrt{\mathcal{C}}+1} \right)^{(d_2-d_1)/2} \left(\frac{k+d_2}{1+d_2} \right)^{(d_2-d_1)/2}.\end{aligned} \tag{4.10}$$

To verify the accuracy of the above approximation, we consider the problem (4.1). When $n = 201$, we have

$$L = 16 \quad \text{and} \quad \alpha = 5.85 \times 10^{-8}.$$

Consequently, $\mathcal{C} = \frac{L}{\alpha} = 2.735 \times 10^8$. Let $k = 10^6$ and substitute them into (4.10), we have $\mathcal{R}_k \approx 5.98 \times 10^6$, while for $\mathcal{E}_{d,k}$ we have

$$\frac{\mathcal{E}_{d_1,k=10^6}}{\mathcal{E}_{d_2,k=10^6}} = 5.96 \times 10^6,$$

which means (4.10) is a good approximation of (4.9).

The above discussion is mainly about the envelope $\mathcal{E}_{d_1,k}$. In terms of what really happens on $\|x_k - x^\star\|$ for d_1 and d_2: from Figure 2, we have that at $k = 10^6$, $\|x_k - x^\star\|$ of d_1 is about 2×10^6 larger than that of d_2. Compared with 5.98×10^6, we can conclude that the above approximation is able to estimate the order of acceleration obtained by lazy-start strategy.

4.2.3 Quantify the advantage of lazy-start

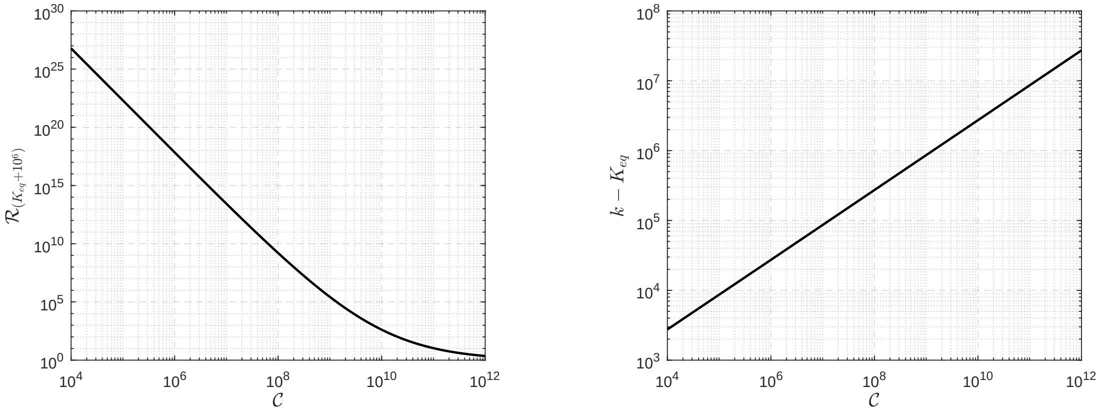
The approximation (4.10) indicates that $\mathcal{R}_k$ is a function of $\mathcal{C}$ and k, in the following we discuss the dependence of $\mathcal{R}_k$ on $\mathcal{C}$ and k from two aspects.

Fix k First consider fixing $k = K_{\text{eq}} + 10^6$ and letting $\mathcal{C} \in [10^4, 10^{12}]$. This setting is to check how much better d_2 is than d_1 in terms of $\|x_k - x^\star\|$ if we run the iteration (4.2) 10^6 more steps after K_{eq}. The obtained value of $\mathcal{R}_k$ is shown below in Figure 5 (a). As we can see, when $\mathcal{C}$ is small, *e.g.* $\mathcal{C} = 10^4$, the advantage can be as large as 10^{27} times and decrease to almost 1 for $\mathcal{C} = 10^{12}$. However it should be noted that for this large condition number, $K_{\text{eq}} + 10^6$ number of iteration steps is far from enough for producing satisfactory outputs.

Fix $\mathcal{R}$ The second aspect is to check for fixed $\mathcal{R}_k = \mathcal{R}$, *e.g.* $\mathcal{R} = 10^5$, how many more steps are needed after K_{eq}. From (4.10), simple calculation yields

$$k - K_{\text{eq}} = \mathcal{R}^{\frac{2}{d_2 - d_1}} \frac{(\sqrt{\mathcal{C}} + 1)(1 + d_2)}{2} - d_2.$$

Let again $\mathcal{C} \in [10^4, 10^{12}]$, the value of $k - K_{\text{eq}}$ is shown in Figure 5 (b). We can observe that when $\mathcal{C} = 10^4$, only around $2,000$ steps are needed, while about 2×10^7 steps for $\mathcal{C} = 10^{12}$.



(a) Value of $\mathcal{R}_k$ when fix $k = K_{\text{eq}} + 10^6$

(b) Value of $k - K_{\text{eq}}$ when fix $\mathcal{R}_k = 10^5$

Figure 5: The dependence of $\mathcal{R}_k$ on the iteration number k and the condition number $\mathcal{C}$: (a) Value of $\mathcal{R}_k$ over $\mathcal{C}$ when fix $k = K_{\text{eq}} + 10^6$; (b) The value of the difference $k - K_{\text{eq}}$ over $\mathcal{C}$ when fix $\mathcal{R}_k = 10^5$.

Remark 4.5. It can be observed from (4.10) that, when fixing $\mathcal{C}$ and k, $\mathcal{R}_k$ increases when d_2 is increasing. This means that if we consider only $\mathcal{R}_k$, then the larger value of d_2 the better. However, one should not do so in practice, as larger d_2 will make the value of K_{eq} also much larger. As a result, proper choices of d_2 should be a trade-off between K_{eq} and $\mathcal{R}_k$, which is the goal of next part.

4.3 Optimal lazy-start parameters

Now we discuss how to practically choose d and the existence of optimal d. The discussion again is delivered through the envelope $\mathcal{E}_{d,k}$ of Proposition 4.4.

4.3.1 Optimal d for $\|x_k - x^\star\|$

We continue using problem (4.1) with $n = 201$, with condition number $\mathcal{C} = 2.735 \times 10^7$. Consider several different values of d which are $d \in [5, 15, 25, 35, 45]$. The values of corresponding $\mathcal{E}_{d,k}$ are plotted in Figure 6 (a). For each $k \in [1, 10^6]$, the minimum of $\mathcal{E}_{d \in [5,15,25,35,45],k}$ is computed and plotted in *red dot line*.

From the plots in Figure 6 (a), it can be observed that for each $d \in [5, 15, 25, 35, 45]$, the corresponding $\mathcal{E}_{d,k}$ is the smallest for certain range of k. For instance, for $d = 5$, $\mathcal{E}_{5,k}$ is the smallest for k between 1 and about 1.75×10^5. This implies that

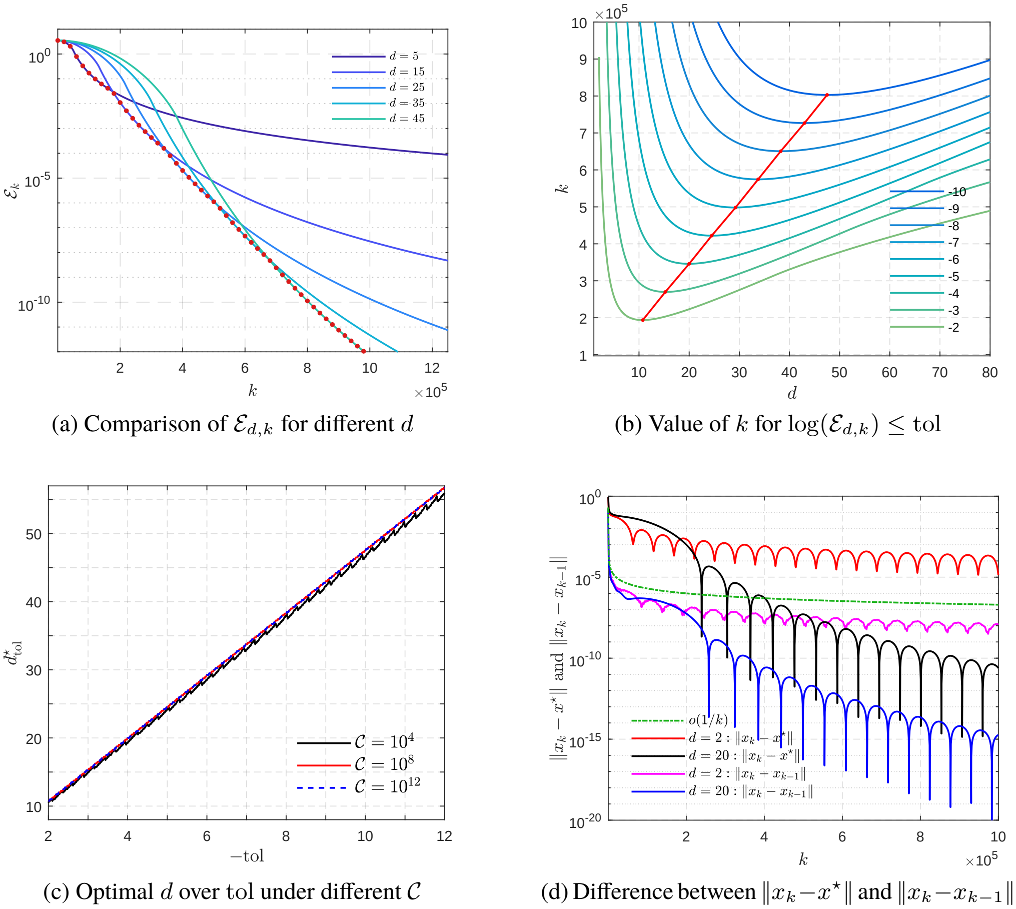
- there exists optimal choice of d;
- The optimal d depends on the accuracy of x_k.

To verify these claims, we consider the following numerical illustration: under a given $\mathrm{tol} \in \{-2, ..., -10\}$, for each $d \in [2, 100]$ compute the minimal number of iterations, *i.e.* k, needed such that

$$\log(\mathcal{E}_{d,k}) \le \mathrm{tol}.$$

The obtained results are shown in Figure 6 (b), from where we can observe that for each $\mathrm{tol} \in \{-2, ..., -10\}$, the corresponding k is a smooth curve that admits a minimal value $k^\star_{\mathrm{tol}}$ for optimal $d^\star_{\mathrm{tol}}$. The red line segment connects all the points of $(d^\star_{\mathrm{tol}}, k^\star_{\mathrm{tol}})$ which almost is a straight line. It indicates that one should

choose small d when tol *is large, and increase the value of d when* tol *is becoming smaller.*



(a) Comparison of $\mathcal{E}_{d,k}$ for different d

(b) Value of k for $\log(\mathcal{E}_{d,k}) \le \mathrm{tol}$

(c) Optimal d over tol under different $\mathcal{C}$

(d) Difference between $\|x_k - x^\star\|$ and $\|x_k - x_{k-1}\|$

Figure 6: Optimal choices of d under different stopping tolerance: (a) Comparison of $\mathcal{E}_{d,k}$ for $d \in [5, 15, 25, 35, 45]$; (b) Number of iteration needed for $\log(\mathcal{E}_{d,k}) \le \mathrm{tol}$; (c) Optimal d over tol for different condition number $\mathcal{C} \in \{10^4, 10^8, 10^{12}\}$. (d) Difference between $\|x_k - x^\star\|$ and $\|x_k - x_{k-1}\|$.

The red line in Figure 6 (b) accounts only for condition number $\mathcal{C} = 2.735 \times 10^7$. In Figure 6 (c), we consider three different condition numbers $\mathcal{C} \in \{10^4, 10^8, 10^{12}\}$ and plot their corresponding optimal choices of d under different tol. Surprisingly, the obtained optimal choices for each $\mathcal{C}$ are almost same, especially for $\mathcal{C} = 10^8, 10^{12}$. From these three lines, we fit the following linear function

$$d^\star_{\mathrm{tol}} = 10.75 + 4.6(-\mathrm{tol} - 2),$$

which can be used to compute the optimal d for a given stopping criterion on $\|x_k - x^\star\|$.

4.3.2 Optimal d for $\|x_k - x_{k-1}\|$

To this point, we have presented the detailed analysis on the advantage of lazy-start strategy. However, the analysis is conducted via the envelope $\mathcal{E}_{d,k}$ of $\|x_k - x^\star\|$ which requires the solution $x^\star$. While in practice, most of time only $\|x_k - x_{k-1}\|$ is available, which makes the above discussion on optimal d not practically useful. Therefore, we discuss briefly below how to adapt the above result to $\|x_k - x_{k-1}\|$.

In Figure 6 (d) we plot both $\|x_k - x^\star\|$ and $\|x_k - x_{k-1}\|$ for the considered problem (4.1) with $d = 2$ and $d = 20$. The red and magenta lines are for $d = 2$ while the black and blue lines are for $d = 20$. It can be observed that $\|x_k - x_{k-1}\|$ is several order smaller than $\|x_k - x^\star\|$, which is caused by the significant decay at the beginning of $\|x_k - x_{k-1}\|$. This is due to the fact that at the beginning stage of the iterates, the convergence of $\|x_k - x_{k-1}\|$ is governed by the $o(1/k)$ rate established in Theorem 3.5; see the green dot-dash line.

If we discard the $o(1/k)$ part of $\|x_k - x_{k-1}\|$, then the remainder can be seen as scaled $\|x_k - x^\star\|$, *i.e.* $\|x_k - x_{k-1}\| = \|x_k - x^\star\|/10^s$ for some $s > 0$. Therefore, if some prior about this shift could be available, then the optimal choice of d would be

$$d^\star_{\text{tol}} = 10.75 + 4.6(-\text{tol} - 2 - s).$$

For a given problem, in practice the value of s can be estimated though the following strategy:

- Run the FISTA iteration for sufficient number of iterations and obtain a rough solution $\tilde{x}$ and also record the residual sequence $\|x_k - x_{k-1}\|$;
- Rerun the iteration again and output the value of $\|x_k - \tilde{x}\|$. Comparing $\|x_k - x_{k-1}\|$ and $\|x_k - \tilde{x}\|$ one can then obtain an estimation of s.

In practice, one can also simply choose $d \in [20, 40]$ which can provide consistent faster performance.

Remark 4.6.

- The discussion of this section has been conducted through FISTA-CD, to extend the result to the case of FISTA-Mod, we may simply take $p = \frac{1}{d}$ and let $q \in]0, 1]$. As we have seen from Figure 1, the correspondence between FISTA-CD and FISTA-Mod is roughly $p = \frac{1}{d}$.
- The discussion of this section considers only the least square problem (4.1) which is very simple. However, this does not mean that lazy-start strategy will fail for more complicated problems such as ($\mathcal{P}$), see Section 7 for evidence. Moreover, owing to the result of [16], many examples of ($\mathcal{P}$) locally around the solution is equivalent to some C^2-smooth problems. As a consequence, though the least square problem is simple, it is representative enough for the discussion.

5 Adaptive acceleration

We have discussed the advantages of the proposed FISTA-Mod scheme, particularly the lazy-start strategy. However, despite the advantage brought by lazy-start, FISTA-Mod/FISTA-CD still suffer the same drawback of FISTA-BT: the oscillation of $\Phi(x_k) - \Phi(x^\star)$ and $\|x_k - x^\star\|$ as shown in Figure 2. Therefore, in this section we discuss adaptive approaches to avoid oscillation. Note that here we only discuss adaptation to strong convexity, and refer to [8] for backtracking strategies for Lipschitz constant L.

The presented acceleration schemes cover two different cases: strong convexity α is explicitly available, α is unknown or non-strongly convex. For the first case, the optimal parameter choices are available. While for the latter, we need to adaptively estimate the strong convexity, hence the parameter choices.

5.1 Strong convexity is available

For this case, we assume that F of ($\mathcal{P}$) is α-strongly convex and R is only convex, and derive the optimal setting of p, q and r for FISTA-Mod. Recall that under step-size γ, the optimal inertial parameter is $a^\star = \frac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}$. From (3.4) the limiting value of a_k, we have that for given $p, q \in]0, 1]$, r should be chosen such that

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

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

$$\begin{aligned} r = f(\alpha, \gamma; p, q) &= 4(1 - p) + 4pa^\star + (p^2 - q)(1 - a^\star)^2 \\ &= 4(1 - p) + \frac{4p(1 - \sqrt{\gamma\alpha})}{1 + \sqrt{\gamma\alpha}} + \frac{4\gamma\alpha(p^2 - q)}{(1 + \sqrt{\gamma\alpha})^2} \le 4. \end{aligned} \tag{5.1}$$

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

Based on the above result, we propose below an generalisation of FISTA-Mod which is able to adapt to the strong convexity of the problem to solve.

Algorithm 3: Strongly convex FISTA-Mod (α-**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
    $t_k = \frac{p + \sqrt{q + r t_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1} - 1}{t_k},$
    $y_k = x_k + a_k(x_k - x_{k-1}),$                                         (5.2)
    $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$
until convergence;
```

Remark 5.1.

- Since $f(\alpha, \gamma; p, q) = 4$ when $\alpha = 0$, the above algorithm mains the $o(1/k^2)$ convergence rate for non-strongly convex case, and in general we have the following convergence property for α-FISTA,
$$\Phi(x_k) - \Phi(x^\star) \leq \mathcal{C}\omega_k,$$
where $\mathcal{C} > 0$ is a constant and $\omega_k = \min\big\{\frac{2L}{p^2(k+1)^2}, (1 - \sqrt{\gamma\alpha})^k\big\}$.
- Recall Remark 3.1, the property of t_k converging to its limit $\tilde{t}$. In practice, it is better to choose $t_0 > \tilde{t}$ as it gives faster practical performance than choosing $t_0 < \tilde{t}$.

Recently, combing FISTA scheme with strong convexity was studied in [8] where the authors also propose an generalisation of FISTA scheme for strongly convex problems. They consider the case that R is α_R-strongly convex and F is α_F-strongly convex, and the whole problem is then $(\alpha = \alpha_R + \alpha_F)$-strongly convex. In [8, Algorithm 1], the following update rule of t_k is considered

$$t_k = \frac{1 - q t_{k-1}^2 + \sqrt{(1 - q t_{k-1}^2)^2 + 4 t_{k-1}^2}}{2} \quad \text{and} \quad a_k = \frac{t_{k-1} - 1}{t_k} \frac{1 + \gamma\alpha_R - t_k\gamma\alpha}{1 - \gamma\alpha_F}, \tag{5.3}$$

where $q = \frac{\gamma\alpha}{1+\gamma\alpha_R}$. As we shall see later in Section 6, the above update rule is equivalent to Nesterov's optimal scheme [23]; see also [11] for discussions.

When $\alpha > 0$, then [8, Algorithm 1] achieves $O((1-\sqrt{q})^k)$ linear convergence rate. When $\alpha_R = 0, \alpha_F > 0$, we have $1 - \sqrt{q} = 1 - \sqrt{\gamma\alpha}$ which means [8, Algorithm 1] and α-FISTA achieves the same optimal rate. However, if both $\alpha_R > 0$ and $\alpha_F \geq 0$, then

$$1 - \sqrt{\frac{\gamma\alpha}{1 + \gamma\alpha_R}} > 1 - \sqrt{\gamma\alpha},$$

which means (5.3) achieves a sub-optimal convergence rate. As a matter of fact, if we transfer the strong convexity of R to F, that is

$$R \overset{\text{def}}{=} R - \frac{\alpha_R}{2}\|x\|^2 \quad \text{and} \quad F \overset{\text{def}}{=} F + \frac{\alpha_R}{2}\|x\|^2.$$

Then R is convex and F is α-strongly convex, and the optimal rate would be $1 - \sqrt{\gamma\alpha}$. Moreover, redefining R does not affect computing $\mathrm{prox}_{\gamma R}$, as it is simply quadratic perturbation of proximity operator [12, Lemma 2.6].

5.2 Strong convexity is not available

The goal of α-FISTA is to avoid the oscillatory behaviour of the FISTA schemes. In the literature, an efficient way to deal with oscillation is the restarting technique from [25]. The basic idea of restarting is that, once the objective function value of $\Phi(x_k)$ is about to increase, the algorithm resets t_k and y_k. Doing so, the algorithm achieves an almost monotonic convergence in terms of $\Phi(x_k) - \Phi(x^\star)$, and can be significantly faster than the original scheme; see [25] or Section 7 for detailed comparisons.

The strong convexity adaptive α-FISTA (Algorithm 3) considers only the situation where the strong convexity is explicitly available, which is very often not the case in practice. Moreover, the oscillatory behaviour is independent of the strong convexity. As a consequence, an adaptive scheme is needed such that the following scenarios can be covered

- Φ is neither *globally* strongly convex nor *locally* strongly convex;
- Φ is *globally* strongly convex with unknown modulus α;
- Φ is *locally* strongly convex with unknown modulus α.

On the other hand, when Φ is strongly convex, estimating the strong convexity in general is time consuming. Therefore, an efficient estimation approach is also needed. To address these problems, we propose a restarting adaptive scheme (Algorithm 4), which combines the restarting technique of [25] and α-FISTA.

```
Algorithm 4: Restarting and Adaptive α-FISTA (Rada-FISTA)
Initial: $p, q \in ]0, 1], r = 4$ and $\xi < 1, t_0 = 1, \gamma = 1/L$ and $x_0 \in \mathcal{H}, x_{-1} = x_0$.
repeat
    • Run FISTA-Mod:
        $t_k = \frac{p + \sqrt{q + r t_{k-1}^2}}{2}, \quad a_k = \frac{t_{k-1} - 1}{t_k},$
        $y_k = x_k + a_k(x_k - x_{k-1}),$
        $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$
    • 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;
```

For the rest of the paper, we shall call Algorithm 4 as "Rada-FISTA". Below we present some discussions:

- Compared to α-FISTA, the main difference of Rada-FISTA is the restarting step which is originally proposed in [25]. Such a strategy can successfully avoid the oscillatory behaviour of $\Phi(x_k) - \Phi(x^\star)$;
- We provide two different option for the restarting step. For both options, once restarts, we reset y_k as in [25]. Meanwhile, we also rescale the value of r by a factor ξ which is strictly smaller than 1. The purpose of rescaling is to approximate the optimal choice of r in (5.1);
- The difference between the two options is that t_k is not reset to 1 in "Option I". Doing so, "Option I" will restart for more times than "Option II", however it will achieve faster practical performance; see Section 7 the numerical experiments. It is worth noting that, for the restarting FISTA of [25], removing resetting t_k could also lead to an acceleration.

We provide a very simple way on how to choose parameter ξ: let k be the iteration number when the criterion $(y_k - x_{k+1})^T (x_{k+1} - x_k) \geq 0$ is triggered for the first time, we then have the corresponding a_k, let $m > 1$ be some large enough constant, then one can simply set $\xi = \sqrt[m]{a_k}$.

5.3 Greedy FISTA

We conclude this section by discussing how to further improve the performance of restarting technique, achieving an even faster performance than Rada-FISTA and restarting FISTA [25].

The oscillation of FISTA schemes is caused by the fact that $a_k \to 1$. For the restarting scheme [25], resetting t_k to 1 forces a_k: increase from 0 again, become close enough to 1 and cause next oscillation, then the scheme restarts. For such loop, if we can shorten the gap between two restarts, then extra acceleration can be obtained. It turns out that using constant a_k (close or equal to 1) can achieve this goal. Therefore, we propose the following restarting scheme.

```
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:
        $y_k = x_k + (x_k - x_{k-1}),$
        $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$        (5.4)
    • 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;
```

We abuse the notation by calling the above algorithm "Greedy FISTA", which uses constant inertial parameter $a_k \equiv 1$ for the momentum term:

- A larger step-size (than $1/L$) is chosen for γ, which can further shorten the oscillation period;
- As such large step-size may lead to divergence, we add a "safeguard" step to ensure the convergence. This step shrinkages the value of γ when certain condition (*e.g.* $\|x_{k+1} - x_k\| \geq S\|x_1 - x_0\|$) is satisfied. Eventually we will have $\gamma = 1/L$ is safeguard is triggered for sufficient number of steps, and the convergence of the objective function value $\Phi(x_k)$ can be guaranteed.

In practice, we find that $\gamma \in [1/L, 1.3/L]$ provides faster performance than Rada-FISTA and restarting FISTA of [25]; See Section 7 for more detailed comparisons.

6 Nesterov's accelerated scheme

In this section, we turn to Nesterov's accelerated gradient method [23] and extend the above results to this algorithm. In the book [23], Nesterov introduces several different acceleration schemes, in the following we mainly focus on the "Constant Step Scheme, III". Applying this scheme to solve ($\mathcal{P}$), we obtain the following accelerated proximal gradient method (APG).

```
Algorithm 6: Accelerated proximal gradient (APG)
Initial: $\tau \in [0, 1], \theta_0 = 1, \gamma = 1/L$ and $x_0 \in \mathcal{H}, x_{-1} = x_0$.
repeat
    Estimate the local strong convexity $\alpha_k$;
        $\theta_k$ solves $\theta_k^2 = (1 - \theta_k)\theta_{k-1}^2 + \tau\theta_k,$
        $a_k = \frac{\theta_{k-1}(1 - \theta_{k-1})}{\theta_{k-1}^2 + \theta_k},$
        $y_k = x_k + a_k(x_k - x_{k-1}),$
        $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma \nabla F(y_k)\big).$
until convergence;
```

When the problem ($\mathcal{P}$) is α-strongly convex, then by setting $\tau = \sqrt{\alpha/L}$ and $\theta_0 = \tau$, we have

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

and the iterate achieves the optimal convergence speed, *i.e.* $1 - \sqrt{\gamma\alpha}$, as we have already discussed in the previous sections. In the rest of this section, we first build connections between the parameters update of APG with α-FISTA, and then extend the lazy-start strategy to APG.

6.1 Connection with α-FISTA

Consider the following equation of θ parametrised by $0 \leq \tau \leq \sigma \leq 1$, which recovers the θ_k update of APG for $\sigma = 1$,

$$\theta^2 + (\sigma\theta_{k-1}^2 - \tau)\theta - \theta_{k-1}^2 = 0. \tag{6.1}$$

The definition of a_k implies $\theta_k \in [0, 1]$ for all $k \geq 1$. Therefore, the θ_k we seek from above equation (6.1) reads

$$\theta_k = \frac{-(\sigma\theta_{k-1}^2 - \tau) + \sqrt{(\sigma\theta_{k-1}^2 - \tau)^2 + 4\theta_{k-1}^2}}{2}. \tag{6.2}$$

It is then easy to verify that θ_k is convergent and $\lim_{k\to+\infty} \theta_k = \sqrt{\frac{\tau}{\sigma}}$. Back to (6.2), we have

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

Letting $t_k = 1/\theta_k$ and substituting back to the above equation lead to

$$t_k = \frac{(\sigma - \tau t_{k-1}^2) + \sqrt{(\sigma - \tau t_{k-1}^2)^2 + 4t_{k-1}^2}}{2}. \tag{6.3}$$

Note that the update rule (5.3) of [8] is a special case of above equation with $\sigma = 1$ and $\tau = \frac{\gamma\alpha}{1+\gamma\alpha_R}$. Moreover,

$$t_k \to \begin{cases} +\infty : \tau = 0, \\ \sqrt{\frac{\sigma}{\tau}} : \tau \in]0, 1]. \end{cases}$$

Depending on the choices of σ, τ, we have

- When $(\sigma, \tau) = (1, 0)$, APG is equivalent to the original FISTA-BT scheme;
- When $(\sigma, \tau) = (1, \gamma\alpha)$, APG is equivalent to [8, Algorithm 1] for adapting to strong convexity.

Building upon the above connection, we can extend the previous result of FISTA-Mod to the case of APG.

Remark 6.1. Let $\tau = 0$ in (6.3), comparing with the t_k update in (3.5), we have that (6.3) is a special case of (3.5) with $p = \sigma$ and $q = \sigma^2$.

6.2 A modified APG

Extending the FISTA-Mod (Algorithm 2) and α-FISTA (Algorithm 3) to the case of APG, we propose the following modified APG scheme which we name as "mAPG".

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 $x_0 \in \mathcal{H}, x_{-1} = x_0$.
repeat
    $\theta_k$ solves $\theta_k^2 = (1 - \sigma\theta_k)\theta_{k-1}^2 + \tau\theta_k$,
    $a_k = \frac{\theta_{k-1}(1 - \theta_{k-1})}{\theta_{k-1}^2 + \theta_k}$,                    (6.4)
    $y_k = x_k + a_k(x_k - x_{k-1})$,
    $x_{k+1} = \mathrm{prox}_{\gamma R}\big(y_k - \gamma\nabla F(y_k)\big)$.
until convergence;
```

Non-strongly convex case For the case Φ is only convex, we have $\tau = 0$, then θ_k is the root of the equation

$$\theta^2 + \sigma\theta_{k-1}^2\theta - \theta_{k-1}^2 = 0.$$

Owing to Section 6.1, we have that mAPG is equivalent to FISTA-Mod with $p = \sigma$ and $q = \sigma^2$. Therefore, we have the following convergence result for mAPG which is an extension of Theorems 3.3 and 3.5.

Corollary 6.2. *For mAPG scheme Algorithm 7, let* $\tau = 0$ *and* $\sigma \in]0, 1]$, *then*

- *For the objective function value,*

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

  If moreover $\sigma < 1$, we have $\Phi(x_k) - \Phi(x^\star) = o(1/k^2)$.
- *Let $\sigma < 1$, then there exists an $x^\star \in \mathrm{Argmin}(\Phi)$ to which the sequence $\{x_k\}_{k\in\mathbb{N}}$ converges weakly. Moreover, $\|x_k - x_{k-1}\| = o(1/k)$.*

Remark 6.3. We can also design a lazy-start strategy for mAPG. Given the correspondence between σ of mAPG and p of FISTA-Mod, owing to Proposition 4.1, we obtain the lazy-start mAPG by choosing $\sigma \in [\frac{1}{80}, \frac{1}{10}]$.

Strongly convex case When the problem ($\mathcal{P}$) is strongly convex with modulus $\alpha > 0$, as $\tau = \gamma\alpha\sigma$, then according to Section 6.1, we have

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

which means that mAPG achieves the optimal convergence rate $1 - \sqrt{\gamma\alpha}$.

Remark 6.4. We can also extend the Rada-FISTA to APG, as it is quite trivial, we shall forgo the details here.

7 Numerical experiments

Now we present numerical experiments of problems arising from inverse problems, signal/image processing, machine learning and computer vision to demonstrate the performance of the proposed schemes. Throughout the section, the following schemes and corresponding settings are compared:

- The original FISTA-BT scheme [6];
- The proposed FISTA-Mod (Algorithm 2) with $p = 1/20$ and $q = 1/2$, *i.e.* the lazy-start strategy;
- The restarting FISTA of [25];
- The Rada-FISTA scheme (Algorithm 4);
- The greedy FISTA (Algorithm 5) with $\gamma = 1.3/L$ and $S = 1, \xi = 0.96$.

The α-FISTA (Algorithm 3) is not considered here, except in Section 7.1, since most of the problems considered are only locally strong convex along certain set [16]. The corresponding MATLAB source code for reproducing the experiments is available at: `https://github.com/jliang993/Faster-FISTA`.

All the schemes are running with same initial point, which is $x_0 = \mathbf{1} \times 10^4$ for the least square problem and $x_0 = \mathbf{0}$ for all other problems. In terms of comparison criterion, we mainly focus on $\|x_k - x^\star\|$ where $x^\star \in \mathrm{Argmin}(\Phi)$ is a global minimiser of the optimisation problem.

7.1 Least square (4.1) continue

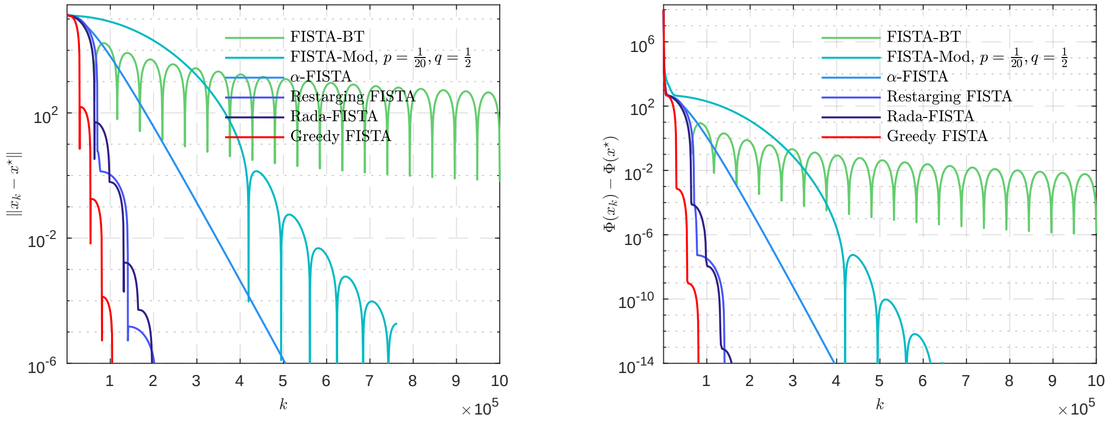
First we continue with the least square estimation (4.1) discussed in Section 4, and present a comparison of different schemes in terms of both $\|x_k - x^\star\|$ and $\Phi(x_k) - \Phi(x^\star)$. Since this problem is strongly convex, the optimal scheme (*i.e.* α-FISTA) is also considered for comparison.

The obtained results are shown in Figure 7, with $\|x_k - x^\star\|$ on the left and $\Phi(x_k) - \Phi(x^\star)$ on the right. From these comparisons, we obtain the following observations:

- FISTA-BT is faster than FISTA-Mod for $k \leq 3 \times 10^5$, and becoming increasing slower afterwards. This agrees with our previous discussion in Figure 6 that each parameter choice (of p and q, and d for FISTA-CD) is the fastest for certain accuracy;
- α-FISTA is the only scheme whose performance is monotonic in terms of both $\|x_k - x^\star\|$ and $\Phi(x_k) - \Phi(x^\star)$. It is also faster than both FISTA-BT and FISTA-Mod;
- The three restarting adaptive schemes are the fastest among tested schemes, with greedy FISTA being faster than the other two.

7.2 Linear inverse problem and regression problems

From now on, we turn to dealing with problems that are only locally strongly convex around the solution of the problem. We refer to [16] for a detailed characterisation of such local neighbourhood.



(a) $\|x_k - x^\star\|$ (b) $\Phi(x_k) - \Phi(x^\star)$

Figure 7: Comparison of different FISTA schemes for least square problem (4.1). (a): convergence of $\|x_k - x^\star\|$; (b) convergence of $\Phi(x_k) - \Phi(x^\star)$.

Linear inverse problem Consider the following regularised least square problem

$$\min_{x\in\mathbb{R}^n} \mu R(x) + \frac{1}{2}\|\mathcal{K}x - f\|^2, \tag{7.1}$$

where $\mu > 0$ is the trade-off parameter, R is the regularisation term. The forward model of (7.1) reads

$$f = \mathcal{K}x_{\text{ob}} + w, \tag{7.2}$$

where $x_{\text{ob}} \in \mathbb{R}^n$ is the original object that obeys certain prior (*e.g.* sparsity and piece-wise constant), $f \in \mathbb{R}^m$ is the observation, $\mathcal{K} : \mathbb{R}^n \to \mathbb{R}^m$ is some linear operator, and $w \in \mathbb{R}^m$ stands for noise. In the experiments, we consider R being ℓ_∞-norm and total variation [31]. Here $\mathcal{K}$ is generated from the standard Gaussian ensemble and the following parameters:

ℓ_∞-norm	$(m, n) = (1020, 1024)$, x_{ob} has 32 saturated entries;
Total variation	$(m, n) = (256, 1024)$, ∇x_{ob} is 32-sparse.

Sparse logistic regression A sparse logistic regression problem for binary classification is also considered. Let $(h_i, l_i) \in \mathbb{R}^n \times \{\pm 1\}$, $i = 1, \cdots, m$ be the training set, where $h_i \in \mathbb{R}^n$ is the feature vector of each data sample, and l_i is the binary label. The formulation of sparse logistic regression reads

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \log\big(1 + e^{-l_i h_i^T x}\big), \tag{7.3}$$

where $\mu > 0$ and is set to be 10^{-2} in the numerical test. The `australian` data set from LIBSVM[2] is considered.

The observation are shown in Figure 8. Though these problems are only locally strongly convex around the solution, the observations are quite close to those of least square problem discussed above:

- The lazy-start FISTA-Mod is slower than FISTA-BT at the beginning, and eventually becomes much faster. For ℓ_∞-norm, it is more than 10 times faster if we need the precision to be $\|x_k - x^\star\| \le 10^{-10}$;
- The restarting adaptive schemes are the fastest ones, and the greedy FISTA is the fastest of all.

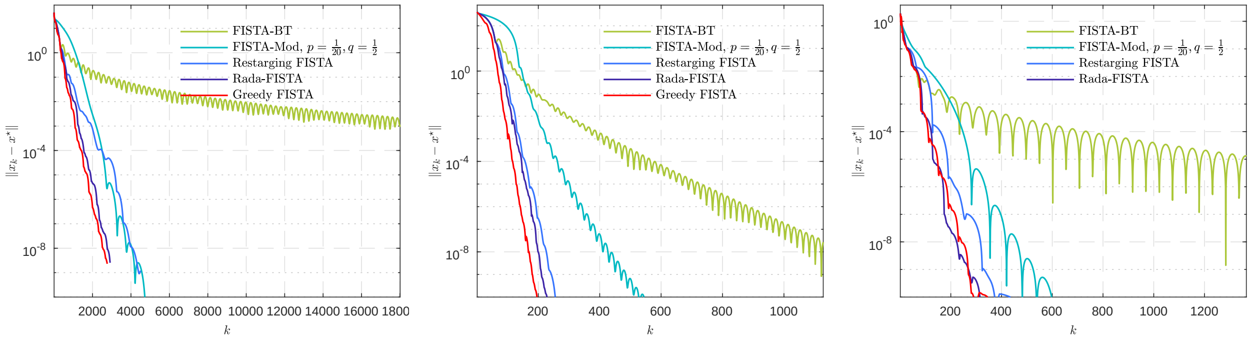
7.3 Principal component pursuit

Lastly, we consider the principal component pursuit (PCP) problem [9], and apply it to decompose a video sequence into background and foreground.

Assume that a real matrix $f \in \mathbb{R}^{m\times n}$ can be written as

$$f = x_{\text{l,ob}} + x_{\text{s,ob}} + w,$$

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) ℓ_∞-norm (b) Total variation (c) Sparse logistic regression

Figure 8: Comparison of different FISTA schemes for linear inverse problems and sparse logistic regression. (a) ℓ_∞-norm; (b) Total variation; (c) Sparse logistic regression.

where $x_{\mathrm{l,ob}}$ is low–rank, $x_{\mathrm{s,ob}}$ is sparse and w is the noise. The PCP proposed in [9] attempts to recover $(x_{\mathrm{l,ob}}, x_{\mathrm{s,ob}})$ to a good approximation, by solving the following convex optimization problem

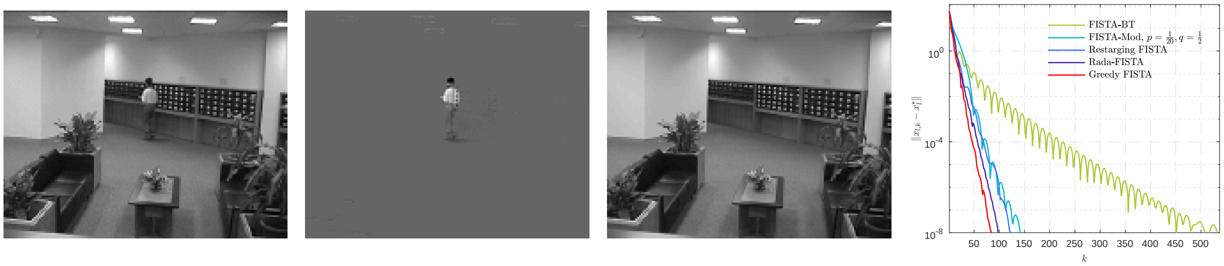
$$\min_{x_{\mathrm{l}}, x_{\mathrm{s}} \in \mathbb{R}^{m\times n}} \ \frac{1}{2}\|f - x_{\mathrm{l}} - x_{\mathrm{s}}\|_F^2 + \mu\|x_{\mathrm{s}}\|_1 + \nu\|x_{\mathrm{l}}\|_*, \tag{7.4}$$

where $\|\cdot\|_F$ is the Frobenius norm. Observe that for fixed x_{l}, the minimizer of (7.4) is $x_{\mathrm{s}}^\star = \mathrm{prox}_{\mu\|\cdot\|_1}(f - x_{\mathrm{l}})$. Thus, (7.4) is equivalent to

$$\min_{x_{\mathrm{l}} \in \mathbb{R}^{m\times n}} \ {}^1\big(\mu\|\cdot\|_1\big)(f - x_{\mathrm{l}}) + \nu\|x_{\mathrm{l}}\|_*, \tag{7.5}$$

where ${}^1\big(\mu\|\cdot\|_1\big)(f - x_{\mathrm{l}}) = \min_z \frac{1}{2}\|f - x_{\mathrm{l}} - z\|_F^2 + \mu\|z\|_1$ is the Moreau Envelope of $\mu\|\cdot\|_1$ of index 1, and hence has 1-Lipschitz continuous gradient.

We use the video sequence from [13] and the obtained result is demonstrated in Figure 9. Again, we obtain very consistent observations with the above examples. Moreover, the performance of lazy-start FISTA-Mod is very close to the restarting adaptive schemes.



(a) Original frame (b) Sparse component (c) Low-rank component (d) Performance comparison

Figure 9: Comparison of different FISTA schemes for principal component pursuit problem. (a) original frame; (b) foreground; (c) background; (d) performance comparison.

References

[1] H. Attouch, A. Cabot, Z. Chbani, and H. Riahi. Inertial forward–backward algorithms with perturbations: Application to tikhonov regularization. *Journal of Optimization Theory and Applications*, 179(1):1–36, 2018.

[2] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated Forward–Backward method is actually $o(k^{-2})$. Technical Report arXiv:1510.08740, 2015.

[3] H. Attouch, J. Peypouquet, and P. Redont. On the fast convergence of an inertial gradient-like dynamics with vanishing viscosity. Technical Report arXiv:1507.04782, 2015.

[4] J. B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornés etn-cycliquement monotones. *Israel Journal of Mathematics*, 26(2):137–150, 1977.

[5] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[6] 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.

[7] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[8] L. Calatroni and A. Chambolle. Backtracking strategies for accelerated descent methods with smooth composite objectives. *arXiv preprint arXiv:1709.09004*, 2017.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[10] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[11] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[12] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[13] L. Li, W. Huang, I. Y. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[14] J. Liang. *Convergence rates of first-order operator splitting methods*. PhD thesis, Normandie Université; GREYC CNRS UMR 6072, 2016.

[15] J. Liang, J. Fadili, and G. Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1-2):403–434, 2016.

[16] J. Liang, J. Fadili, and G. Peyré. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

[17] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[18] D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.

[19] C. Molinari, J. Liang, and J. Fadili. Convergence rates of forward–douglas–rachford splitting method. *arXiv preprint arXiv:1801.01088*, 2018.

[20] A. Moudafi and M. Oliny. Convergence of a splitting inertial proximal method for monotone operators. *Journal of Computational and Applied Mathematics*, 155(2):447–454, 2003.

[21] A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[22] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[23] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[24] Y. Nesterov. Gradient methods for minimizing composite objective function. 2007.

[25] B. O'Donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[26] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

[27] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[28] B. T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

[29] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[30] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[31] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.