MATH 590 2023 FALL REPORT

HONGDA LI

November 18, 2023

Abstract

In this paper, we review the paper written by Walkington [1] 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 give a literature review on the frontier for both the theories and applications around FISTA. We correct one misconception that occurred in Walkington [1] 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 [2, theorem 10.7.7] by eliminating an identity used in their proof. Finally, we use the Forward Backward Envelope to adapt smooth non-smooth additive objective to Nesterov's generic accelerated gradient algorithm [3, 2.2.7].

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 [1], 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),

$$f(u) = \int_0^1 \frac{1}{2}(u - \hat{u})^2 + \alpha|u'|dt.$$

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[1]. We derive the

approximation of the integral. Denote $s_i = u_i - \hat{u}_i$.

$$\begin{aligned}
\frac{1}{2}\int_0^1 (u-\hat{u})^2 dt + \alpha \int_0^1 |u'| dt &\approx \frac{1}{2}\sum_{i=0}^{N}\left(\frac{s_i^2 + s_{i+1}^2}{2}\right) h_{i+1} + \alpha \sum_{i=1}^{N}\left|\frac{u_i - u_{i-1}}{h_{i+1}}\right| \\
&\triangleright \text{ let } C \in \mathbb{R}^{N\times(N+1)} \text{ be upper bi-diagonal with } (1,-1) \\
&= \frac{1}{2}\left(\frac{s_0^2 h_1}{2} + \frac{s_N^2 h_N}{2} + \sum_{i=1}^{N-1} s_i^2 h_i\right) + \alpha\|Cu\|_1 \\
&\triangleright \text{ using } D \in \mathbb{R}^{N\times(N+1)}, \\
&\triangleright \; D := \operatorname{diag}(h_1/2, h_1, h_2, \cdots, h_N, h_N/2) \\
&= \frac{1}{2}\langle u-\hat{u}, D(u-\hat{u})\rangle + \alpha\|Cu\|_1.
\end{aligned}$$

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

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

With the assumption that D is positive definite, we have

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

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 is obtained by

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

$-g(\lambda)$ is easier to optimize, and obtaining the primal solution is also simple since D^{-1} is a diagonal matrix.

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 [4] can solve with a convergence rate of $\mathcal{O}(1/k^2)$ on the objective value of the function. However, highlighted in Walkington[1], 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. It sparks the discussion in this paper on the variants of FISTA, hoping to provide some insights on why Nesterov's momentum-based method's inability 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[4]. 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[5] 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. Presenting FISTA and its variants is the second part. The third part reviews the algorithmic tricks and improvements applied to the APG. Section 3 addresses a mistake made in Walkington's writing [1, 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 [2, 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 [6]. They pioneer the theories of TV minimization by solving PDE. They discussed the empirical observation that the L1 regularization term produces sharper images. Walkington [1] 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 [7, 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 [5]. For a complete survey of the state of arts computational methods applied to TV minimizations, see Chambolle [8]. 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.[9]. For frontier work of applying non-convex penalty term and its theoretical guarantee consult [10], [11].

Variants of FISTA

Walkington's writing on the method of V-FISTA and accelerated gradient[1, section 4, section 3] consists of proofs that are too short and uninformative for good understanding. The frustration motivates us to look for better proofs of the algorithm's convergence rate in other literature. It was a surprise that Walkington did not cite Nesterov's new book[12]. We contextualize Walkington's approach with Amir Beck's book[2] and Nesterov's book [12].

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

1: **Input:** $(g, h, x^{(0)})$
2: $y^{(0)} = x^{(0)}$, $\kappa = L/\sigma$
3: **for** $k = 0, 1, \cdots$ **do**
4: $\quad x^{(k+1)} = \mathcal{T}_L y^{(k)}$
5: $\quad y^{(k+1)} = x^{(k+1)} + \theta_{k+1}(x^{(k+1)} - x^{(k)})$
6: $\quad$ Execute subroutine $\mathcal{S}$.
7: **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. Chaning $\mathcal{T}_L, \theta_{k+1}$ and $\mathcal{S}$, produce different variants of FISTA.

1. Original FISTA proposed by Beck [4] considers $\theta_{k+1} = (t_k - 1)/t_{k+1}$, $t_{k+1}(t_{k+1} - 1) = t_k^2$, with $\mathcal{T}_L x = \text{prox}_{L^{-1}h}(x - L^{-1}\nabla g(x))$ and $t_0 = 1$. This achieves $\mathcal{O}(1/k^2)$ on the objective value. Our literature review didn't discover proofs 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. A version of FISTA where the iterates have weak convergence was proved in Chambolle, Dossal [13]. It's a variant where $\theta_{k+1} = (n + a - 1)/a$, for $a > 2$. $\mathcal{T}_L$ 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 [2, 10.7.7][1, 3.3]. $\mathcal{T}_L$ is the same as (1.).

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

$$\sigma \approx \langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle / \|x^{(k+1)} - x^{(k)}\|^2.$$

   It yields excellent results for the numerical experiments.

5. MFISTA in Beck[5] 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 Hessian. 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, Aujol et al.[14] proposed an automatic restart algorithm that achieves faster convergence without knowing the prior strong convexity (or weaker quadratic growth condition) parameter σ. Later, they developed the idea into a parameter-free algorithm in their work [15]. Earlier attempts proved a fast linear convergence rate under quadratic growth conditions by triggering the restart of FISTA based on the gradient mapping norm. See [16][16][17]. 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. [18] 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 [19], 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[3], 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 [20]. Finally, for a theoretical underpinning of Nesterov's generic method in his book, consult Ahn and Sra[21]. 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.

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[12]. Walkington[1] rephrased his work with one crucial mistake in understanding Nesterov's claim. We detail Nesterov's claim and provide context for understanding the mistakes in Walkington.

3.1 First-order Method

The following is rephrased from Assumption [3, 2.1.4].

Definition 1 (First Order Method)**.** We are in $\mathbb{R}^n$ for now. Given $x^{(0)} \in \mathbb{R}^n$, an iterative algorithm generates sequence of $\left(x^{(n)}\right)_{n\in\mathbb{N}}$ in the space. All classes $\in \mathrm{GA}^{1\text{st}}$ satisfy that

$$x^{(k+1)} \in \left\{x^{(0)}\right\} + \operatorname{span}\left\{\nabla f\left(x^{(i)}\right)\right\}_{i=1}^{k-1}.$$

Let $\mathcal{A}_f^k x^{(0)}$ denotes the solution of the k-th iterate $x^{(k)}$ generated by an algorithm $\mathcal{A} \in \mathrm{GA}^1$, with initial guess $x^{(0)}$.

4 FISTA Under Strong Convexity

4.1 Subsection

5 Numerical Experiments

A Appendix

This is a new section.

A.1 Subsection

This is a subsection.

A.2 Cute Subsection

References

[1] W. Noel, "Nesterov's Method for Convex Optimization," *SIAM Review*, vol. 65, no. 2, pp. 539–562. [Online]. Available: https://epubs-siam-org.eu1.proxy.openathens.net/doi/epdf/10.1137/21M1390037

[2] A. Beck, *First-Order Methods in Optimization | SIAM Publications Library*, ser. MOS-SIAM Series in Optimization. SIAM. [Online]. Available: https://epubs.siam.org/doi/book/10.1137/1.9781611974997

[3] Y. Nesterov, "Lecture on Convex Optimizations Chapter 2, Smooth Convex Optimization," in *Lectures on Convex Optimization*, ser. Springer Optimization and Its Applications, Y. Nesterov, Ed. Cham: Springer International Publishing, 2018, pp. 59–137. [Online]. Available: https://doi.org/10.1007/978-3-319-91578-4_2

[4] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009. [Online]. Available: http://epubs.siam.org/doi/10.1137/080716542

[5] ——, "Fast Gradient-Based Algorithms for Constrained Total Variation Image Denoising and Deblurring Problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009, conference Name: IEEE Transactions on Image Processing. [Online]. Available: https://ieeexplore.ieee.org/document/5173518

[6] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, Nov. 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016727899290242F

[7] T. Goldstein, C. Studer, and R. Baraniuk, "A Field Guide to Forward-Backward Splitting with a FASTA Implementation," Dec. 2016, arXiv:1411.3406 [cs]. [Online]. Available: http://arxiv.org/abs/1411.3406

[8] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," *Acta Numerica*, vol. 25, pp. 161–319, 2016, publisher: Cambridge University Press (CUP). [Online]. Available: https://hal.science/hal-01346507

[9] 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, Jul. 2010, pp. 263–340. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/9783110226157.263/html

[10] C. An, H.-N. Wu, and X. Yuan, "Enhanced total variation minimization for stable image reconstruction," *Inverse Problems*, vol. 39, no. 7, p. 075005, Jul. 2023, arXiv:2201.02979 [cs, eess, math]. [Online]. Available: http://arxiv.org/abs/2201.02979

[11] ——, "The springback penalty for robust signal recovery," *Applied and Computational Harmonic Analysis*, vol. 61, pp. 319–346, Nov. 2022, arXiv:2110.06754 [cs, math]. [Online]. Available: http://arxiv.org/abs/2110.06754

[12] Y. Nesterov, *Lectures on Convex Optimization*, ser. Springer Optimization and Its Applications. Cham: Springer International Publishing, 2018, vol. 137. [Online]. Available: http://link.springer.com/10.1007/978-3-319-91578-4

[13] A. Chambolle and C. Dossal, "On the Convergence of the Iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm"," *Journal of Optimization Theory and Applications*, vol. 166, no. 3, pp. 968–982, Sep. 2015. [Online]. Available: https://doi.org/10.1007/s10957-015-0746-4

[14] J.-F. Aujol, C. H. Dossal, H. Labarrière, and A. Rondepierre, "FISTA restart using an automatic estimation of the growth parameter," May 2022. [Online]. Available: https://hal.science/hal-03153525

[15] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre, "Parameter-Free FISTA by Adaptive Restart and Backtracking," *arXiv.org*, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.14323v1

[16] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9029983

[17] O. Fercoq and Z. Qu, "Adaptive restart of accelerated gradient methods under local quadratic growth condition," *IMA Journal of Numerical Analysis*, vol. 39, no. 4, pp. 2069–2095, Oct. 2019, arXiv:1709.02300 [math]. [Online]. Available: http://arxiv.org/abs/1709.02300

[18] W. Su, S. Boyd, and E. J. Candes, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," *arXiv.org*, Mar. 2015. [Online]. Available: https://arxiv.org/abs/1503.01243v2

[19] 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*, vol. 26, no. 3, pp. 1824–1834, Jan. 2016, publisher: Society for Industrial and Applied Mathematics. [Online]. Available: https://epubs.siam.org/doi/10.1137/15M1046095

[20] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Mathematical Programming*, vol. 175, no. 1, pp. 69–107, May 2019. [Online]. Available: https://doi.org/10.1007/s10107-018-1232-1

[21] K. Ahn and S. Sra, "Understanding Nesterov's Acceleration via Proximal Point Method," Jun. 2022, arXiv:2005.08304 [cs, math]. [Online]. Available: http://arxiv.org/abs/2005.08304