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 FISTA method and some of its variants with numerical experiments on the total variation minimization problem, in addition we propose a heuristic estimation of strong convexity parameter and demonstrate that it converges faster when applied. We give literature review on the frontier theoretical development on the FISTA algorithm. We correct one misconception occured 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 the their proof.

1 Preliminaries

In this section, We present and model the problem of denoising one dimension signal. The dual objective function of the problem derived in this section motivates the use of Accelerated Proximal Gradient with smooth, non-sooth composite objective function. The following content are mostly summarized from Walkington [1], section 1, and section 4. They are supplemented by my own writings.

1.1 Signal Denoising in One Dimension

Let a one dimensional singal be $u : [0, 1] \mapsto \mathbb{R}$ and u experiences absolute continuity. This class of absolutely continuous function can model discrete signal because digital signal are piecewise constant. Let $\hat{u}$ denotes an observation of u corrupted by noise. The denoised signal is the minimizer of f(u),

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

A practical approach on modern computing devices would necessitate discritization of the integral. We use trapzidal 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, hence the formulation below is slightly more general than Walkington[1]. We derive the

approximation of the integral by doing

$$\text{Denote } s_i = u_i - \hat{u}_i,$$

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

The above formulation suggests smooth, non-smooth additive composite objective for $f(u)$. This type of optimization method can be solved via the Proximal Gradient method and its variants. Unfortunately the non-smooth part $\alpha\|Cu\|_1$ presents computational difficulty if matrix C is unfriendly for proximal resolvent operator. One way to bypass the difficulty involves reformulating with $p = Cu$, and solve 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 admit hyper box indicator function that makes the resolvent friendlier because proximal operator of indicator is projection, in the case of projecting onto hyper box, the operator is simple. Given dual variable λ, primal is obtained by

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

At this point, we had a formulation such that, solving $-g(u)$ is an easy task with the smooth non-smooth additive objective, and obtaining the primal solution is simple as well 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 propsed 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 here that for the class of strongly convex objectives, Proximal Gradient without acceleration has a better theoretical convergence results than the accelerated version. This surprising facts hints at a fundamental difference between methods with, and without Nesterov's acceleration. It sparks the discussion in this paper on the variants of FISTA in hope of providing some insights on the reasons for Nesterove's momentum based method's inability to adapt the convergence rate with objective has strong convexity. For the terminologies, we use FISTA to specifically refers to the proximal gradient method presented by Beck and Teboulle[3]. We use Accelerated Proximal Gradient method (APG) to refer to the class of first order acceleration algorithms developed/inspired from FISTA.

1.3 Outline of the Paper

Section 2 consists of 3 parts. Reviewing of iteratures on the problem of total variation minimization for image/signal denoising and deblurring is the first part. Presenting FISTA and its variants is the second part. Reviewing the algorithmic tricks and improvements applied to the APG is the third parts. Section 3 addresses a mistake made in Walkington's writing [1, theorem 2.4]. We will talk about what a first order method is and the fact that the lower complexity bound on the objective value and iterates for a fixed iteration, is achieved by a different function. We discuss how this 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 and it removed an equality to strengthens interpretibility and generality. Section 5 presents plots of convergence and results of applying variants of Accelerated Proximal Methods to the Total Variation problem.

2 Literatures Review

2.1 Total Variation Minimizations

Total Variation (TV) minimization method was introduced by Rudin-Osher and Fatemi in [4]. They pioneer the theories of TV minimizations by solving PDE. They discussed the empirical observation that L1 regularizations term produces sharper images. Walkington [1] a basic formulation of one dimensional signal denoising. However it's important 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 a L1 penalty terms for 2D signal recovery

is in [6]. For a full survey of state of arts computational methods applied to TV minimizations see Chambolle [7]. For a detailed exposition of mathematical theories regarding variational analysis on different type of TV problems and statistical inferences based interpretations of 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

For different Variants of FISTA, we specifically refer to first order acceleration method based on Nesterov's Framework, adhering to Nesterov's lectures [11] and Beck's book [2, chapter 10]. Nesterov's lectures focuses on establishing the most general frameworks for this type of methods. Nesterov approach doesn't handle function that is non-smooth, at least not directly. However, in his lectures, he derived a generic accelerated gradient algorithm (2.2.7) based on the idea of accelerating sequences.

Nesterov's derivation of the linear convergence rate, and sub-linear convergences rate, were completed 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 were done in lemma 2.2.4, without the assumption of a minimizer, and for Lipschitz smooth function with and without strong convexity. Amir Beck's book had the algorithm V-FISTA presented in 10.7.7, it's an algorithm that is exactly the same as Nesterov's algorithm: 2.2.22, with the addition of a proximal operator.

Algorithm 1 V-FISTA

1: **Input:** $(f, g, x^{(0)})$

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, "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