MATH 590 2023 FALL REPORT

HONGDA LI

November 16, 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 theoretical development of the FISTA algorithm. 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.

1 Preliminaries

In this section, We motivate the discussion about 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

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

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

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

$-g(\lambda)$ 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 [3] 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 $\mathcal{O}((1-1/\kappa)^k)$ 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 Nesterove'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[3]. 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 or Nesterov Accelerated gradient 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 [4]. 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] 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 [5, 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 [6]. For a complete survey of the state of arts computational methods applied to TV minimizations, see Chambolle [7]. 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.[8]. For frontier work of applying non-convex penalty term and its theoretical guarantee consult [9], [10].

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[11]. We contextualize Walkington's approach with Amir Beck's book[2] and Nesterov's book [11].

For different Variants of FISTA, we specifically refer to the first-order acceleration method based on Nesterov's Framework, adhering to Nesterov's lecture [12] and Beck's book [2, chapter 10]. The algorithm that all three writers were talking about is the V-FISTA algorithm (algorithm 1).

Algorithm 1 V-FISTA

```
Input: $(f, g, x^{(0)})$
$y^{(0)} = x^{(0)}$, $\kappa = L/\sigma$
$x^{(k+1)} = \text{prox}_{1/Lg}(y^{(k)} - (1/L)\nabla f(y^{(k)}))$
$y^{(k+1)} = x^{(k+1)} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)(x^{(k+1)} - x^{(k)})$
```

The V-FISTA algorithm (algorithm 1), when applied to an objective function of the type $F = f + g$, where f is prox friendly and g is L-Lipschitz and σ-strongly convex, it achieves a convergece rate of $\mathcal{O}((1-1/\sqrt{\kappa})^k)$ for the function objective. This convergent rate is faster than $\mathcal{O}((1-1/\kappa)^k)$ for the proximal gradient. The parameter $\kappa = L/\sigma$ is carefully chosen to achieve a linear convergence rate.

The V-FISTA algorithm contains more parameters. A L-Lipschitz smooth convex function g, is σ-strongly by adding $\sigma\|\cdot\|^2/2$. Hence, define $g_\sigma(x) = f(x) + \sigma\|x\|^2/2$ then we have $\lim_{\sigma\to 0} g_\sigma(x) = f(x)$. The function transitions smoothly from a strongly convex function to just a convex function. It's tempted to think that if we take the limit of $\sigma \to 0$ on V-FISTA (algorithm 1) would yield the FISTA algorithm, analogous to the function $f_\sigma \to f$. However, this is not the case since $\lim_{\sigma\to 0} \kappa = \infty$, making

$$\lim_{\sigma\to 0} \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1.$$

This didn't result in FISTA. Finding one generic algorithm to derive and describe both FISTA and V-FISTA would be of good interest.

As it turns out, Nesterov[12] derived the smooth counterpart of both algorithms from one generic algorithm[12, (2.2.7)], but his argument is not directly applicable to a function of smooth and non-smooth composite. Nesterov's lectures focus on establishing the most general frameworks for this method class. His approach doesn't handle function that is non-smooth, at least not directly. Nesterov's completed derivation of the linear convergence rate and sub-linear convergence rate is in theorem 2.2.2, based on the results of lemma 2.2.4. We emphasize that in Nesterov's writing, his derivation of linear and sub-linear convergence of the function objective was done in lemma 2.2.4, without the assumption of a minimizer, for Lipschitz smooth function with and without/without strong convexity, in one single proof. Amir Beck's book had the algorithm V-FISTA presented in 10.7.7; it's an algorithm that is the same as Nesterov's algorithm: 2.2.22, with the addition of a proximal operator. Unfortunately, Amir Beck doesn't have a unified framework of descriptions of both FISTA and V-FISTA.

There are other variants of FISTA. the MFISTA method proposed by Beck [3, theorem 5.1], was empirically more robust and stable. However, it still retains the $\mathcal{O}(1/k^2)$ convergence rate at that time. The idea of restarting FISTA, however, refuses to die. More recently, the work of [13] proposed a condition of restarting FISTA such that a global linear convergence rate can be achieved based on the quadratic growth condition, a weaker condition than strong convexity. Faster forward to the frontier, Aujol et al.[14] proposed an automatic restart algorithm that achieves faster convergence without knowing the strong convexity (or the weaker quadratic growth condition) parameter in prior. They then developed the idea into a parameter-free algorithm in their work [15].

Simultaneously, looking for a unified framework behind Nesterov's acceleration theory continues. One of the recent theoretical improvements by[16] that is based on the new writing from Nesterov's book, derived and unified a lot of variants of Nesterov's acceleration method using the proximal point method proposed by Rockefeller.

3 Nesterov's Lower Bound Clarified

3.1 title

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] 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

[4] 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

[5] 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

[6] A. Beck and M. Teboulle, “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

[7] 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

[8] 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

[9] 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

[10] ——, “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

[11] 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

[12] ——, “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

[13] T. Alamo, P. Krupa, and D. Limon, “Restart FISTA with Global Linear Convergence,” Dec. 2019, arXiv:1906.09126 [math]. [Online]. Available: http://arxiv.org/abs/1906.09126

[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] 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