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

Hongda Li

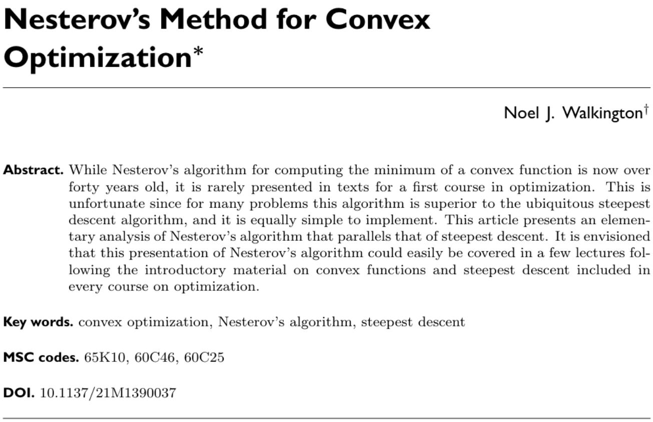
UBC Okangan

November 20, 2023

# ToC

1. Introduction
2. TV Minimizations
3. Literature Review
4. Nesterov Lower Bound
5. V-FISTA Under Strong Convexity
6. Numerical Results
7. References

# 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:

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

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

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

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

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

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

# 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

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

So

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

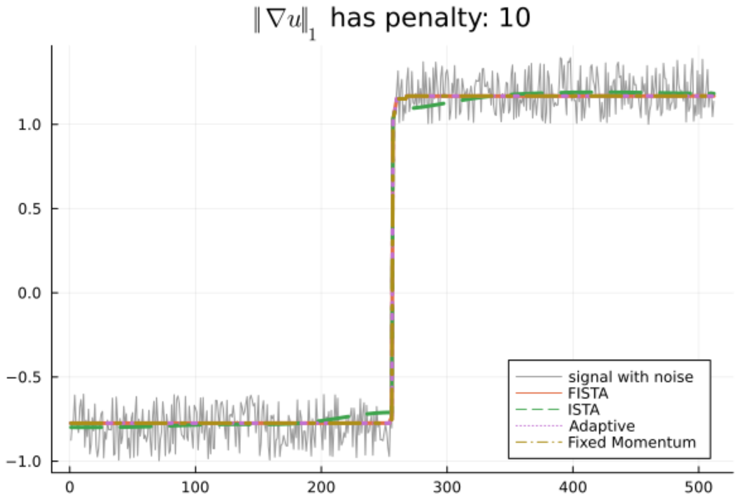
# Dual

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

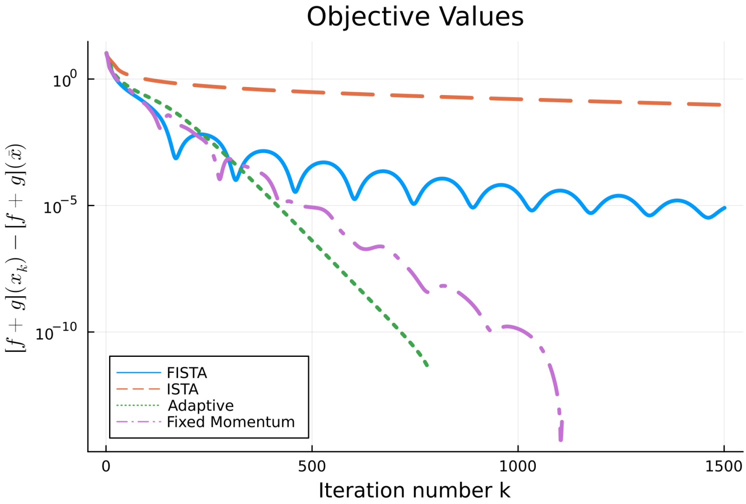
- Fact: $u = \hat{u} + D^{-1}C^T\lambda$, for the primal.
- $D^{-1}$ is Positive Definite and diagonal, very easy to invert.
- $-g(\lambda)$ would be strongly convex.

# Numerical Results

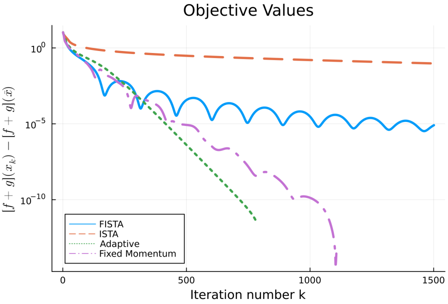
Implemented with Julia[3], with several variants of FISTA, we have



# 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 $\sigma$ strongly on the smooth part of the FB splitting objective.

# Generic FISTA

We introduce the below algorithm 1 to expedite presentation.

**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
```

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

# References

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