A Discussion on The Nesterov Momentum and Variants of FISTA with TV Minmizations Application

HONGDA LI

December 8, 2023

Abstract

In this paper, we review the paper written by Walkington [23] on the topic of proximal gradient with Nesterov accelerations. We compare the performance of the FISTA method and some of its variants with numerical experiments on the total variation minimization problem; in addition, we propose a heuristic estimation of the strong convexity parameter and demonstrate that it converges faster when applied. We use the Julia programming language [11] for the numerical implementation; see the GitHub repository here. We give a literature review on the frontier for both the theories and applications around FISTA. We correct one misconception in Walkington [23] regarding Nesterov's proof of lower bound on the optimality of first-order algorithms. We present a better proof of linear convergence of FISTA under strong convexity assumption from Beck [8, theorem 10.7.7] by eliminating an identity used in their proof.

1 Preliminaries

In this section, We initiate the discussion by denoising a one-dimensional signal. The dual objective function of the problem derived in this section motivates the use of Accelerated Proximal Gradient with smooth, non-smooth composite objective function. We summarized it from Walkington [23], sections 1 and 4.

1.1 Signal Denoising in One Dimension

Let a one dimensional singal be $u : [0, 1] \mapsto \mathbb{R}$ and u. Regularizing the derivative of the signal using the L1 norm helps recover the digital signal if it's a piecewise constant function. This is called a Total Variation (TV) minimization. Let $\hat{u}$ denote an observation of u corrupted by noise. The denoised signal is the minimizer of f(u),

$$f(u) = \int_0^1 \frac{1}{2}\left(u(t) - \hat{u}(t)\right)^2 + \alpha|u'(t)|dt.$$

Practical implementations for modern computing devices would necessitate discretization of the integral.

We use the trapezoidal rule and second-order forward difference for the derivative. Let $\hat{u} \in \mathbb{R}^{N+1}$, a vector in the form of $\hat{u} = [\hat{u}_0 \; \cdots \; \hat{u}_N]$, let $t_0 < \cdots < t_N$ be a sequence of time corresponded to each observation of $\hat{u}_i$. The time intervals are $h_i = t_i - t_{i-1}$ for $i = 1, \cdots, N$, not necessarily equally spaced, making this formulation below is slightly more general than Walkington[23]. We derive the approximation of the integral. Denote $s_i = u_i - \hat{u}_i$. Let $C \in \mathbb{R}^{N\times(N+1)}$ be upper bi-diagonal with $(1/h_i, -1/h_i)$, and $D \in \mathbb{R}^{N\times(N+1)}$ be $\text{diag}(h_1/2, h_1, h_2, \cdots, h_N, h_N/2)$, then

$$\begin{aligned}\frac{1}{2}\int_0^1 (u-\hat{u})^2 dt + \alpha\int_0^1 |u'|dt &\approx \frac{1}{2}\sum_{i=0}^{N}\left(\frac{s_i^2+s_{i+1}^2}{2}\right)h_{i+1} + \alpha\sum_{i=1}^{N}\left|\frac{u_i-u_{i-1}}{h_i}\right| \\ &= \frac{1}{2}\left(\frac{s_0^2 h_1}{2} + \frac{s_N^2 h_N}{2} + \sum_{i=1}^{N-1} s_i^2 h_i\right) + \alpha\|Cu\|_1 \\ &= \frac{1}{2}\langle u-\hat{u}, D(u-\hat{u})\rangle + \alpha\|Cu\|_1.\end{aligned}$$

The above formulation suggests a smooth, non-smooth additive composite objective for $f(u)$. The Proximal Gradient method and its variants can solve this optimization problem. Unfortunately, the non-smooth part $\alpha\|Cu\|_1$ presents computational difficulty if matrix C is unfriendly for the prox operator. One way to bypass the difficulty involves reformulating with $p = Cu$ and solving the dual problem.

Dual Reformulation

Let $p = Cu$, $C \in \mathbb{R}^{(N+1)\times N}$ with $D \in \mathbb{R}^{(N+1)\times(N+1)}$. We reformulate it into

$$\min_{u\in\mathbb{R}^{N+1}}\left\{\underbrace{\frac{1}{2}\langle(u-\hat{u}), D(u-\hat{u})\rangle}_{f(u)} + \underbrace{\alpha\|p\|_1}_{h(p)}\ \middle|\ p = Cu\right\},$$

producing Lagrangian of the form

$$\mathcal{L}((u,p),\lambda) = f(u) + h(p) + \langle\lambda, p - Cu\rangle.$$

The dual is

$$\begin{aligned}-g(-\lambda) &:= \inf_{(u,p)\in\mathbb{R}^{N+1}\times\mathbb{R}^N}\{\mathcal{L}((u,p),\lambda)\} \\ &= \inf_{(u,p)\in\mathbb{R}^{N+1}\times\mathbb{R}^N}\{f(u) + h(p) + \langle\lambda, p - Cu\rangle\} \\ &= \inf_{u\in\mathbb{R}^{N+1}}\left\{f(u) - \langle\lambda, Cu\rangle + \inf_{p\in\mathbb{R}^N}\{h(p) + \langle\lambda, p\rangle\}\right\} \\ &= -f^\star(-C^T\lambda) - h^\star(-\lambda).\end{aligned}$$

With the assumption that D is positive definite, we have

$$\begin{aligned}-g(-\lambda) &= -\frac{1}{2}\|C^T\lambda\|^2_{D^{-1}} - \langle\hat{u}, -C^T\lambda\rangle - \delta_{[-\alpha,\alpha]^N}(-\lambda) \\ g(\lambda) &= \frac{1}{2}\|C^T\lambda\|^2_{D^{-1}} - \langle C^T\lambda, u\rangle + \delta_{[-\alpha,\alpha]^N}(\lambda).\end{aligned}$$

Observe that the above admits a hyperbox indicator function that makes the prox operator friendlier because the proximal operator of the indicator is projection; in the case of projecting onto the box, the operator is simple. Given dual variable λ, primal variables are obtained by

$$\begin{aligned}(u^+, p^+) &= \text{argmin}_{(u,p)}\mathcal{L}((u,p),\lambda) \\ \partial_u\mathcal{L}((u,p),\lambda)|_{u=u^+} &= D(u^+ - \hat{u}) - C^T\lambda = \mathbf{0} \\ \implies u^+ &= \hat{u} + D^{-1}C^T\lambda.\end{aligned}$$

$-g(\lambda)$ is easier to optimize, and obtaining the primal solution is also simple since D^{-1} is a diagonal matrix. Strong duality asserts the above relations between the primal and dual variables, and their objectives would be the same.

1.2 FISTA has Worse Convergence Guarantee for Strongly Convex Objectives

The dual objective for a total variation minimization problem is a strongly convex and Lipschitz smooth function because of the norm induced by the positive definite matrix D^{-1}. It's in a form where FISTA proposed by [10] can solve with a convergence rate of $\mathcal{O}(1/k^2)$ on the objective value of the function. However, highlighted in Walkington[23], the proximal gradient method without acceleration achieves $\mathcal{O}\left((1-1/\kappa)^k\right)$ convergence rate. Which is faster. The parameter κ is the condition number; in this case, it would be L/α, where L is the Lipschitz smooth constant of $g(u)$ and α is the strong convexity constant for $g(u)$.

We emphasize that the Proximal Gradient without acceleration has better theoretical convergence results than the accelerated version for the class of strongly convex objectives. This paper sparks the discussion on the variants of FISTA, hoping to provide some insights on why the inability of Nesterov's momentum-based method to adapt the convergence rate with objective has strong convexity. For the terminologies, we use FISTA to refer to the proximal gradient method presented by Beck and Teboulle[10]. We use the Accelerated Proximal Gradient method (APG) to refer to the first-order acceleration algorithms developed/inspired by FISTA.

Finally, whether the original FISTA[9] or Nesterov Accelerated gradient from 1983 has linear convergence with the presence of strong convexity (or potentially other weaker conditions) is not known during our research and literature review.

1.3 Outline of the Paper

Section 2 consists of 3 parts. The first part reviews the literature on the problem of Total Variation (TV) minimization for image/signal denoising and deblurring. The second part presents FISTA and its variants. The third part reviews the algorithmic tricks and improvements applied to the APG. Section 3 addresses a mistake made in Walkington's writing [23, theorem 2.4]. We will discuss a first-order method and how a different function achieves the lower complexity bound on the objective value and iterates for a fixed iteration. We discuss how omitting the details of this theorem creates potential misconceptions of other frontier research ideas. Section 4 presents a proof that I adapted from Amir Beck's writing [8, theorem 10.7.7]. The proof is slightly more general, removing one equality to strengthen interpretability and generality. Section 5 presents plots of convergence and results of applying variants of APG to the TV problem.

2 Literatures Review

2.1 Total Variation Minimizations

Rudin-Osher and Fatemi introduced the Total Variation (TV) minimization method in [24]. They pioneer the theories of TV minimization by solving PDE. They discussed the empirical observation that the L1 regularization term produces sharper images. Walkington [23] gives a basic formulation of one-dimensional signal denoising. However, it's essential to keep in mind that this is a problem that motivates a variety of modern computational methods and theories. We will list some of them for context.

Goldstein et al. in [17, 3.2.1] showcased the dual reformulation of a 2D signal recovery with $\|\nabla u\|$ as the regularizations term. We note that this norm is without the squared. A more hardcore, detailed coverage of reformulating the dual with L1 penalty terms for 2D signal recovery is in [9]. For a complete survey of the state of arts computational methods applied to TV minimizations, see Chambolle [15]. For a detailed exposition of mathematical theories regarding variational analysis on different types of TV problems and statistical inferences-based interpretations of the TV regularization term, consult the work by Chambolle et al.[13]. For frontier work of applying non-convex penalty term and its theoretical guarantee consult [4], [3].

Variants of FISTA

Different variants of FISTA differ by the sequence involved for their momentum method. Choosing different parameters in algorithm 1 produces variants of FISTA.

Algorithm 1 Generic FISTA

```
Input: $(g, h, x^{(0)})$
$y^{(0)} = x^{(0)}$, $\kappa = L/\sigma$
for $k = 0, 1, \cdots$ do
    $x^{(k+1)} = T_L y^{(k)}$
    $y^{(k+1)} = x^{(k+1)} + \theta_{k+1}(x^{(k+1)} - x^{(k)})$
    Execute subroutine $\mathcal{S}$.
end for
```

The scope of algorithm 1 considers the additive composition of convex smooth and nonsmooth $F = g + h$ function with g being a L-smooth function. Changing T_L, θ_{k+1} and $\mathcal{S}$, produce different variants of FISTA.

(1.) Original FISTA proposed by Beck [10] considers $\theta_{k+1} = (t_k - 1)/t_{k+1}$, $t_{k+1}(t_{k+1} - 1) = t_k^2$, with $T_L x = \text{prox}_{L^{-1}h}(x - L^{-1}\nabla g(x))$ and $t_0 = 1$. It achieves $\mathcal{O}(1/k^2)$ on the objective value. Our literature review didn't discover proof for the convergence of the iterates. We also didn't find proofs for a convergence rate faster than $\mathcal{O}((1 - 1/\kappa)^k)$ under strong convexity.

(2.) This is a variant where $t_{k+1} = (n + a - 1)/a, \theta_{k+1} = (t_k - 1)/(t_{k+1})$, for $a > 2$. $\mathcal{T}_L$ is the same as (1.). In [14, Theorem 4.1] Chambolle, Dossal proved the iterates exhibit convergence. T_L is the same as (1.).

(3.) Using $\theta_{k+1} = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ where $\kappa = L/\sigma$, with σ being the strong convexity constant produces V-FISTA in Beck [8, 10.7.7][23, 3.3]. T_L is the same as (1.).

(4.) We proposed A modification which is based on (3.), but it estimates σ, the strong convexity constant based $x^{(k)}, x^{(k+1)}, \nabla g(x^{(k+1)}), g(x^{(k)})$ using

$$\sigma \le \langle \nabla g(y^{(k+1)}) - \nabla g(y^{(k)}), y^{(k+1)} - y^{(k)} \rangle / \|y^{(k+1)} - y^{(k)}\|^2.$$

It yields excellent results in our numerical experiments. See Section 5.

(5.) MFISTA in Beck[9] is produced by adding $\mathcal{S}$ to be the procedure

$$(y^{(k+1)}, t_{k+1}) = \begin{cases} (x^{(k+1)}, 1) & F(y^{(k+1)}) > F(x^{(k+1)}), \\ (y^{(k+1)}, t_{k+1}) & \text{else.} \end{cases}$$

This condition asserts a monotone decrease in the objective value. If the objective with momentum increases, it resets the momentum for the next iteration. It has a convergence rate of $\mathcal{O}(1/k^2)$ back then. Our review of the literature didn't confirm the existence of a proof where it has a faster convergence than $\mathcal{O}((1-1/\kappa)^k)$ under the presence of strong convexity.

For our problem posed in section 1, the V-FISTA algorithm variant (3.), when applied to the dual objective, achieves a convergence rate of $\mathcal{O}((1-1/\sqrt{\kappa})^k)$ for the function objective. For larger κ, this produces a significantly better convergence rate than gradient descent, which is $\mathcal{O}((1-1/\kappa)^k)$, and FISTA, which is $\mathcal{O}(1/k^2)$.

Variants (3.) is simple; unfortunately, obtaining σ in itself could be prohibitively expensive and require knowledge about the inverse of Hessian (in the case of our TV Minimization problem). Underestimation of σ slows down the convergence rate. This observation sparks research interest in a method that achieves approximately $\mathcal{O}((1-1/\sqrt{k})^k)$ linear convergence rate for strongly convex objectives and still retains $\mathcal{O}(1/k^2)$ for Lipschitz-smooth functions in general. One of the other interests is designing a unified theoretical framework to describe all variants of APG.

To address the first issue, people use the idea of restarting FISTA. Earlier attempts proved an asymptotic fast linear convergence rate under quadratic growth conditions by triggering the restart of FISTA based on the gradient mapping norm. See [2][16]. Aujol et al.[7] proposed an automatic restart algorithm that achieves fast linear convergence without knowing the prior strong convexity (or weaker quadratic growth condition) parameter σ. Their bound is not asymptotic. Later, they developed the idea into a parameter-free algorithm in their work [6]. The interests gather around restarting FISTA because spending too much computational effort would be competing against the Proximal Quasi-Newton method, questioning the use of momentum in the first place.

On the theoretical side, Su et al. [25] identifies a second-order differential equation with the exact limit of (1.). A dynamical system understanding of FISTA and APG, in general, enables a wider variety of mathematical tools. For example, in Attouch and Peypouquet [5], they showed a $o(1/k^2)$ convergence rate of variant (2.) based on the ODE understanding. We emphasize that it's the little-o and not the big-O. In Nesterov[21], he proposed a generic algorithm that can derive variants (1.), (3.), and more using the idea of Estimating Sequences and Functions. He constructed a proof of convergence on his generic algorithm, demonstrating both linear and sub-linear convergence rates (depending on the parameter) on the functions' objective value without assuming the minimizers' existence. For Nesterov's involvement in proving convergence of APG in the non-convex settings, consult [19]. For a theoretical underpinning of Nesterov's generic method in his book, consult Ahn and Sra[1]. They derived a lot of variants of APGs using the Proximal Point method of Rockafellar and discussed the unified theme of a "similar triangle" behind the Nesterov APG method.

Finally, the most recent hardcore idea in theory and practice is from Jang et al. [18]. They squeezed out a constant from the convergence rate of FISTA by formulating the search for a faster algorithm as a QCQP. They call their algorithm OptISTA. Their approach is based on the performance estimation problem (PEP). They showed that their optimality is exact, different from Nesterov's complexity lower bound claim.

3 Nesterov's Lower Bound Clarified

Nesterov discussed his claim of the lower convergence rate for the first-order method on differentiable function in his book[22]. Walkington[23] rephrased his work with one crucial mistake in understanding Nesterov's claim.

A precise understanding is required to prevent confusion and lack of forethought in further research. We detail Nesterov's claim and provide context for understanding the mistakes in Walkington. We comment on how the claim relates to other works at the end of the section. We use the following notations

1. Let $\mathcal{A}_f^k x^{(0)}$ denotes the solution of the k-th iterate $x^{(k)}$ generated by an algorithm $\mathcal{A} \in \mathrm{GA}^{1\text{st}}$, with initial guess $x^{(0)}$, objective function f. With this notation, $x^{(1)} = \mathcal{A}_f x^{(0)}$ and $(\mathcal{A}_f^k x^{(0)})_{k\in\mathbb{N}}$ denotes the sequence generated by $\mathcal{A} \in \mathrm{GA}^{1\text{st}}$, with f.

2. Let $\mathcal{F}_L^{1,1}$ denote the set of convex functions $f : \mathbb{R}^n \mapsto R$ having L-Lipschitz gradient.

3.1 First-order Method

The following is rephrased from Assumption [21, 2.1.4]. We came up with the two examples to illustrate the definition for better understanding.

Definition 1 (First Order Method)**.** We are in $\mathbb{R}^n$ for now. We define $\mathrm{GA}^{1\text{st}}$ to be a set of all first-order algorithms. Given $x^{(0)} \in \mathbb{R}^n$, an iterative algorithm $\mathcal{A} \in \mathrm{GA}^{1\text{st}}$ generates sequence of $x^{(1)}, x^{(2)}, \cdots, x^{(k)}$ in the space satisfying:

$$x^{(j+1)} := \mathcal{A}_f^{j+1} x^{(0)} \in \left\{x^{(0)}\right\} + \operatorname{span}\left\{\nabla f\left(x^{(i)}\right)\right\}_{i=1}^{j} \quad \forall f, \forall 1 \leq j \leq k-1.$$

Example 3.1 (Fixed Step Descent)**.** The method of fixed-step gradient descent, $x^{(k+1)} = x^{(k)} - L^{-1}\nabla f(x^{(k)})$ is $\bar{\mathcal{A}} \in \mathrm{GA}^{1\text{st}}$ achieves a maximal decrease in objective value for all $f \in \mathcal{F}_L^{1,1}$ given $x^{(k)}$, it can be understood as

$$\bar{\mathcal{A}} \in \underset{\mathcal{A}\in\mathrm{GA}^{1\text{st}}}{\operatorname{argmin}} \max_{f\in\mathcal{F}_L^{1,1}} \left\{f\left(\mathcal{A}_f x^{(k)}\right)\right\}.$$

This method is memoryless because it only matters what $x^{(k)}$, prior iterate $x^{(i)}, 1 \leq i \leq k-1$ plays no role.

Example 3.2 (Steepest Descent)**.** Fix some $f, x^{(k)}$, the method of steepest descent would be $\bar{\mathcal{A}} \in \mathrm{GA}^{1\text{st}}$ and it's

$$\bar{\mathcal{A}} \in \underset{\mathcal{A}\in\mathrm{GA}^{1\text{st}}}{\operatorname{argmin}} \left\{f\left(\mathcal{A}_f x^{(k)}\right)\right\}.$$

This method is also memoryless.

Other methods of $\mathrm{GA}^{1\text{st}}$ include Conjugate Gradient, Quasi-Newton, and Gradient Descent with Momentum.

3.2 Lower Complexity Bounds for L-Lipschitz Smooth Function

The following is Nesterov [21, Thm 2.1.7].

Theorem 3.2.1 (Nesterov's Claim of Lower Bound)**.** For any $1 \leq k \leq (n-1)/2$, for all $x^{(0)} \in \mathbb{R}^n$, there exists a Lipschitz smooth convex function in $\mathbb{R}^n$ such that for all algorithm from $\mathrm{GA}^{1\text{st}}$, we have the lower bound for the optimality gap for the function values and its iterates:

$$f\left(x^{(k)}\right) - f^* \geq \frac{3L\|x - x^*\|^2}{32(k+1)^2}, \quad \|x^{(k)} - x^*\|^2 \geq \frac{1}{8}\|x^{(0)} - x^*\|^2.$$

Where x^* is the minimizer of f, so that $f(x^*) = \inf_x f(x)$.

Remark 3.2.1. We emphasize that in Theorem 3.2.1 fixes each k and finds a function f such that the lower bound applies at the k-th iterations. Mathematically, it would mean

$$\forall\, 1 \le k \le \frac{n+1}{2}, x^{(0)} \in \mathbb{R}^n \;\exists f \in \mathcal{F}_L^{1,1} \text{ s.t: } \min_{A\in \mathrm{GA}^{1\mathrm{st}}}\left\{f\left(A_f^k x^{(0)}\right)\right\} - f^* \ge \frac{3L\|x^{(0)}-x^*\|^2}{32(k+1)^2}$$

$$\forall\, 1 \le k \le \frac{n+1}{2}, x^{(0)} \in \mathbb{R}^n \quad \max_{f\in\mathcal{F}_L^{1,1}} \min_{A\in \mathrm{GA}^{1\mathrm{st}}}\left\{f\left(A_f^k x^{(0)}\right) - f^*\right\} \ge \frac{3L\|x^{(0)}-x^*\|^2}{32(k+1)^2}$$

$$\forall x^{(0)} \in \mathbb{R}^n \quad \min_{1\le k\le 1/2(n+1)} \max_{f\in\mathcal{F}_L^{1,1}} \min_{A\in \mathrm{GA}^{1\mathrm{st}}}\left\{f\left(A_f^k x^{(0)}\right) - f^*\right\} \ge \frac{3L\|x^{(0)}-x^*\|^2}{32(1/2(n+1))^2},$$

A function f_k provides the lower bound for fixed $1 \le k \le (n+1)/2$. k parameterized f_k, which Nesterov did in his proof. We emphasize that f_k is different depending on what k is. In addition, observe that minimizer x^* is assumed to exist. We believe that the theorem is generalizable to infinite dimensional Hilbert spaces.

We now quote Walkington [23, theorem 2.4]

Theorem 3.2.2 (Walkington's Claim of Lower Bound)**.** Let X be an infinite-dimensional Hilbert Space and set $x^{(0)} = \mathbf{0}$. There exists a convex function $f : X \mapsto \mathbb{R}$ with Lipschitz gradient and minimum $f(x_*) > -\infty$ such that for any sequence satisfying

$$x_{i+1} \in \operatorname{Span}\left\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \cdots, \nabla f(x^{(i)})\right\}, \quad i = 0, 1, 2, \cdots,$$

there holds

$$\min_{1\le i\le n} f(x_i) - f(x_*) \ge \frac{3L\|x_1 - x_*\|^2}{32(n+1)^2},$$

where L is the Lipschitz constant of the gradient.

Remark 3.2.2. Theorem 3.2.2 and Theorem 3.2.1 is completely different. The former claims there exists a single function from $\mathcal{F}_L^{1,1}(\mathcal{H})$ introduces the lower bound for all values of k, and all algorithms from $\mathrm{GA}^{1\mathrm{st}}$, but the latter didn't claim that. The difference would remain in infinite dimension Hilbert space if we were to generalize Thorem 3.2.1. There is no proof after Walkington's claim; we can't know if he had his way of proving the latter claim. It makes us think it is likely a missed detail in his writing.

3.3 Discussion

Walkington cited Bubeck [12, thm 3.14], and Nesterov's old 2004 book[20] for the lower bound claim. Bubeck has the correct claim, and it's the same as Nesterov. Attouch's claim in [5, thm 1] doesn't contradict Theorem 3.2.1, because he fixed function Φ function and the existence of minimizer x^* is not assumed.

4 FISTA Under Strong Convexity

In this section, we show several claims on the convergence proof of the algorithm of V-FISTA. Beck [8, 10.7.7] inspires the works. We removed one identity from their proof to reveal more transparency and interpretability. The original author intends to convince the reader with as few words as possible. We intend to educate and share thoughts. We present the essential claims here to expedite understanding of the big picture. The proofs with details are in the appendix.

4.1 Setting up the Stage

Starting with algorithm 1, $\theta_k = (t_k - 1)/(t_k + 1)$. We consider $F = g + h$, g is Lipschitz smooth with constant L, and strongly convex with constant σ. h is convex. The parameters, θ_k, and t_k, will be determined as we review the proof. $t_0 = 1$ is the base case for t_k sequence; it represents the fact that there are no accelerations on the first step of the algorithm; its value depends on what we want it to be.

Here is a list of quantities we constructed for a better exposition.

1. $s^{(k)} = x^{(k)} - x^{(k-1)}$, the velocity vector, for all $k \ge 1$.
2. $e^{(k)} = x^{(k)} - \bar{x}$, the error vector at the kth iteration, for $k \ge 0$.
3. $\theta_k = (t_k - 1)/(t_k + 1)$, which is the momentum step size.
4. $u^{(k)} = \bar{x} + t_k(x^{(k-1)} - x^{(k)}) - x^{(k-1)}$, the error term extrapolated with the velocity. We take $u^{(0)} = \bar{x} - x^{(0)}$.
5. $\delta_k = f(x^{(k)}) - f_{\text{opt}}$, with $f_{\text{opt}} = f(\bar{x}) = \inf_x f(x)$.
6. The quantity R_k plays a crucial role in the proof; it's
$$R_k = \frac{\sigma(t_{k+1} - 1)}{2}\left\|x^{(k)} - \bar{x}\right\|^2 - \frac{L - \sigma}{2}\left\|\bar{x} + t_{k+1}\left(x^{(k)} - y^{(k)}\right) - x^{(k)}\right\|^2$$
7. $\kappa = L/\sigma$, the condition number, it would be that $\kappa \ge 0$.
8. $q = \sigma/L$, the reciprical of the condition number. It would be that $q \in (0, 1)$.

4.2 Convergence Claim

We proposed the following alternative for proving the convergence rate of V-FISTA.

Lemma 4.2.1. If there exists a sequence $(t_k)_{k\in\mathbb{N}}, (C_k)_{k\in\mathbb{N}}$ in $\mathbb{R}$ such that

$$\begin{cases} \frac{C_k}{t_{k+1}^2} = \frac{L(1 - t_{k+1}^{-1})}{2t_k^2} \\ R_k + C_k\left\|u^{(k)}\right\|^2 \ge 0, \end{cases} \tag{4.2.1}$$

then

$$\delta_{k+1} \le \left(\prod_{i=0}^{k}(1 - t_k^{-1})\right)\left(\delta_0 + \frac{L}{2t_0^2}\left\|u^{(0)}\right\|^2\right).$$

Proof. See A.1. □

Proposition 4.1. If the sequence $(t_k)_{k\ge\mathbb{N}}$ by t_k satisfies $t_{k+1} = 1 + \frac{t_k^2(1-q)}{t_k+1}$, and $t_{k+1} \ge t_k + 1 - t_k^2 q$ and $t_k > 1$ for all $k \ge 0$, then lemma 4.2.1 stays true.

Proof. See A.2. □

Proposition 4.2. The choice of $t_k = t_{k+1} = \sqrt{L/\sigma}$ $\forall k \ge 0$ makes the proposed condition in proposition 4.2.1 true. Hence, the V-FISTA variant (3.) has a convergence rate bound

$$\delta_{k+1} \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k\left(\delta_0 + \frac{L}{2t_0^2}\left\|u^{(0)}\right\|^2\right).$$

Proof. From proposition 4.1 we have

$$\begin{aligned} \frac{t_{k+1}-1}{t_k^2(1-q)} - \frac{1}{t_k+1} &= 0 \quad \text{since } t_{k+1} = t_k, \\ \frac{t-1}{t^2(1-q)} &= \frac{1}{t+1} \\ t^2(1-q) &= t^2 - 1 \\ t^2 q &= 1 \\ t &= \pm\sqrt{\frac{L}{\sigma}}, \end{aligned}$$

with $t_k > 1$ it has to be $t_k = \sqrt{\frac{L}{\sigma}}$ for all $k \geq 0$. With that we have $t_{k+1} = t_k + 1 - t_k^2 q = t_k + 1 - 1 = t_k$, hence the inequality is also satisfied. □

Remark 4.2.1. We do not claim that these results are new. We came up with this proof and this way of presenting to understand Amir's proof[8, 10.7.7]. If the reader is interested in a proof that aims for genericity and arguably did a better job than the above claims, see Chambolle[15, Appendix B].

5 Numerical Experiments

We consider a signal length of 512, with $t_k = k$, and $t_k = 0$ for $0 \leq t \leq 256$ and $t_k = 1$ for $257 \leq t \leq 512$. We apply algorithm 1, variants (4.) to the TV minimization problem with initial guess $x^{(0)} = \mathbf{0}$, $\theta_1 = 0$. During the execution of the algorithm, we record the following quantities

1. The gradient mapping norm $\left\|y^{(k)} - T(y^{(k)})\right\|$.

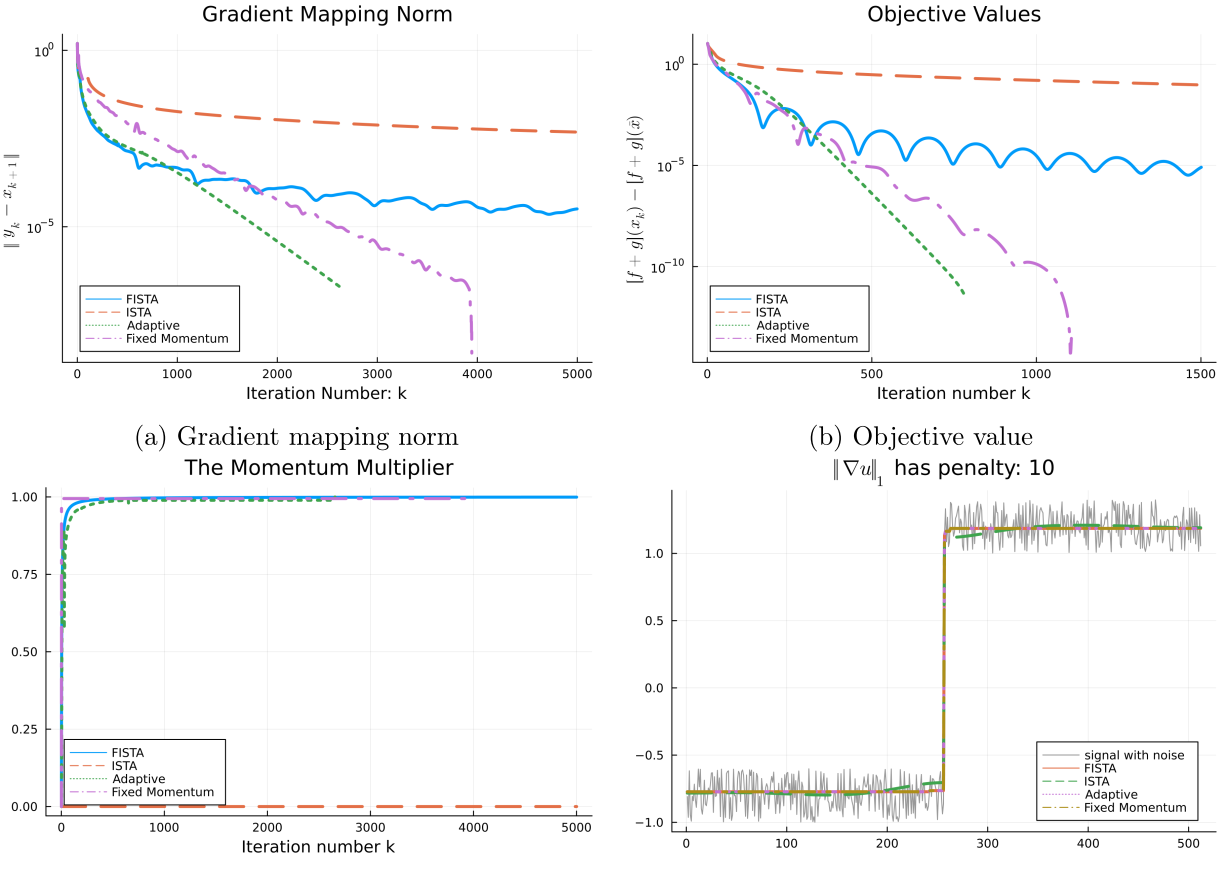
2. The optimality gap $F(x^{(k)}) - F_{\text{opt}}$, we estimate the optimality gap by choosing the smallest $F(x^{(k)})$ to be the minimum value from all competing algorithms, for all $k \geq 0$.

3. The θ_k, i.e., the momentum term.

The algorithm terminates upon $\|y^{(k)} - T(y^{(k)})\| \leq 10^{-10}$, or, a maximum iteration threshold is hit. We use the Julia programming language [11] for the numerical implementation; see the GitHub repository here.

5.1 Subsection

We showcase the results of $\alpha = 10$ in figure 1.



(a) Gradient mapping norm

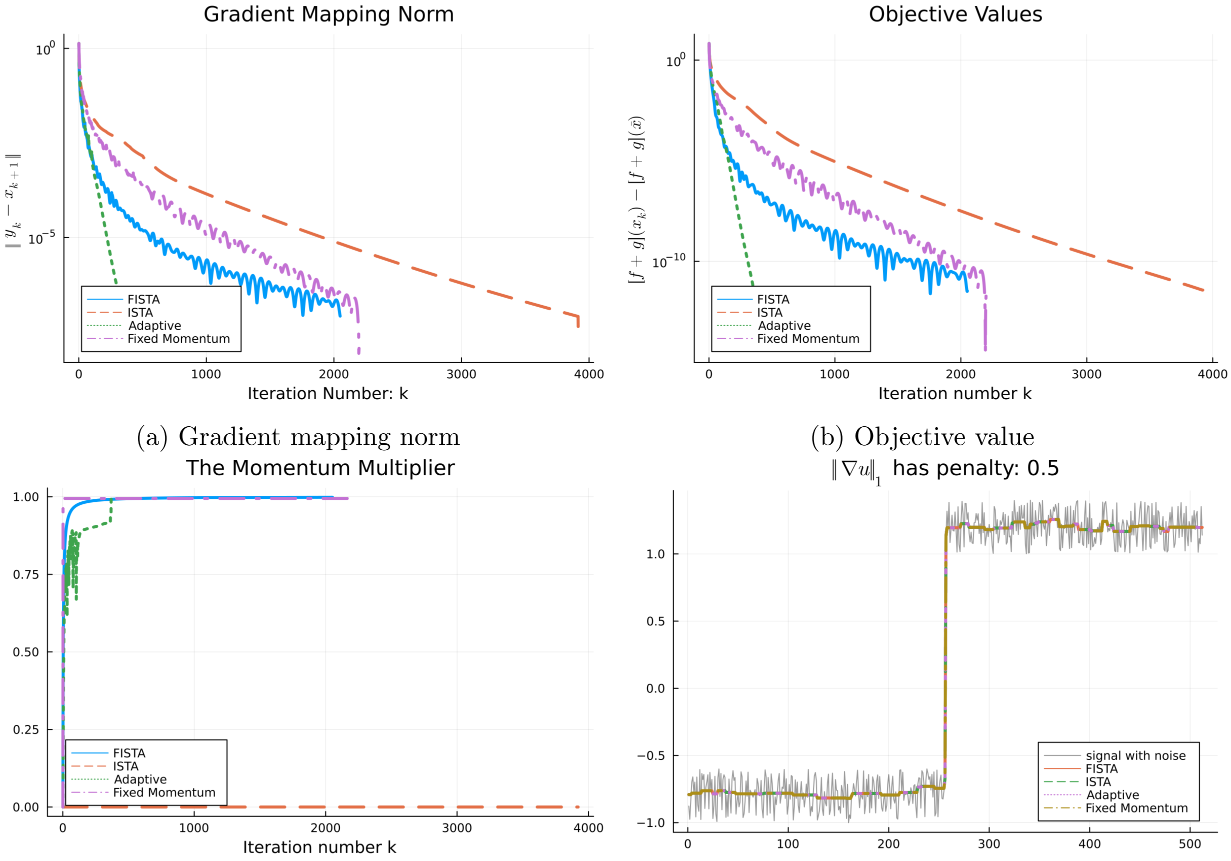
(b) Objective value

(c) Momentum

(d) Signal Recovery

Figure 1: Experiments with $\alpha = 10$, in legend, adaptive refers to our method of spectral momentum, fixed momentum refers to V-FISTA.

Different values of α affect the results for recovering the signal and convergences. A smaller value of α creates difficulty for the convergence for many methods except ISTA. It also severely affects the recovered signal. See figure 2 for more information.



(a) Gradient mapping norm

(b) Objective value

(c) Momentum

(d) Signal Recovery

Figure 2: Experiments with $\alpha = 0.5$, in legend, adaptive refers to our method of spectral momentum, fixed momentum refers to V-FISTA.

Our method of spectral momentum, described as variant (4.), has overwhelmingly fast empirical convergence results for the TV minimization problem. Both its effectiveness and convergence are still under investigation. We call it the spectral momentum method because of the method of Spectral Adapative Stepsize. As described in [17, 4.1], the method uses the same formula to estimate the smallest eigenvalues for the Hessian. However, our method is entirely different because we applied the estimation to the κ term for the momentum on V-FISTA.

Acknowledgement

Professor Shawn Wang at UBCO, who is also my advisor. He had doubts about what I said and asked great questions. That helps me develop a better understanding of Nesterov's lower complexity bound.

A Appendix

A.1 Proof for Lemma 4.2.1

For better notation we use T to denote the proximal gradient operator T_L appeared in Algorithm 1. We need the following lemma to start the proof. See [8, remark 10.17] for the Proximal Gradient Lemma.

Lemma A.1.1 (Proximal Gradient Lemma)**.** Let $F = g + h$ where h is convex closed and proper, g is L-Lipschitz smooth with a constant of L. Let $y \in \mathbb{R}^n$, we define $y^+ = T(y)$, then for any $x \in \mathbb{R}^n$, we have:

$$f(x) - f(y^+) \geq \frac{L}{2}\|x - y^+\|^2 - \frac{L}{2}\|x - y\|^2 + D_g(x, y),$$

Where $D_g(x, y) := g(x) - g(y) - \langle \nabla g(y), x - y \rangle$ is the Bregman Divergence for the smooth part of the sum: g.

The following lemma is from Nesterov's new book [22, thm 2.1.9, (2.1.23)].

Lemma A.1.2 (Strong Convexity and the Cute Formula)**.** Let f be continuous differentiable and μ-strongly convex on $Q \subseteq \mathbb{R}^n$, then for all $x, y \in Q$ and $\alpha \in [0, 1]$, we have the equivalent conditions

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x - y\|^2, \tag{A.1.1}$$

$$\alpha f(x) + (1 - \alpha) f(y) \geq f(\alpha x + (1 - \alpha)) + \frac{\mu\alpha(1 - \alpha)}{2}\|x - y\|^2. \tag{A.1.2}$$

Proof of Lemma 4.2.1. For the sequence $(t_k)_{k \in \mathbb{N} \cup \{0\}}$, we consider substituting x, y with:

1. $x = t_{k+1}^{-1}\bar{x} + (1 - t_{k+1}^{-1})x^{(k)}, y = y^{(k)}$.
2. $\bar{x} \in \underset{x}{\operatorname{argmin}}\, F(x)$ and $f(\bar{x}) = F_{\text{opt}}$.

From the Proximal Gradient Lemma, using strong convexity of g, $D_g(x, y) \geq \frac{\sigma}{2}\|y - x\|^2$ for all x, y then

$$\begin{aligned} F(x) - F \circ T(y) &\geq \frac{L}{2}\|x - T(y)\|^2 - \frac{L}{2}\|x - y\|^2 + D_g(x, y) && \text{(A.1.3)} \\ &\geq \frac{L}{2}\|x - Ty\|^2 - \frac{L - \sigma}{2}\|x - y\|^2. && \text{(A.1.4)} \end{aligned}$$

Expandind what we substituted into A.1, we can simplify and get

$$\begin{aligned} x - T(y) &= t_{k+1}^{-1}\bar{x} + (1 - t_{k+1}^{-1})x^{(k)} - Ty^{(k)} \\ &= t_{k+1}^{-1}\bar{x} + (1 - t_{k+1}^{-1})x^{(k)} - x^{(k+1)} \\ &= t_{k+1}^{-1}\left(\bar{x} + (t_{k+1} - 1)x^{(k)} - t_{k+1}x^{(k+1)}\right) \\ &= t_{k+1}^{-1}\left(\bar{x} + t_{k+1}\left(x^{(k)} - x^{(k+1)}\right) - x^{(k)}\right), \\ x - y &= t_{k+1}^{-1}\bar{x} + (1 - t_{k+1}^{-1})x^{(k)} - y^{(k)} \\ &= t_{k+1}^{-1}\left(\bar{x} + (t_{k+1} - 1)x^{(k)} - t_{k+1}y^{(k)}\right) \\ &= t_{k+1}^{-1}\left(\bar{x} + t_{k+1}\left(x^{(k)} - y^{(k)}\right) - x^{(k)}\right), \end{aligned}$$

rewriting, the RHS of A.1 so

$$\frac{L}{2}\left\|t_{k+1}^{-1}\left(\bar{x}+t_{k+1}\left(x^{(k)}-x^{(k+1)}\right)-x^{(k)}\right)\right\|^2-\frac{(L-\sigma)}{2}\left\|t_{k+1}^{-1}\left(\bar{x}+t_{k+1}\left(x^{(k)}-y^{(k)}\right)-x^{(k)}\right)\right\|^2.$$

The LHS of A.1 can be bounded using strong convexity of F.

$$\begin{aligned}
F(x)-F\circ T(y) &= F(x)-F\left(x^{(k+1)}\right)\\
&= F\left(t_{k+1}^{-1}x^{(k)}+(1-t_{k+1}^{-1})\bar{x}\right)-F\left(x^{(k+1)}\right)\\
&\leq t_{k+1}^{-1}F_{\text{opt}}+(1-t_{k+1}^{-1})F\left(x^{(k)}\right)-\frac{\sigma}{2}t_k^{-1}\left(1-t_{k+1}^{-1}\right)\left\|x^{(k)}-\bar{x}\right\|^2-F\left(x^{(k+1)}\right)\\
&= t_{k+1}^{-1}\left(F_{\text{opt}}-F\left(x^{(k)}\right)\right)+F\left(x^{(k)}\right)-F\left(x^{(k+1)}\right)-\frac{\sigma}{2}t_{k+1}^{-1}\left(1-t_{k+1}^{-1}\right)\left\|x^{(k)}-\bar{x}\right\|^2.
\end{aligned}$$

Denote $\delta_k := F\left(x^{(k)}\right)-F_{\text{opt}}$,

$$F(x)-F\circ T(y)\leq -t_{k+1}^{-1}\delta_k+\delta_k-\delta_{k+1}-\frac{\sigma}{2}t_{k+1}^{-1}\left(1-t_{k+1}^{-1}\right)\left\|x^{(k)}-\bar{x}\right\|^2.$$

With the above we present the full form of A.1 so

$$\begin{aligned}
&(1-t_{k+1}^{-1})\delta_k-\delta_{k+1}-\frac{\sigma}{2}t_{k+1}^{-1}\left(1-t_{k+1}^{-1}\right)\left\|x^{(k)}-\bar{x}\right\|^2\\
&\quad\geq\frac{L}{2}\left\|t_{k+1}^{-1}\left(\bar{x}+t_{k+1}\left(x^{(k)}-x^{(k+1)}\right)-x^{(k)}\right)\right\|^2-\frac{(L-\sigma)}{2}\left\|t_{k+1}^{-1}\left(\bar{x}+t_{k+1}\left(x^{(k)}-y^{(k)}\right)-x^{(k)}\right)\right\|^2.\\
&(t_{k+1}^2-t_{k+1})\delta_k-t_{k+1}^2\delta_{k+1}-\frac{\sigma}{2}(t_{k+1}-1)\left\|x^{(k)}-\bar{x}\right\|^2\\
&\quad\geq\frac{L}{2}\left\|\bar{x}+t_{k+1}\left(x^{(k)}-x^{(k+1)}\right)-x^{(k)}\right\|^2-\frac{(L-\sigma)}{2}\left\|\bar{x}+t_{k+1}\left(x^{(k)}-y^{(k)}\right)-x^{(k)}\right\|^2\\
&(t_{k+1}^2-t_{k+1})\delta_k-t_{k+1}^2\delta_{k+1}\underbrace{-\frac{\sigma(t_{k+1}-1)}{2}\left\|x^{(k)}-\bar{x}\right\|^2+\frac{L-\sigma}{2}\left\|\bar{x}+t_{k+1}\left(x^{(k)}-y^{(k)}\right)-x^{(k)}\right\|^2}_{-R_k}\\
&\quad\geq\frac{L}{2}\left\|\bar{x}+t_{k+1}\left(x^{(k)}-x^{(k+1)}\right)-x^{(k)}\right\|^2.
\end{aligned}\tag{A.1.5}$$

Recall the quantities from section 4, and from the algorithm

$$\begin{aligned}
y^{(k)} &= x^{(k)}+\theta_k(x^{(k)}-x^{(k-1)})\\
y^{(k)}-x^{(k)} &= \theta_k(x^{(k)}-x^{(k-1)})=\theta_k s^{(k)}.
\end{aligned}$$

Simplifying,

$$\begin{aligned}
&\frac{L-\sigma}{2}\left\|\bar{x}+t_{k+1}\left(x^{(k)}-y^{(k)}\right)-x^{(k)}\right\|^2\\
=&\frac{L-\sigma}{2}\left\|\bar{x}-x^{(k)}-t_{k+1}\theta_k s^{(k)}\right\|^2\\
=&\frac{L-\sigma}{2}\left\|e^{(k)}+t_{k+1}\theta_k s^{(k)}\right\|^2,
\end{aligned}$$

Observe that $u^{(k+1)} = \bar{x} + t_{k+1}(x^{(k)} - x^{(k+1)}) - x^{(k)}$, it is in the norm on the RHS of A.1.5. $u^{(k)}$ has representation by $e^{(k)}, x^{(k)}$.

$$\begin{aligned}
u^{(k)} &= \bar{x} + t_k \left(x^{(k-1)} - x^{(k)}\right) - x^{(k-1)} \\
&= \bar{x} + (t_k - 1)\left(x^{(k-1)} - x^{(k)}\right) + \left(x^{(k-1)} - x^{(k)}\right) - x^{(k-1)} \\
&= \bar{x} + (t_k - 1)\left(x^{(k-1)} - x^{(k)}\right) - x^{(k)} \\
&= -e^{(k)} - (t_k - 1)s^{(k)},
\end{aligned}$$

with these simplifications, we will be able to write down A.1.5 as

$$R_k = \frac{\sigma(t_{k+1} - 1)}{2}\left\|e^{(k)}\right\|^2 - \frac{L - \sigma}{2}\left\|e^{(k)} + t_{k+1}\theta_k s^{(k)}\right\|^2, \tag{A.1.6}$$

$$t_{k+1}(t_{k+1} - 1)\delta_k - R_k \geq t_{k+1}^2\delta_{k+1} + \frac{L}{2}\left\|u^{(k+1)}\right\|^2. \tag{A.1.7}$$

We are now prepared for deriving a bound on the convergence rate of the algorithm.

$$\begin{aligned}
t_{k+1}^2\delta_{k+1} + \frac{L}{2}\left\|u^{(k+1)}\right\|^2 - C_k\left\|u^{(k)}\right\|^2 &\leq t_{k+1}(t_{k+1} - 1)\delta_k - R_k - C_k\left\|u^{(k)}\right\|^2 \\
t_{k+1}^2\delta_{k+1} + \frac{L}{2}\left\|u^{(k+1)}\right\|^2 - C_k\left\|u^{(k)}\right\|^2 &\leq t_{k+1}(t_{k+1} - 1)\delta_k \\
\delta_{k+1} \leq (1 - t_{k+1}^{-1})\delta_k + \frac{C_k}{t_{k+1}^2}\left\|u^{(k)}\right\|^2 &- \frac{L}{2t_{k+1}^2}\left\|u^{(k+1)}\right\|^2.
\end{aligned}$$

Going from the first inequality to the second we used $R_k + C_k\left\|u^{(k)}\right\| \geq 0$ in lemma 4.2.1 Going from the second inequality to the third we devided both sides by t_{k+1}^2 with rearrangement. Let's use $\frac{C_k}{t_{k+1}^2} = \frac{L(1 - t_{k+1}^{-1})}{2t_k^2}$ from lemma 4.2.1 so

$$\begin{aligned}
\delta_{k+1} &\leq (1 - t_{k+1}^{-1})\delta_k + \frac{L(1 - t_{k+1}^{-1})}{2t_k^2}\left\|u^{(k)}\right\|^2 - \frac{L}{2t_{k+1}^2}\left\|u^{(k+1)}\right\|^2 \\
\delta_{k+1} &\leq (1 - t_{k+1}^{-1})\left(\delta_k + \frac{L}{2t_k^2}\left\|u^{(k)}\right\|^2\right) - \frac{L}{2t_{k+1}^2}\left\|u^{(k+1)}\right\|^2 \\
&\triangleright \text{ unrolling recursion yield} \\
\delta_{k+1} &\leq \left(\prod_{i=0}^{k}(1 - t_k^{-1})\right)\left(\delta_0 + \frac{L}{2t_0^2}\left\|u^{(0)}\right\|^2\right).
\end{aligned}$$

We are done. □

A.2 Proof for Proposition 4.1

Proof. Observe that condition from lemma 4.2.1 implies

$$\begin{aligned}
R_k + C_k\left\|u^{(k)}\right\|^2 &\geq 0 \\
\frac{\sigma(t_{k+1} - 1)}{2}\left\|e^{(k)}\right\|^2 - \frac{L - \sigma}{2}\left\|e^{(k)} + t_{k+1}\theta_k s^{(k)}\right\|^2 + \frac{Lt_{k+1}(t_{k+1} - 1)}{2t_k^2}\left\|u^{(k)}\right\|^2 &\geq 0 \\
\frac{\sigma(t_{k+1} - 1)}{L - \sigma}\left\|e^{(k)}\right\|^2 - \left\|e^{(k)} + t_{k+1}\theta_k s^{(k)}\right\|^2 + \frac{Lt_{k+1}(t_{k+1} - 1)}{t_k^2(L - \sigma)}\left\|e^{(k)} + (t_k - 1)s^{(k)}\right\|^2 &\geq 0
\end{aligned}$$

From the first to the second line, divide $\frac{L-\sigma}{2}$, From the second to the thrid line recall $u^{(k)} = -e^{(k)} - (t_k - 1)s^{(k)}$. Since the vector quantities $e^{(k)}, s^{(k)}$ share the same superscript, we may ignore it. Expanding the expression would yield quantities $\langle s, e\rangle, \|s\|^2, \|e\|^2$, we list them for each term all

$$
\begin{aligned}
&-\|e + t_{k+1}\theta_k s\|^2 = -\left(\|e\|^2 + t_{k+1}^2\theta_k^2\|s\|^2 + 2\theta_k t_{k+1}\langle e, s\rangle\right), \\
&\frac{Lt_{k+1}(t_{k+1}-1)}{t_k^2(L-\sigma)}\|e + (t_k - 1)s\|^2 = \frac{Lt_{k+1}(t_{k+1}-1)}{t_k^2(L-\sigma)}\left(\|e\|^2 + (t_k-1)^2\|s\|^2 + 2(t_k - 1)\langle e, s\rangle\right), \\
&\frac{\sigma(t_{k+1}-1)}{L-\sigma}\|e\|^2.
\end{aligned}
$$

Grouping each of the terms $\|e\|^2, \|s\|^2, \langle e, s\rangle$, we compute their coefficients with $q = \sigma/L$,

$$
\begin{aligned}
\|e\|^2 \text{ has: } \quad & \frac{\sigma(t_{k+1}-1)}{L-\sigma} + \frac{Lt_{k+1}(t_{k+1}-1)}{t_k^2(L-\sigma)} - 1 \\
&= (t_{k+1}-1)\left(\frac{q}{1-q} + \frac{t_{k+1}}{t_k^2(1-q)}\right) - 1 \\
&= \frac{t_{k+1}-1}{1-q}\left(q + \frac{t_{k+1}}{t_k^2}\right) - 1 \\
\|s\|^2 \text{ has: } \quad & \frac{Lt_{k+1}(t_{k+1}-1)}{t_k^2(L-\sigma)}(t_k-1)^2 - t_{k+1}^2\theta_k^2 \\
&= \frac{t_{k+1}(t_{k+1}-1)}{t_k^2(1-q)}(t_k-1)^2 - t_{k+1}^2\left(\frac{t_k-1}{t_k+1}\right)^2 \\
&= (t_k-1)^2\left(\frac{t_{k+1}(t_{k+1}-1)}{t_k^2(1-q)} - \frac{t_{k+1}^2}{(t_k+1)^2}\right) \\
\langle s, e\rangle \text{ has: } \quad & \frac{2Lt_{k+1}(t_{k+1}-1)}{t_k^2(L-\sigma)}(t_k-1) - 2\theta_k t_{k+1} \\
&= 2(t_k-1)t_{k+1}\left(\frac{t_{k+1}-1}{t_k^2(1-q)} - \frac{1}{t_k+1}\right),
\end{aligned}
$$

To satisfy the assumption, it would be great to have the coefficients for $\langle s, e\rangle$ to be zero, and the coefficients of $\|e\|^2, \|s\|^2$ to be a positive quantities.

For the coefficient of $\langle s, e\rangle$ to be zero, it would imply the condition

$$
\frac{t_{k+1}-1}{t_k^2(1-q)} = \frac{1}{t_k+1} \tag{A.2.1}
$$

$$
t_{k+1} = \frac{t_k^2(1-q)}{t_k+1} + 1, \tag{A.2.2}
$$

we assume that $t_k \neq 1$ for all $k \geq 0$. Next, we consider the non-negativity condition for the coefficients of $\|e\|^2$, so

$$
\begin{aligned}
\frac{t_{k+1}-1}{1-q}\left(q + \frac{t_{k+1}}{t_k^2}\right) - 1 \geq 0, &\quad \text{using } \frac{t_{k+1}-1}{1-q} = \frac{t_k^2}{t_k+1} \\
\frac{t_k^2}{t_k+1}\left(q + \frac{t_{k+1}}{t_k^2}\right) > 1, &\quad \text{using } t_k > 1, \forall k \geq 0, \text{ from proposition 4.1}, \\
t_k^2 q + t_{k+1} \geq t_k + 1 & \\
t_{k+1} \geq t_k + 1 - t_k^2 q. &
\end{aligned} \tag{A.2.3}
$$

Similarly the non-negatvity constraints for $\|s^{(k)}\|^2$ would yield

$$\begin{aligned}
(t_k-1)^2\left(\frac{t_{k+1}(t_{k+1}-1)}{t_k^2(1-q)}-\frac{t_{k+1}^2}{(t_k+1)^2}\right) &\geq 0, \text{ by } \frac{t_{k+1}-1}{t_k^2(1-q)}=\frac{1}{t_k+1}\\
(t_k-1)^2\left(\frac{t_{k+1}}{t_k+1}-\frac{t_{k+1}^2}{(t_k+1)^2}\right) &\geq 0, \ (t_k-1)>0, \text{ so we devided by } (t_k-1)^2\\
\frac{t_{k+1}}{t_k+1} &\geq \frac{t_{k+1}^2}{(t_k+1)^2}\\
\frac{1}{t_k+1} &\geq \frac{t_{k+1}}{(t_k+1)^2}\\
1 &\geq \frac{t_{k+1}}{t_k+1}\\
t_k+1 &\geq t_{k+1}.
\end{aligned}$$

Now, under the assumption of $t_k > 1$, the above condition would be redundant because A.2.1 has

$$\begin{aligned}
\frac{t_{k+1}-1}{t_k^2(1-q)}-\frac{1}{t_k+1} &= 0\\
t_{k+1}-1-\frac{t_k^2(1-q)}{t_k+1} &= 0\\
t_{k+1} &= 1+\frac{t_k^2(1-q)}{t_k+1}\\
&\leq 1+t_k^2/(t_k+1)\\
&\leq 1+t_k^2/t_k = 1+t_k.
\end{aligned}$$

We are done. With (A.2.3), (A.2.1), and $t_k > 1$ we can use the generic convergence rate stated in Proposition 4.1. □

References

[1] K. Ahn and S. Sra, *Understanding Nesterov's Acceleration via Proximal Point Method*, June 2022. arXiv:2005.08304 [cs, math].

[2] T. Alamo, P. Krupa, and D. Limon, *Gradient Based Restart FISTA*, in 2019 IEEE 58th Conference on Decision and Control (CDC), Dec. 2019, pp. 3936–3941. ISSN: 2576-2370.

[3] C. An, H.-N. Wu, and X. Yuan, *The Springback Penalty for Robust Signal Recovery*, Applied and Computational Harmonic Analysis, 61 (2022), pp. 319–346. arXiv:2110.06754 [cs, math].

[4] ——, *Enhanced Total Variation Minimization for Stable Image Reconstruction*, Inverse Problems, 39 (2023), p. 075005. arXiv:2201.02979 [cs, eess, math].

[5] H. Attouch and J. Peypouquet, *The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method is Actually Faster Than \$1/k^2\$*, SIAM Journal on Optimization, 26 (2016), pp. 1824–1834. Publisher: Society for Industrial and Applied Mathematics.

[6] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre, *Parameter-Free FISTA by Adaptive Restart and Backtracking*, arXiv.org, (2023).

[7] J.-F. Aujol, C. H. Dossal, H. Labarrière, and A. Rondepierre, *FISTA restart rsing an automatic estimation of the growth parameter*, (2022).

[8] A. Beck, *First-Order Methods in Optimization | SIAM Publications Library*, MOS-SIAM Series in Optimization, SIAM.

[9] A. Beck and M. Teboulle, *Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems*, IEEE Transactions on Image Processing, 18 (2009), pp. 2419–2434. Conference Name: IEEE Transactions on Image Processing.

[10] ——, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[11] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A Fresh Approach to Numerical Computing*, SIAM Review, 59 (2017), pp. 65–98. Publisher: Society for Industrial and Applied Mathematics.

[12] S. Bubeck, *Convex Optimization: Algorithms and Complexity*, (2015). arXiv:1405.4980 [cs, math, stat].

[13] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, *An Introduction to Total Variation for Image Analysis*, in Theoretical Foundations and Numerical Methods for Sparse Recovery, M. Fornasier, ed., DE GRUYTER, July 2010, pp. 263–340.

[14] 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 (2015), pp. 968–982.

[15] A. Chambolle and T. Pock, *An Introduction to Continuous Optimization for Imaging*, Acta Numerica, 25 (2016), pp. 161–319. Publisher: Cambridge University Press (CUP).

[16] O. Fercoq and Z. Qu, *Adaptive restart of accelerated gradient methods under local quadratic growth condition*, IMA Journal of Numerical Analysis, 39 (2019), pp. 2069–2095. arXiv:1709.02300 [math].

[17] T. Goldstein, C. Studer, and R. Baraniuk, *A Field Guide to Forward-Backward Splitting with a FASTA Implementation*, Dec. 2016. arXiv:1411.3406 [cs].

[18] U. Jang, S. D. Gupta, and E. K. Ryu, *Computer-Assisted Design of Accelerated Composite Optimization Methods: OptISTA*, May 2023. arXiv:2305.15704 [math].

[19] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175 (2019), pp. 69–107.

[20] Y. Nesterov, *Introductory Lectures on Convex Optimization*, vol. 87 of Applied Optimization, Springer US, Boston, MA, 2004.

[21] ——, *Lecture on Convex Optimizations Chapter 2, Smooth Convex Optimization*, in Lectures on Convex Optimization, Y. Nesterov, ed., Springer Optimization and Its Applications, Springer International Publishing, Cham, 2018, pp. 59–137.

[22] ——, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, Cham, 2018.

[23] W. Noel, *Nesterov's Method for Convex Optimization*, SIAM Review, 65, pp. 539–562.

[24] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear Total Variation Based Noise Removal Algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.

[25] W. Su, S. Boyd, and E. J. Candes, *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights*, arXiv.org, (2015).