MATH 590 2023 FALL REPORT

HONGDA LI

November 15, 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 Introduction

In this section, We present and model the problem of denoising one dimension signal to motivate the use of Accelerated 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 Modeling for 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) defined as

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

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

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

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. This approach is possible when D is positive semi-definite, which in our case, it is.

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

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

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 Algorithmic Approach

The dual objective 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, it's highlighted in Walkington[1], the proximal gradient method without acceleration achieves $\mathcal{O}((1-1/\kappa)^k)$ convergence rate. The parameter κ is the condition number, in this case it would be L/α, where L is the lipschitz smooth constant of g and α is the strong convexity constant for g. 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 counter intuitive fact 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.

1.3 Outline of the Paper

In

2 Literatures Review

2.1 Subsections

3 Nesterov's Lower Bound Clarified

3.1 title

4 A Better Proof for FISTA Under Strong Convexity

4.1 Subsection

5 A Modified FISTA Under Strong Convexity

This is the Bleeh Bleeh Bleeh I am not Listening section.

6 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