

# Proximal Gradient: Convergence, Implementations and Applications

Hongda Li

UBC Okanagan

November 24, 2022

## 1 Introduction and Proximal Operators

- Taxonomy of Proximal type of Methods
- The Proximal Operator
- Strong Smoothness
- A Major Assumption

## 2 Envelope and Secret Sauce

- Forward Backwards Envelope
- The Proximal Gradient Algorithm
- The Secret Sauce

## 3 Numerical Experiments

- LASSO
- Image Deconvolution with Noise

## 4 Conclusion

## 5 References

# Sum of 2 Functions

## Sum of 2

$$\min_x g(x) + h(x) \quad (1)$$

1. Throughout this presentation, we assume the objective of a function  $f$  is the sum of 2 functions.
2. We are interested in the paper: FISTA (Fast Iterative-Shrinkage Algorithm) by Beck and Teboulle [1].
1. When  $h = \delta_Q$  with  $Q$  closed and convex with  $Q \subseteq \text{ri} \circ \text{dom}(h)$ , we use projected subgradient.
2. When  $g$  is **strongly smooth** and  $h$  is **closed convex proper** whose proximal oracle is easy to compute, we consider the use of FISTA.

## What is FISTA

Simply put, the FISTA algorithm is the non-smooth analogy of gradient descent with Nesterov Momentum.

We will be going over these things in the presentations.

1. Derive the proximal gradient operator under standard convexity and regularity assumptions for the function  $g, h$ .
2. State one important lemma that arose during the proof for the proximal gradient method that is later useful for the proof of the FISTA.
3. Derive the FISTA algorithm and construct the sequence of the Nesterov Momentum during the proof using a template algorithm.
3. Numerical Experiments!

## What is FISTA

Simply put, the FISTA algorithm is the non-smooth analogy of gradient descent with Nesterov Momentum.

We will be going over these things in the presentations.

1. Derive the proximal gradient operator under standard convexity and regularity assumptions for the function  $g, h$ .
2. State one important lemma that arose during the proof for the proximal gradient method that is later useful for the proof of the FISTA.
3. Derive the FISTA algorithm and construct the sequence of the Nesterov Momentum during the proof using a template algorithm.
3. Numerical Experiments!

## What is FISTA

Simply put, the FISTA algorithm is the non-smooth analogy of gradient descent with Nesterov Momentum.

We will be going over these things in the presentations.

1. Derive the proximal gradient operator under standard convexity and regularity assumptions for the function  $g, h$ .
2. State one important lemma that arose during the proof for the proximal gradient method that is later useful for the proof of the FISTA.
3. Derive the FISTA algorithm and construct the sequence of the Nesterov Momentum during the proof using a template algorithm.
3. Numerical Experiments!

## What is FISTA

Simply put, the FISTA algorithm is the non-smooth analogy of gradient descent with Nesterov Momentum.

We will be going over these things in the presentations.

1. Derive the proximal gradient operator under standard convexity and regularity assumptions for the function  $g, h$ .
2. State one important lemma that arose during the proof for the proximal gradient method that is later useful for the proof of the FISTA.
3. Derive the FISTA algorithm and construct the sequence of the Nesterov Momentum during the proof using a template algorithm.
3. Numerical Experiments!

# Proximal Operator Definition

## Definition

The Proximal Operator Let  $f$  be convex closed and proper, then the proximal operator parameterized by  $\alpha > 0$  is a non-expansive mapping defined as:

$$\text{prox}_{f,\alpha}(x) := \arg \min_y \left\{ f(y) + \frac{1}{2\alpha} \|y - x\|^2 \right\}.$$



# Prox is the Resolvent of Subgradient

## Lemma (Resolvent of the Subgradient)

*When the function  $f$  is convex closed and proper, the  $\text{prox}_{\alpha, f}$  can be viewed as the following operator  $(I + \alpha \partial f)^{-1}$ .*

Proof.

$$0 \in \partial \left[ f(y) + \frac{1}{2\alpha} \|y - x\|^2 \right] (y^+)$$

$$0 \in \partial f(y^+) + \frac{1}{\alpha} (y^+ - x)$$

$$\frac{x}{\alpha} \in (\partial f + \alpha^{-1} I)(y^+)$$

$$x \in (\alpha \partial f + I)(y^+)$$

$$y \in (\alpha \partial f + I)^{-1}(x).$$



# An Example of Prox

## Definition (Soft Thresholding)

For some  $x \in \mathbb{R}$ , the proximal operator of the absolute value is:

$$\text{prox}_{\lambda \|\cdot\|_1, t}(x) = \text{sign}(x) \max(|x| - t\lambda, 0).$$

One could interpret the sign operator as projecting  $x$  onto the interval  $[-1, 1]$  and the  $\max(|x| - t\lambda, 0)$  as the distance of the point  $x$  to the interval  $[-t\lambda, t\lambda]$ .

## Definition (Strong Smoothness)

A differentiable function  $g$  is called strongly smooth with a constant  $\alpha$  then it satisfies:

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq \frac{\alpha}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{E}. \quad (2)$$

## Remark

The absolute value sign can be removed and replaced with  $0 \leq$  on the left when the function  $g$  is a convex function.

# Equivalence of Strong Smoothness and Lipschitz Gradient

## Theorem (Lipschitz Gradient Equivalence under Convexity)

*Suppose  $g$  is differentiable on the entire of  $\mathbb{E}$ . It is closed convex proper. It is strongly smooth with parameter  $\alpha$  if and only if the gradient  $\nabla g$  is globally Lipschitz continuous with a parameter of  $\alpha$  and  $g$  is closed and convex.*

$$\|\nabla g(x) - \nabla g(y)\| \leq \alpha \|x - y\| \quad \forall x, y \in \mathbb{E}$$

## Proof.

Using line integral, we can prove Lipschitz gradient implies strong smoothness without convexity. The converse requires convexity and applying generalized Cauchy Inequality to (iv) in Theorem 5.8 for Beck's textbook [1]. □

# A Major Assumption

## Assumption (Convex Smooth Nonsmooth with Bounded Minimizers)

*We will assume that  $g : \mathbb{E} \mapsto \mathbb{R}$  is **strongly smooth** with constant  $L_g$  and  $h : \mathbb{E} \mapsto \bar{\mathbb{R}}$  is **closed convex and proper**. We define  $f := g + h$  to be the summed function and  $ri \circ \text{dom}(g) \cap ri \circ \text{dom}(h) \neq \emptyset$ . We also assume that a set of minimizers exists for the function  $f$  and that the set is bounded. Denote the minimizer using  $\bar{x}$ .*

# Envelope and Upper Bounding Functions

## Upper Bounding Function

With assumption 1, we construct upper bounding functions at the point  $x$  evaluated at  $y$  for the function  $f$ , and it is given by:

$$g(x) + \nabla g(x)^T(y - x) + \frac{\beta}{2}\|y - x\|^2 + h(y) =: m_x(y|\beta) \quad \forall y \in \mathbb{E},$$

In brief, suppose we are at the point  $x$  of the iterations; we are minimizing the function  $m_x(y|\beta)$  to obtain the next point for our iterations.

## Theorem (Minimizer of the Envelope)

*The minimizer for the envelope has a closed form, and it is  $\text{prox}_{h,\beta^{-1}}(x + \beta^{-1}\nabla g(x))$ , with assumption 1.*

# The Prox Gradient Operator

## Proof.

Minimizer of the Envelope We consider minimizing the envelope; zero is in the subgradient of the upper bounding function  $m_x(y|\beta)$ .

$$\mathbf{0} \in \nabla g(x) + \beta(y - x) + \partial h(y)$$

$$\nabla g(x) + \beta x \in \beta y + \partial h(y)$$

$$-\beta^{-1} \nabla g(x) + x \in y + \beta^{-1} \partial h(y)$$

$$-\beta^{-1} \nabla g(x) + x \in [I + \beta^{-1} \partial h](y)$$

$$\implies [I + \beta^{-1} \partial h]^{-1}(-\beta^{-1} \nabla g(x) + x) \ni y,$$

recall lemma 2, it's the operator  $\text{prox}_{h, \beta^{-1}}(x + \beta^{-1} \nabla g(x))$ . □

# Prox Step and the Proximal Gradient Algorithm

## The Prox Step

For simplicity we will be calling the point  $\text{prox}_{h,\beta^{-1}}(x + \beta^{-1}\nabla g(x))$  “the prox step”, and we denote it as  $\mathcal{P}_{\beta^{-1}}^{g,h}(x)$  when there is no ambiguity we simply use  $\mathcal{P}x$ .

## The Proximal Gradient Method

---

### Algorithm Proximal Gradient With Fixed Step-sizes

---

```
1: Input:  $g, h$ , smooth and nonsmooth,  $L$  stepsize,  $x^{(0)}$  an initial guess of solution.  
2: for  $k = 1, 2, \dots, N$  do  
3:    $x^{(k+1)} = \mathcal{P}_{L^{-1}}^{g,h}x^{(k)}$   
4:   if  $x^{(k+1)}, x^{(k)}$  close enough then  
5:     Break  
6:   end if  
7: end for
```

---



# The Proximal Gradient Method

1. (Proved in my report.) It converges *Monotonically* for stepsize  $L \geq L_g$ .
2. (Proved in my report.) It has a convergence rate of  $\mathcal{O}(1/k)$  on the optimality gap  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  where  $\bar{x}$  is one of the minimizers for  $f$  satisfying assumption 1.
3. A line search routine can be applied, and the stepsize should satisfy the conditions:  $m_x(\mathcal{P}_x|L) \leq f(x)$ .

# The Proximal Gradient Method

1. (Proved in my report.) It converges *Monotonically* for stepsize  $L \geq L_g$ .
2. (Proved in my report.) It has a convergence rate of  $\mathcal{O}(1/k)$  on the optimality gap  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  where  $\bar{x}$  is one of the minimizers for  $f$  satisfying assumption 1.
3. A line search routine can be applied, and the stepsize should satisfy the conditions:  $m_x(\mathcal{P}_x|L) \leq f(x)$ .

# The Proximal Gradient Method

1. (Proved in my report.) It converges *Monotonically* for stepsize  $L \geq L_g$ .
2. (Proved in my report.) It has a convergence rate of  $\mathcal{O}(1/k)$  on the optimality gap  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  where  $\bar{x}$  is one of the minimizers for  $f$  satisfying assumption 1.
3. A line search routine can be applied, and the stepsize should satisfy the conditions:  $m_x(\mathcal{P}_x|L) \leq f(x)$ .

# The Accelerated Proximal Gradient Method

## Momentum Template Method

---

**Algorithm** Template Proximal Gradient Method With Momentum

---

- 1: **Input:**  $x^{(0)}, x^{(-1)}, L, h, g$ ; 2 initial guesses and stepsize  $L$
  - 2:  $y^{(0)} = x^{(0)} + \theta_k(x^{(0)} - x^{(-1)})$
  - 3: **for**  $k = 1, \dots, N$  **do**
  - 4:    $x^{(k)} = \text{prox}_{h, l^{-1}}(y^{(k)} + l^{-1}\nabla g(y^{(k)})) =: \mathcal{P}y^{(k)}$
  - 5:    $y^{(k+1)} = x^{(k)} + \theta_k(x^{(k)} - x^{(k-1)})$
  - 6: **end for**
-

## Lemma (Prox Step 2 Points)

*With assumption 1, and  $\beta^{-1} > L_g$  still being our stepsize for algorithm 1, let  $y \in \mathbb{E}$  and define  $y^+ = \mathcal{P}_{\beta^{-1}}^{g,h}(y)$  we have for any  $x \in \mathbb{E}$ :*

$$f(x) - f(y^+) \geq \frac{\beta}{2} \|y^+ - y\|^2 + \beta \langle y - x, y^+ - y \rangle.$$

1. It is equivalent to the lemma 2.3 in the FISTA paper[1]
2. Proof for convergence without and without the momentum uses this lemma.

## Lemma (Prox Step 2 Points)

*With assumption 1, and  $\beta^{-1} > L_g$  still being our stepsize for algorithm 1, let  $y \in \mathbb{E}$  and define  $y^+ = \mathcal{P}_{\beta^{-1}}^{g,h}(y)$  we have for any  $x \in \mathbb{E}$ :*

$$f(x) - f(y^+) \geq \frac{\beta}{2} \|y^+ - y\|^2 + \beta \langle y - x, y^+ - y \rangle.$$

1. It is equivalent to the lemma 2.3 in the FISTA paper[1]
2. Proof for convergence without and without the momentum uses this lemma.

# Some Physical Quantities

1.  $v^{(k)} = x^{(k)} - x^{(k-1)}$  is the velocity term.
2.  $\bar{v}^{(k)} = \theta_k v^{(k)}$  is the weighed velocity term.
3.  $e^{(k)} := x^{(k)} - \bar{x}$ , where  $\bar{x} \in \arg \min_x (f(x))$ , where  $\bar{x}$  might not be unique.
4.  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  which represent the optimality gap at step  $k$ .

# Some Physical Quantities

1.  $v^{(k)} = x^{(k)} - x^{(k-1)}$  is the velocity term.
2.  $\bar{v}^{(k)} = \theta_k v^{(k)}$  is the weighed velocity term.
3.  $e^{(k)} := x^{(k)} - \bar{x}$ , where  $\bar{x} \in \arg \min_x (f(x))$ , where  $\bar{x}$  might not be unique.
4.  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  which represent the optimality gap at step  $k$ .



# Some Physical Quantities

1.  $v^{(k)} = x^{(k)} - x^{(k-1)}$  is the velocity term.
2.  $\bar{v}^{(k)} = \theta_k v^{(k)}$  is the weighed velocity term.
3.  $e^{(k)} := x^{(k)} - \bar{x}$ , where  $\bar{x} \in \arg \min_x (f(x))$ , where  $\bar{x}$  might not be unique.
4.  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  which represent the optimality gap at step  $k$ .

# Some Physical Quantities

1.  $v^{(k)} = x^{(k)} - x^{(k-1)}$  is the velocity term.
2.  $\bar{v}^{(k)} = \theta_k v^{(k)}$  is the weighed velocity term.
3.  $e^{(k)} := x^{(k)} - \bar{x}$ , where  $\bar{x} \in \arg \min_x (f(x))$ , where  $\bar{x}$  might not be unique.
4.  $\Delta_k := f(x^{(k)}) - f(\bar{x})$  which represent the optimality gap at step  $k$ .

Substitute  $x = x^{(k)}, y = y^{(k+1)}$  to Prox Step 2 Points

$$f(x^{(k)}) - f \circ \mathcal{P}_Y^{(k+1)} \geq \frac{L}{2} \| \mathcal{P}_Y^{(k+1)} - y^{(k+1)} \|^2 + L \langle y^{(k+1)} - x^{(k)}, \mathcal{P}_Y^{(k+1)} - y^{(k+1)} \rangle$$

$$[*1] \implies 2L^{-1}(\Delta_k - \Delta_{k+1}) \geq \|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle x^{(k+1)} - y^{(k+1)}, y^{(k+1)} - x^{(k)} \rangle$$

$$[*2] \implies 2L^{-1}(\Delta_k - \Delta_{k+1}) \geq \|v^{(k+1)} - \bar{v}^{(k)}\|^2 + 2\langle v^{(k+1)} - \bar{v}^{(k)}, \bar{v}^{(k)} \rangle \quad (*)$$

Substitute  $x = \bar{x}, y = y^{(k+1)}$  to Prox Step 2 Points

$$-2L^{-1}\Delta_{k+1} \geq \|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)} \rangle$$

$$-2L^{-1}\Delta_{k+1} \geq \|v^{(k+1)} - \bar{v}^{(k)}\|^2 + 2\langle v^{(k+1)} - \bar{v}^{(k)}, e^{(k)} + \bar{v}^{(k)} \rangle. \quad (*)$$

## Linear Combinations

Assume some linear combinations of the term  $(*)$ ,  $(\star)$  with:  $(t_k - 1) \geq 0$  for all  $k$ , then  $(t_{k+1} - 1)(*) + (\star)$  is:

$$\begin{aligned} & 2L^{-1}((t_{k+1} - 1)\Delta_k - t_{k+1}\Delta_{k+1}) \\ & \geq t_{k+1} \|\mathbf{v}^{(k+1)} - \bar{\mathbf{v}}^{(k)}\|^2 + 2\langle t_{k+1}(\mathbf{v}^{(k+1)} - \bar{\mathbf{v}}^{(k)}), \mathbf{e}^{(k)} + t_{k+1}\bar{\mathbf{v}}^{(k)} \rangle, \quad (**) \end{aligned}$$

1. No more monotonicity property.
2. The quantity on the right side bounds the weight differences of  $\Delta_k, \Delta_{k+1}$ .
3. **What if the expression can match the form of**  
 $a_k - a_{k+1} \geq b_{k+1} - b_k \quad \forall k \in \mathbb{N}?$

## 2 Sequences

### Lemma

*2 Bounded Sequences Consider the sequences  $a_k, b_k \geq 0$  for  $k \in \mathbb{N}$  with  $a_1 + b_1 \leq c$ . Inductively the two sequences satisfy  $a_k - a_{k+1} \leq b_{k+1} - b_k$ , which describes a sequence with oscillations bounded by the difference of another sequence. Consider the telescoping sum:*

$$\begin{aligned} a_k - a_{k+1} &\geq b_{k+1} - b_k \quad \forall k \in \mathbb{N} \\ \implies -\sum_{k=1}^N a_{k+1} - a_k &\geq \sum_{k=1}^N b_{k+1} - b_k \\ -(a_{N+1} - a_1) &\geq b_{N+1} - b_1 \\ c \geq a_1 + b_1 &\geq b_{N+1} + a_{N+1} \\ \implies c &\geq a_{N+1}. \end{aligned}$$

## We can Match the Template to That Form

Yes, it does, and it is:

$$\begin{aligned}
 & 2L^{-1}((t_{k+1}^2 - t_{k+1})\Delta_k - t_{k+1}^2\Delta_{k+1}) \\
 & \geq t_{k+1}^2\|v^{(k+1)} - \bar{v}^{(k)}\|^2 + 2\langle t_{k+1}^2(v^{(k+1)} - \bar{v}^{(k)}), e^{(k)} + t_{k+1}\bar{v}^{(k)} \rangle \\
 & = \|t_{k+1}(v^{(k+1)} - \bar{v}^{(k)})\|^2 + 2\langle t_{k+1}^2(v^{(k+1)} - \bar{v}^{(k)}), e^{(k)} + t_{k+1}\bar{v}^{(k)} \rangle \\
 & = \|t_{k+1}v^{(k+1)} - t_{k+1}\bar{v}^{(k)} + e^{(k)} + t_{k+1}\bar{v}^{(k)}\|^2 - \|e^{(k)} - t_{k+1}\bar{v}^{(k)}\|^2 \\
 & = \|t_{k+1}v^{(k+1)} + e^{(k)}\|^2 - \|e^{(k)} - t_{k+1}\bar{v}^{(k)}\|^2 \\
 [1] \implies & = \|t_{k+1}v^{(k+1)} + e^{(k)}\|^2 - \|v^{(k)} + e^{(k-1)} + t_{k+1}\bar{v}^{(k)}\|^2 \\
 & = \|t_{k+1}v^{(k+1)} + e^{(k)}\|^2 - \|e^{(k-1)} + (t_{k+1}\theta_k + 1)v^{(k)}\|^2, \tag{**}
 \end{aligned}$$

To match, we need:

$$\begin{cases} t_{k+1}^2 - t_{k+1} = t_k^2, \\ t_k = t_{k+1}\theta_k + 1. \end{cases} \tag{***}$$

## Nesterov Momentum Sequences

The Nesterov momentum sequence solves  $(\star \star \star)$ ! This is the sequence:

$$t_k = \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$
$$\theta_k = \frac{t_k - 1}{t_{k+1}}, \quad (\star \star \star)$$

1. It has the property that  $t_k \geq (k + 1)/2$ .
2. I continued from here to prove the  $\mathcal{O}(1/k^2)$  convergence rate of FISTA. Read my report on that one.

## The Lasso Problem

Lasso minimizes the 2-norm objective with one norm penalty.

$$\min_x \left\{ \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 \right\}$$

And the prox for  $\|\cdot\|_1$  is given by:

$$(\text{prox}_{\lambda\|\cdot\|_1, t}(x))_i = \text{sign}(x_i) \max(|x_i| - t\lambda, 0),$$

For our experiment:

1.  $A = \text{diag}(\text{linspace}(0, 2, 128))$ .
2. Vector  $b$  is the diagonal of  $A$  and every odd index is changed into  $\epsilon \sim N(0, 10^{-3})$ .



# Results

The plot of  $\Delta_k$ :

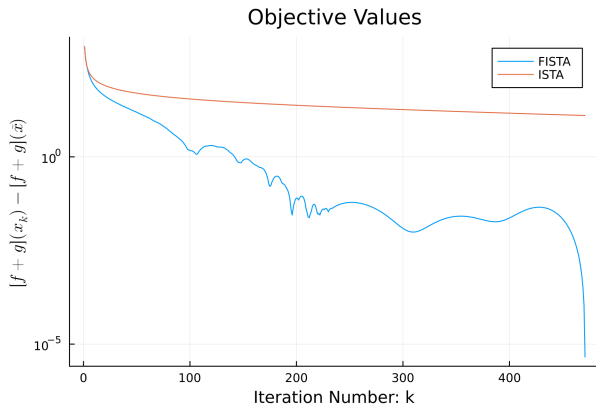
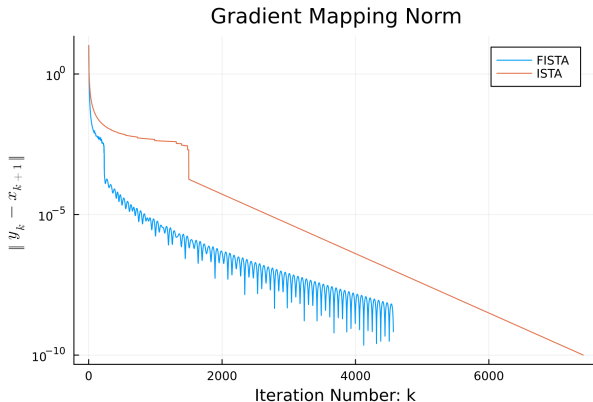


Figure: The left is the objective value of the function during all iterations.

# Results

The plot of  $\|y^{(k)} - x^{(k+1)}\|_\infty$ :



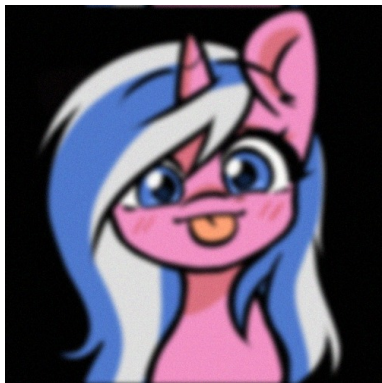
# Experiment Setup

Given an image that is convoluted by a Gaussian kernel with some gaussian noise, we want to recover the image, given the parameters for convolutions.

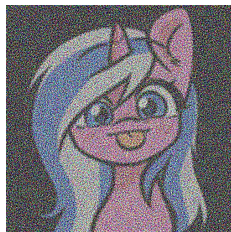
1. Gaussian blur with a discrete 15 by 15 kernel is a linear transform represented by a sparse matrix  $A$  in the computer.
2. When an image is 500 by 500 with 3 color channels,  $A$  is  $750000 \times 750000$ .
3. Let the noise be on all normalized colors values with  $N(0, 10^{-2})$
4. We let  $\lambda = \alpha \times (3 \times 500^2)^{-1}$ .
5. Implemented in Julia, and the code is too long to be shown here.

# The Blurred Image

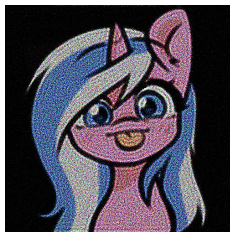
We consider blurring the image of a pink unicorn that I own.



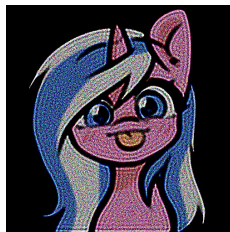
**Figure:** The image is blurred by the Gaussian Blurred matrix  $A$  with a tiny amount of noise on the level of  $2 \times 10^{-2}$  that is barely observable. Zoom in to observe the tiny amount of Gaussian noise on top of the blur.



(a)



(b)



(c)

**Figure:** (a)  $\alpha = 0$ , without any one norm penalty, is not robust to the additional noise. (b)  $\alpha = 0.01$ , there is a tiny amount of  $\lambda$ . (c)  $\alpha = 0.1$ , it is more penalty compared to (a).

## The Paper's Contribution

Beck's contribution involves proving that the Nesterov accelerations work for the proximal gradient method. Beck popularized the use of momentum in the broader context to improve the convergence of algorithms such as the proximal gradient method.

## My Contribution

I wrote the same proof in a different way for Proximal gradient in both the accelerated case and the unaccelerated case. I avoided using the assumption of Nesterov Momentum term and instead used a template method and the idea of form match to derive the sequence in the middle of the convergence proof.



Amir Beck and Marc Teboulle.

A fast iterative shrinkage-thresholding algorithm for linear inverse problems, 2009.