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

Hongda Li

UBC Okangan

November 27, 2023

Presentation Overview

1. Paper Under Review
2. TV Minimizations
3. Literature Review
4. Nesterov Lower Bound
5. V-FISTA Under Strong Convexity
6. 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]

Formulation of Total Variational Minimization

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

Variational Formulation

$$f(u) = \int_0^1 \frac{1}{2}(u(t) - \hat{u}(t))^2 + \alpha|u'(t)|dt.$$

- Minimizing $f(u)$ with pentalty term constant $\alpha > 0$ yield a recovered signal.
- Penalization helps recover an original signal u that is assumed to be piecewise constant with bounded variation.

Formulation of Total Variational Minimization

Variational Formulation

$$f(u) = \int_0^1 \frac{1}{2}(u(t) - \hat{u}(t))^2 + \alpha|u'(t)|dt.$$

Solving the problem on modern computing platforms requires discretization of the problem. Here is a list of helpful quantities for discretized formulation.

- Let $u = [u_0 \ \cdots \ u_N]^T \in$, be a vector. This is the variable we optimize to get a solution.
- Let $t_0 < \cdots < t_i < \cdots < t_N$ be a sequence of time for each observed value: u_i.
- Let $h_i = t_i - t_{i-1}$ be the elapse of time between u_i, u_{i-1}, for all $i \geq 1$.
- Let $C \in \mathbb{R}^{N\times(N+1)}$ be a bi-diagonal matrix with $-1/h_i, 1/h_i$ on the diagonal and the super diagonal. This is a finite difference matrix.
- Let $D = \text{diag}(h_1/2, h_1, \cdots, h_N, h_N/2) \in \mathbb{R}^{(N+1)\times(N+1)}$.

Discretized Model

The following would be the discretized model formulation. [1] 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. Hence we reformulate and use the lagrangian dual.

[1] See our report for the omitted details.

Discretized Model

The following would be the discretized model formulation. [1] 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. Hence we reformulate and use the lagrangian dual.

[1] See our report for the omitted details.

Dual

The Dual Reformulation

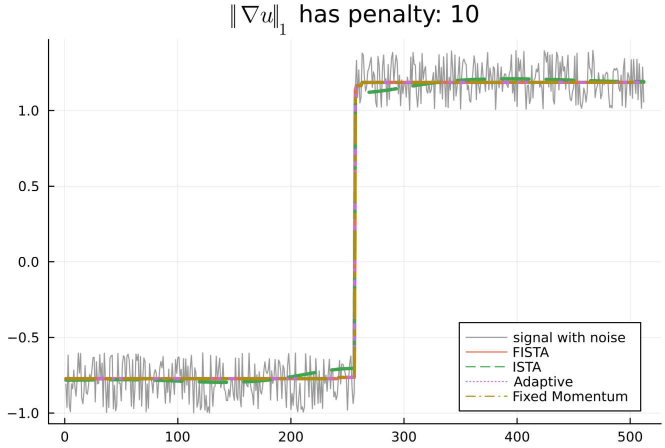
$$-g(\lambda) = -\frac{1}{2}\|C^T\lambda\|_{D^{-1}}^2 - \langle \hat{u}, C^T\lambda\rangle - \delta_{[-\alpha,\alpha]^N}(\lambda).$$

See my report for omitted details.

- 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.

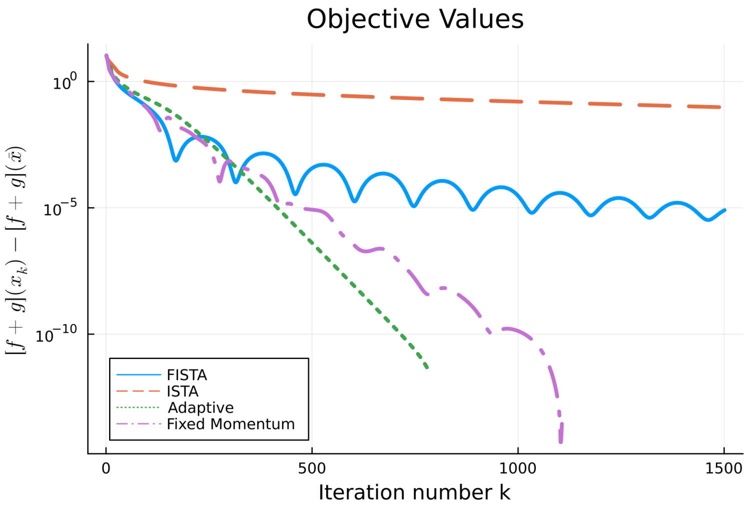
Numerical Results

We implemented in Julia [2], with several variants of FISTA, producing

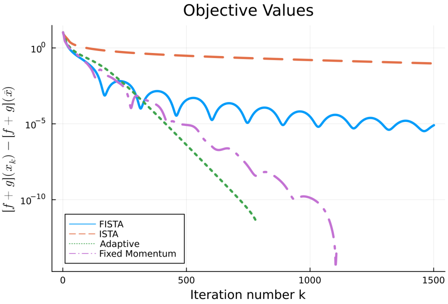


See GitHub repository here for our code implementation. See our report for more numerical experiments.

Convergence Results



One Big Bummer



Main observations

1. FISTA is non-robust to strong convexity; the above suggests that it experiences the same $\mathcal{O}(1/k^2)$.
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: $\quad x^{(k+1)} = T_L y^{(k)}$; $T_L(x) = \mathrm{prox}_{L^{-1}h}(x - L^{-1}\nabla g(x))$
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

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

Original FISTA proposed by Beck [3] 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.
- $T_L(x) = \text{prox}_{L^{-1}h}(x - L^{-1}\nabla g(x))$.

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 Chambolle, Dossal [4],

- $t_{k+1} = (n + a - 1)/a, \theta_{k+1} = (t_k - 1)/t_{k+1}$, for $a > 2$.
- Proved in Chambolle, Dossal [4, thm 4.1], its iterates of this version of FISTA exhibit 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 [5], and also Nesterov [6].

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

Spectral Momentum 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 [7, 4.1], also known as (Barzilai-Borwein method) when applied to gradient descent. However, we used it on the momentum term. Its efficacy was demonstrated at the start of the talk. We are not sure why it works so well.

MFISTA Variants (5.)

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**

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

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

Recall $\theta_{k+1} = (t_k - 1)/t_{k+1},\ t_{k+1}(t_{k+1} - 1) = t_k^2,\ t_0 = 1.$

MFISTA Drawbacks

MFISTA was an early attempt at improving FISTA.

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

However, the idea of restarting FISTA refuses to go away.

- Asymptotic linear convergence of conditional restart by Alamo et al. [9][10] et al., and Fercoq [11], under strong convexity, or local quadratic growth.
- Parameter Free FISTA with automatic restart by Aujol et al. [12], linear convergence $\mathcal{O}(\exp(-K\sqrt{\kappa}n))$ (n is the iteration counter) on quadratic growth with proofs, no parameters needed.

Theories of Nesterov Momentum

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

1. Su et al. [13] identified a second-order ODE that is the exact limit of the variant (1.).
2. In Attouch and Peypouquet [14], they showed a $o(1/k^2)$ convergence rate of variant (2.) based on the ODE understanding.
3. In Nesterov [6], he proposed a generic algorithm that can derive variants (1.), (3.), and more.
4. In Ahn and Sra [15] they derived a lot of variants of APGs using the Proximal Point method of Rockafellar and discussed the unified theme of a "similar triangle."

The Most Recent and Hardcore Developement

The paper is titled "Computer-Assisted Design of Accelerated Composite Optimization Methods: OptISTA" by Jang et al. [16]. They updated the manuscript On Nov 1st, 2023.

- They also proved a lower bound for their OptISTA, showing its exact optimality.
- They claim they had concluded the search for the fastest first-order method.

That leads to the question: "Why do we call Nesterov's acceleration method optimal?" What is optimality for an algorithm?

First-Order Methods

What we can do is to understand at least Nesterov's claim on lower complexity bound and why we believe there is a mistake in Walkington [theorem 2.4][1].

Definition (First Order Method)

We are in $\mathbb{R}^n$ for now. Given $x^{(0)} \in \mathbb{R}^n$, an iterative algorithm generates sequence of $\left(x^{(n)}\right)_{n\in\mathbb{N}}$ in the space. All $\mathcal{A} \in \text{GA}^{\text{1st}}$ satisfy that

$$x^{(j+1)} := \mathcal{A}_f^{j+1} x^{(0)} \in \left\{x^{(0)}\right\} + \operatorname{span}\left\{\nabla f\left(x^{(i)}\right)\right\}_{i=1}^{j} \quad \forall f, \forall 1 \le j \le k-1.$$

We adapted the above definition from Nesterov [6, 2.1.4]. We came up with two examples for the definition.

Fixed-Step Descent

Example (Fixed Step Descent)

The method of fixed-step gradient descent, $x^{(k+1)} = x^{(k)} - L^{-1}\nabla f(x^{(k)})$ is $\bar{\mathcal{A}} \in \mathrm{GA}^{1\mathrm{st}}$ achieves a maximal decrease in objective value for all $f \in \mathcal{F}_L^{1,1}$ for some $x^{(k)} \in \mathbb{R}^n$, it can be understood as

$$\bar{\mathcal{A}} \in \underset{\mathcal{A}\in \mathrm{GA}^{1\mathrm{st}}}{\operatorname{argmin}} \max_{f\in\mathcal{F}_L^{1,1}} \left\{ f\left(\mathcal{A}_f x^{(k)}\right)\right\}. \quad \triangleright \text{ Only a single iteration.}$$

Example (Steepest Descent)

Fix some $f, x^{(k)}$, the method of steepest descent would be $\bar{\mathcal{A}} \in \mathrm{GA}^{1\mathrm{st}}$ and it's

$$\bar{\mathcal{A}} \in \underset{\mathcal{A}\in \mathrm{GA}^{1\mathrm{st}}}{\operatorname{argmin}} \left\{ f\left(\mathcal{A}_f x^{(k)}\right)\right\}.$$

Nesterov Lower Complexity Bound

The following is Nesterov [6, Thm 2.1.7].

Theorem (Nesterov's Claim of Lower Bound)

For any $1 \leq k \leq (n-1)/2$, for all $x^{(0)} \in \mathbb{R}^n$, there exists a Lipschitz smooth convex function in $\mathbb{R}^n$ such that for all algorithm from GA^{1st}, we have the lower bound for the optimality gap for the function values and its iterates:

$$f\left(x^{(k)}\right) - f^* \geq \frac{3L\|x - x^*\|^2}{32(k+1)^2}, \quad \|x^{(k)} - x^*\|^2 \geq \frac{1}{8}\|x^{(0)} - x^*\|^2.$$

Where x^* is the minimizer of f, so that $f(x^*) = \inf_x f(x)$.

We emphasize that it didn't fix the function f, but it fixes the iteration counter k.

Walkington's claim of Lower Bound

We now quote Walkington [1, theorem 2.4]

Theorem (Walkington's Claim of Lower Bound)

Let X be an infinite-dimensional Hilbert Space and set $x^{(0)} = \mathbf{0}$. There exists a convex function $f : X \mapsto \mathbb{R}$ with Lipschitz gradient and minimum $f(x_*) > -\infty$ such that for any sequence satisfying

$$x_{i+1} \in Span\left\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \cdots, \nabla f(x^{(i)})\right\}, \quad i = 0, 1, 2, \cdots,$$

there holds

$$\min_{1 \le i \le n} f(x_i) - f(x_*) \ge \frac{3L\|x_1 - x_*\|^2}{32(n+1)^2},$$

where L is the Lipschitz constant of the gradient.

The Differences Between Their Claims

- Walkingon claimed there exists a single function from $\mathcal{F}_L^{1,1}(\mathcal{H})$ that introduces the lower bound for all values of k, and all algorithms from $\mathrm{GA}^{1\mathrm{st}}$, but the latter didn't claim that.
- The difference would remain in infinite dimension Hilbert space if we generalize Nesterov's lower bound claim.
- No contradiction from Attouch and Peypouquet [14] on $o(1/k^2)$ because they fixed Φ in the proof.

There is no proof after Walkington's claim; we can't know if he had his way of proving the latter claim. It makes us think it is likely a missed detail in his writing.

Preparing for Strong Convexity Convergence of V-FISTA

To understand the proof, we introduce the following quantities for simplicity.

1. $s^{(k)} = x^{(k)} - x^{(k-1)}$, the velocity vector, for all $k \geq 1$.
2. $e^{(k)} = x^{(k)} - \bar{x}$, the error vector at the kth iteration, for $k \geq 0$.
3. $\theta_k = (t_k - 1)/(t_k + 1)$, which is the momentum step size.
4. $u^{(k)} = \bar{x} + t_k(x^{(k-1)} - x^{(k)}) - x^{(k-1)}$, the error term extrapolated with the velocity. We take $u^{(0)} = \bar{x} - x^{(0)}$.
5. $\delta_k = f(x^{(k)}) - f_{\text{opt}}$, with $f_{\text{opt}} = f(\bar{x}) = \inf_x f(x)$.
6. The quantity R_k plays a crucial role in the proof; it's
$$R_k = \frac{\sigma(t_{k+1} - 1)}{2} \left\| x^{(k)} - \bar{x} \right\|^2 - \frac{L - \sigma}{2} \left\| \bar{x} + t_{k+1} \left(x^{(k)} - y^{(k)} \right) - x^{(k)} \right\|^2$$
7. $\kappa = L/\sigma$, the condition number, it would be that $\kappa \geq 0$.
8. $q = \sigma/L$, the reciprical of the condition number. It would be that $q \in (0, 1)$.

V-FISTA Generic Convergence

Lemma (Generic Convergence Results)

If there exists a sequence $(t_k)_{k\in\mathbb{N}}, C_k$ *such that*

$$\begin{cases} \frac{C_k}{t_{k+1}^2} = \frac{L(1-t_{k+1}^{-1})}{2t_k^2} \\ R_k + C_k \left\| u^{(k)} \right\|^2 \geq 0, \end{cases} \tag{1}$$

then

$$\delta_{k+1} \leq \left(\prod_{i=0}^{k} (1 - t_k^{-1}) \right) \left(\delta_0 + \frac{L}{2t_0^2} \left\| u^{(0)} \right\|^2 \right).$$

Proof.

In the appendix of the report. □

A Sufficient Condition for Convergence

Proposition (Sufficient Conditions for Generic Convergence)

If the sequence $(t_k)_{k\geq\mathbb{N}}$ by t_k satisfies

1. $t_{k+1} = 1 + \frac{t_k^2(1-q)}{t_k+1}$,
2. and $t_{k+1} \geq t_k + 1 - t_k^2 q$,
3. and $t_k > 1$ for all $k \geq 0$,

make lemma 4 (Generic Convergence Results) true.

Proof.

In the appendix of the report. □

Fixed Momentum Convergence

Proposition

The choice of $t_k = t_{k+1} = \sqrt{\frac{L}{\sigma}}$ for all $k \geq 0$, then proposition (Sufficient Conditions for Generic Convergence) is true. Hence, the V-FISTA variant (3.) has a convergence rate bound

$$\delta_{k+1} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(\delta_0 + \frac{L}{2t_0^2}\left\|u^{(0)}\right\|^2\right).$$

Proof.

we have

$$\frac{t_{k+1}-1}{t_k^2(1-q)} - \frac{1}{t_k+1} = 0 \quad \triangleright \text{ by } t_{k+1} = t_k, \text{ then}$$

$$\frac{t-1}{t^2(1-q)} = \frac{1}{t+1}$$

$$t^2(1-q) = t^2 - 1$$

$$t^2 q = 1$$

$$t = \pm\sqrt{\frac{L}{\sigma}},$$

with $t_k > 1$ it has to be $t_k = \sqrt{\frac{L}{\sigma}}$ for all $k \geq 0$. With that we have $t_{k+1} = t_k + 1 - t_k^2 q = t_k + 1 - 1 = t_k$, hence the inequality is also satisfied. $\square$

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

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

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

References II

- 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
- 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
- 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

W. Su, S. Boyd, and E. J. Candes, “A Differential Equation for Modeling Nesterov’s Accelerated Gradient Method: Theory and Insights,” *arXiv.org*, Mar. 2015. [Online]. Available: https://arxiv.org/abs/1503.01243v2

H. Attouch and J. Peypouquet, “The Rate of Convergence of Nesterov’s Accelerated Forward-Backward Method is Actually Faster Than \$1/k^2\$,” *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1824–1834, Jan. 2016, publisher: Society for Industrial and Applied Mathematics. [Online]. Available: https://epubs.siam.org/doi/10.1137/15M1046095

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

References VI

U. Jang, S. D. Gupta, and E. K. Ryu, "Computer-Assisted Design of Accelerated Composite Optimization Methods: OptISTA," May 2023, arXiv:2305.15704 [math]. [Online]. Available: http://arxiv.org/abs/2305.15704