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

Hongda Li

UBC Okangan

November 20, 2023

ToC

Paper Under Review

Nesterov's Method for Convex Optimization*

Noel J. Walkington[†]

Abstract. While Nesterov's algorithm for computing the minimum of a convex function is now over forty years old, it is rarely presented in texts for a first course in optimization. This is unfortunate since for many problems this algorithm is superior to the ubiquitous steepest descent algorithm, and it is equally simple to implement. This article presents an elementary analysis of Nesterov's algorithm that parallels that of steepest descent. It is envisioned that this presentation of Nesterov's algorithm could easily be covered in a few lectures following the introductory material on convex functions and steepest descent included in every course on optimization.

Key words. convex optimization, Nesterov's algorithm, steepest descent

MSC codes. 65K10, 60C46, 60C25

DOI. 10.1137/21M1390037

Noel J. Walkington, SIAM REVIEW Jun 2023 Education, Volume 65 Number 2, pp. 539-562. [1]

Presentation Outline and Objective

1. Introducing the Application of TV Minimization for Signal Recovery.
2. Literature Review.
3. Nesterov lower bound complexity claim clarified.
4. Our proof for V-FISTA convergence under strong convexity inspired by [2, 10.7.7]
5. Some exciting numerical results for our method, which we refer to as "The method of Spectral Momentum".

Total Variance Minimization Formulation

Total Variance Minimization (TV) problem recovers the digital signal from observations of a signal with noise. Let $u : [0, 1] \mapsto \mathbb{R}$ be the signal and $\hat{u}$ be a noisy observation, then

Variational Formulation

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

- Minimizing $f(u)$ with pentalty term constant $\alpha > 0$ yield a recovered signal.
- Original signal u is assumed to be piecewise constant with finite many pieces.
- Sparsity is imposed on u', making u' to be Dirac Delta function.

Discritizations

Implementations on modern computing platforms **necessitate discretization** of signal u to $\mathbb{R}^{N+1}$. With $s_i = u_i - \hat{u}_i$, $h_k = t_k - t_{k-1}, k \geq 1$ using the trapezoid rule and first-order forward difference yields:

$$\frac{1}{2}\int_0^1 (u - \hat{u})^2 dt + \alpha \int_0^1 |u'| dt \approx \frac{1}{2}\sum_{i=0}^{N}\left(\frac{s_i^2 + s_{i+1}^2}{2}\right) h_{i+1} + \alpha \sum_{i=1}^{N}\left|\frac{u_i - u_{i-1}}{h_{i+1}}\right|$$

$\triangleright$ let $C \in \mathbb{R}^{N\times(N+1)}$ be upper bi-diagonal with $(1, -1)$

$$= \frac{1}{2}\left(\frac{s_0^2 h_1}{2} + \frac{s_N^2 h_N}{2} + \sum_{i=1}^{N-1} s_i^2 h_i\right) + \alpha\|Cu\|_1$$

$\triangleright$ using $D \in \mathbb{R}^{N\times(N+1)}$,

$\triangleright$ $D := \text{diag}(h_1/2, h_1, h_2, \cdots, h_N, h_N/2)$

$$= \frac{1}{2}\langle u - \hat{u}, D(u - \hat{u})\rangle + \alpha\|Cu\|_1.$$

Discretized Model

Recall D is diagonal, strictly positive entry, $C \in \mathbb{R}^{N \times N+1}$ is bidiagonal.

Discretized Formulation

$$f(u) = \frac{1}{2}\langle u - \hat{u}, D(u - \hat{u})\rangle + \alpha\|Cu\|_1.$$

If we were to use the Forward-Backward(FB) splitting, then we have unresolved implementation difficulties:

1. ADMM, Chambolle Pock, would apply; however, when using the FB Splitting, $\alpha\|Cu\|_1$ would be prox unfriendly.
2. Prox over $\alpha\|Cu\|_1$ is possible with D being bi-diagonal, but it would be a hassle if done for generic C.

Discretized Model

Recall D is diagonal, strictly positive entry, $C \in \mathbb{R}^{N\times N+1}$ is bidiagonal.

Discretized Formulation

$$f(u) = \frac{1}{2}\langle u - \hat{u}, D(u - \hat{u})\rangle + \alpha\|Cu\|_1.$$

If we were to use the Forward-Backward(FB) splitting, then we have unresolved implementation difficulties:

1. ADMM, Chambolle Pock, would apply; however, when using the FB Splitting, $\alpha\|Cu\|_1$ would be prox unfriendly.
2. Prox over $\alpha\|Cu\|_1$ is possible with D being bi-diagonal, but it would be a hassle if done for generic C.

Remedy via Lagrangian 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 Problem is

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

So

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

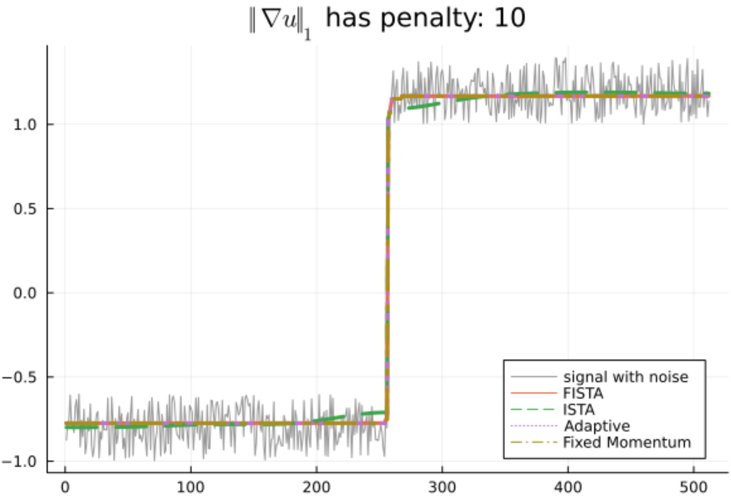
Dual

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

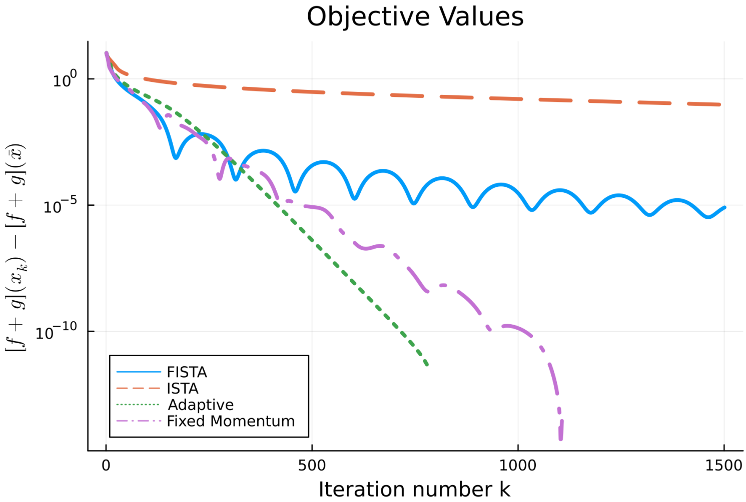
- Fact: $u = \hat{u} + D^{-1}C^T\lambda$, for the primal.
- D^{-1} is Positive Definite and diagonal, very easy to invert.
- $-g(\lambda)$ would be strongly convex.

Numerical Results

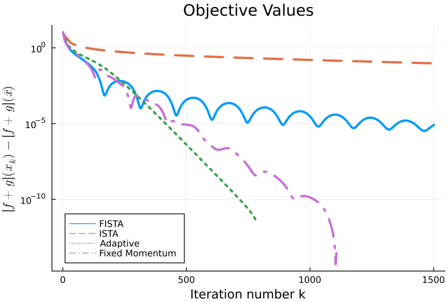
Implemented with Julia[3], with several variants of FISTA, we have



One Big Bummer



One Big Bummer



Main observations

1. FISTA is non-robust to strong convexity; it experiences the same $(1/k^2)$.
2. However, ISTA would be $\mathcal{O}(1-1/\kappa)^k)$ under strong convexity, with $\kappa = L/\sigma$, for L-Lipschitz smooth and σ strongly on the smooth part of the FB splitting objective.

Generic FISTA

We introduce the below algorithm 1 to expedite presentation. Consider Smooth, non-smooth Additive composite objective $F = g + h$.

A Good Template

Algorithm Generic FISTA

1: **Input:** $(g, h, x^{(0)})$
2: $y^{(0)} = x^{(0)}$, $\kappa = L/\sigma$
3: **for** $k = 0, 1, \cdots$ **do**
4: $\quad x^{(k+1)} = T_L y^{(k)}$
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**

Changing T_L, θ_{k+1}, and $\mathcal{S}$ yield different variants.

Variant (1.), FISTA Original

Algorithm Generic FISTA

1: **Input:** $(g, h, x^{(0)})$
2: $y^{(0)} = x^{(0)}$, $\kappa = L/\sigma$
3: **for** $k = 0, 1, \cdots$ **do**
4: $\quad x^{(k+1)} = T_L y^{(k)}$
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**

Original FISTA proposed by Beck [4] has

- $\theta_{k+1} = (t_k - 1)/t_{k+1}$, $t_{k+1}(t_{k+1} - 1) = t_k^2$, $t_0 = 1$.
- It achieves $\mathcal{O}(1/k^2)$ on the objective value; it doesn't improve for strongly convex function g.
- No known proof for the convergence of the iterates.

Variant (2.), Chambolle, Dossal

Algorithm 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

From Chamebolle, Dossal [5],

- $\theta_{k+1} = (n + a - 1)/a$, for $a > 2$.
- Proved in Chambolle, Dossal [5], its iterates of this version of FISTA exhibit weak convergence.
- T_L is the same as (1.); it experiences the same convergence rate for the function objective.

Variant (3.) (Also known as V-FISTA), From Beck

Algorithm 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

Proposed in Beck[2], and also Nesterov[6].

- has $\theta_{k+1} = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$.
- Only for strong convexity g, (Or the weaker condition such as Quadratic Growth).
- Has $\mathcal{O}((1 - 1/\sqrt{\kappa})^k)$, for both objective value and iterates.
- T_L is still the same FB splitting.

Motivates Variants (4.)

V-FISTA is simple but requires knowledge of σ, the strong convexity index; it is a troublemaker.

- Exact estimation would involve inverting the Hessian or a closed-form formula tailored for a specific problem.
- Over estimation of σ invalidates the linear convergence results.
- Under estimation of σ slows down the linear-convergence.

To overcome this, we propose a real-time overestimation of σ using

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

We call this **Spectral Momentum**. The same formula is used for spectral stepsizes, an adaptive stepsize scheme for gradient descent[7, 4.1]. Its efficacy was demonstrated at the start of the talk. We are not sure why it works so well.

MFISTA

Algorithm 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

MFISTA in Beck[8] is produced by adding $\mathcal{S}$ to be the procedure

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

MFISTA

MFISTA is an early attempt at improving FISTA. However, the idea of restarting refuses to die.

1. Requires frequent computing of the objective, slowing it by a constant factor compared to FISTA.
2. There is no better convergence rate than $\mathcal{O}(1/k^2)$ from our research.

FISTA Restart Refuses to Die

In our opinion, the interests gather around restarting FISTA because spending too much computational effort would compete against the Proximal Quasi-Newton method, questioning the use of momentum in the first place. Hence, the recent development

- Asymptotic linear convergence of conditional restart by Alamo et al. [9][10, text] et al., and Fercoq [11], under strong convexity, or local quadratic growth.
- Parameter Free FISTA with automatical restart by Aujol et al.[12], fast linear convergence on quadratic growth with proofs, no parameters needed.

Theories of Nesterov Momentum

Frontier developments of theories for Nesterov Momentum are not dying either.

References I

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

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

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, Jan. 2017, publisher: Society for Industrial and Applied Mathematics. [Online]. Available: https://epubs.siam.org/doi/10.1137/141000671

References II

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

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

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

References III

- 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
- 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
- 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

References IV

——, "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

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

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