Proximal Gradient: Convergence, Implementations and Applications

November 12, 2022

Abstract

We review the proximal gradient and accelerated proximal gradient algorithm convergence rate under different assumptions. We then demonstrate with numerical experiments involving Lasso and deblurring images.

1 Introduction

We are concerned with the problem type:

$$\min_x g(x) + h(x), \tag{1.0.1}$$

where the objectives can be splitted into the sum of two functions. Algorithms are developed for solving optimization problems of this format. We will list some of the algorithms with their convergence rate under different assumptions.

1. The projected subgradient algorithm solves $h = \delta_Q$ where Q is a closed convex set and g is closed and convex with $Q \subseteq \mathrm{ri} \circ \mathrm{dom}(g)$. The algorithm generates a sequence of $x^{(k)}$ where the average weighted by step size of all these solutions has a convergence rate of $\mathcal{O}(1/\sqrt{k})$.

2. The proximal Graidient algorithm are used for strongly smooth function g and a convex, closed and proper function h that has an easy to compute proximal oracle. Under convexity assumption for h, the optimal and the minimizer exists and the convergence rate is $\mathcal{O}(1/k)$.

3. The Accelerated Proximal Algorithm, which is a modified version of the proximal gradient that uses Nesterov Momentum and converges with $\mathcal{O}(1/k^2)$ with an additional convexity assumption on g. The convergence can be even faster when more additional assumptions are added to g.

We introduce the setup of the algorithms, and provides proofs for some of the convergence results of the Proximal Gradient algorithms. Possible variants with step sizes will be discussed. We also implemented some example of the algorihm in Julia.

For this report, most of the content can be found in Amir's Beck and textbook [1], and the proof for the convergence of Accelerated Proximal Gradient method closely follows the original paper for FISTA[2] by Amir and Marc Teboulle. Other additional and specific materials references will be in the discussion.

(**EDIT HERE**) Section .. of the paper introduces...

2 Preliminaries

The proximal operator is a crucial component for the algorithm and its non-expansive property is relevant to the convergence of Proximal Gradient under the non-convex case. We won't go into detail for the non-convex case. Under the assumption of convexity for f, the property of the strongly smooth function is more relevant.

2.1 The Proximal Operator

Definition 1 (Proximal Operator and Moreau Envelope)**.** A Moreau Envelope $\text{env}_{\alpha,f}(x)$, $\text{prox}_{\alpha,f}$ the proximal operator are defined for some function f:

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

The proximal operator is a singleton when the function f is convex, proper and closed due to the strong convexity of $f(y) + 1/(2\alpha)\|y-x\|^2$. Observe that $\text{env}_{\alpha,f}(x) = (f \square \frac{1}{2\alpha}\|\cdot\|^2)(x)$, hence the infimal convolution gives us the interpretation that the epigraphs of the envelope is adding between the epigraph of these 2 functions. This conceptualization will help with the intuitive understanding of many proximal algorithm. In addition please observe the following identities:

$$\text{prox}_{f/\alpha,1} = \text{prox}_{f,\alpha}$$
$$\alpha^{-1}\text{env}_{\alpha f,1}(x) = \text{env}_{f,\alpha}(x).$$

Proposition 2.1 (Proximal Operator is a Nonexpansive Mapping)**.** Let $f : \mathbb{E} \mapsto \bar{\mathbb{R}}$ be a closed, convex proper function. then $\text{prox}_f(x)$, with $\alpha = 1$ is a singleton for every point $x \in \mathbb{E}$. Moreoever, for any points $x, y \in \mathbb{E}$ the estimate holds:

$$\|\text{prox}_f(x) - \text{prox}_f(y)\|_\star^2 \leq \langle \text{prox}_f(x) - \text{prox}_f(y), x - y \rangle.$$

Using the identity that $\text{prox}_{f/\alpha} = \text{prox}_{f,\alpha}$, the proximal gradient is nonexpansive for all value of α.

Lemma 2.1.1 (The Alternatie Form of Proximal Operator)**.** When the function f is convex closed and proper, the $\text{prox}_{\alpha,f}$ can be viewed the following operator $(I + \alpha\partial f)^{-1}$, which is also, a single valued operator that sometimes has a nice closed form solution to it.

Proof.

$$
\begin{aligned}
0 &\in \partial \left[f(y) + \frac{1}{2\alpha}\|y - x\|^2 \middle| y \right] (y^+) \\
0 &\in \partial f(y^+) + \frac{1}{\alpha}(y^+ - x) \\
\frac{x}{\alpha} &\in (\partial f + \alpha^{-1} I)(y^+) \\
x &\in (\alpha \partial f + I)(y^+) \\
y &\in (\alpha \partial f + I)^{-1}(x).
\end{aligned}
$$

□

2.2 The Strong Smoothness

Definition 2 (Strong Smoothness)**.** A function g is called smooth with a constant α then it satisfies:

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq \frac{\alpha}{2}\|x - y\|^2 \quad \forall x, y \in \mathbb{E}. \tag{2.2.1}$$

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

3 Proximal Gradient and Forward Backward Envelope

We introduce the algorithm through the forward backward envelope, which helps with the intuitive undestanding with this algorithm. We then state some of the important properties. The name forward backward envelope is credit to numerical method that simulates gradient dynamical system that is the summation of a stiff and nonstiff dynamics by using forward Euler on the nonstiff part, and backward Euler on the stiff part. We won't go into detail regarding this specific interpretation of the proximal gradient method.

Assumption 1. 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 $\mathrm{ri} \circ \mathrm{dom}(f) \cap \mathrm{ri} \circ \mathrm{dom}(g) \neq \emptyset$.

3.1 Poximal Gradient Minimizes the Forward Backward Envelope

First, we follow the intuitive idea of constructing a upper bounding function $m_x(y)$, a surrogate function if one prefers for $g + h$ with $\beta \geq L_g$:

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

this function $m_x(y|\beta)$ is a strongly convex function and it's equal to $g + h$ at x, and larger than it on every other points. The *envelope function*, defined as $m^+(y|\beta) := \min_y\{m_x(y|\beta)\}$

minimizes the upper bounding function, and the function m^+ is lower than $g+h$ on all points and its minimizer takes the following form:

$$\arg\min_y\{m_x(y)\} = \arg\min_y \left\{ g(x) + \nabla g(x)^T(y-x) + \frac{\beta}{2}\|y-x\|^2 + h(y) \right\}.$$

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

Proof. We consider the fact that, to minimize the envelope, zero is in the subgradient of the upper bounding function $m_x(y|\beta)$.

$$\begin{aligned} \mathbf{0} &\in \nabla g(x) + \beta(y-x) + \partial h(y) \\ \nabla g(x) + \beta x &\in \beta y + \partial h(y) \\ -\beta^{-1}\nabla g(x) + x &\in y + \beta^{-1}\partial h(y) \\ -\beta^{-1}\nabla g(x) + x &\in [I + \beta^{-1}\partial h](y) \\ \implies [I + \beta^{-1}\partial h]^{-1}(-\beta^{-1}\nabla g(x) + x) &\ni y, \end{aligned}$$

using lemma 2.1.1, the RHS is the operator $\text{prox}_{h,\beta^{-1}}(x + \beta^{-1}\nabla g(x))$. □

Remark 3.1.1. The minimizer: $\text{prox}_{h,\beta^{-1}}(x + \beta^{-1}\nabla g(x))$ is I call the proximal step. It will make the envelope $m_x(y|\beta)$ strictly lower than $f(x)$, and it this is not true, then x is a minimizer of x. This will be made clear next.

3.2 Fixed Point of the Prox Step

Denote the prox step $\mathcal{P}^{g,h}_{\beta^{-1}}(x) = \text{prox}_{h,\beta^{-1}}(x - \beta^{-1}\nabla g(x))$, in most context without ambiguity it will be simply denoted as $\mathcal{P}x$. The fixed point of $\mathcal{P}$ is a point x such that $x = \mathcal{P}x$. If this is true, then x is the minimizer of f, we denoted as $\bar{x}$. To see how this is true consider any x^+ such that $x^+ = \mathcal{P}x$, using subgradient of the envelope:

$$\begin{aligned} \mathbf{0} &\in \nabla g(x) + \beta(x^+ - x) + \partial h(x^+) \\ \beta(x - x^+) &\in \partial h(x^+) + \nabla g(x^+) \\ x = x^+ \implies \mathbf{0} &\in \partial h(x^+) + \nabla g(x^+), \end{aligned}$$

and therefore, if x^+ is fixed point of $\mathcal{P}$ then it is one of the local minimizers of the function f. Conversely, if x^+ is not a fixed point of x, then it has to make the objective value of the upper bounding function $m_x(y|\beta)$ decreases because it's a strongly convex function. However, this doesn't necessarily mean that the prox step can decrease the value of the function f. We explain more about this in the next subsection.

Remark 3.2.1. The operator $\beta(x - \mathcal{P}x)$ is called the gradient mapping in Amiar's Book [1], and it has many more important properties that are useful for the convergence proof of proximal gradient method under many different context. Please observe that, if the function $h \equiv 0$, the gradient mapping is simply the gradient of the function g. We won't go into the details here unfortunately cause that is outside of the scope.

3.3 Step-Sizes that Ensures Monotone Descent Property

With assumption 1, only a specific size of step-size can guarantee a decrease in the function value for the minimizers the minimizes the envelope, which we will call it the proximal step, give by $\text{prox}_{h,\beta^{-1}}(x - \beta^{-1}\nabla g(x))$.

Theorem 2 (Stepsize that Ensures Monotone Decrease)**.** The step size L^{-1} of the proximal gradient that guarantee a decrease in the objective value has to satisfies: $L \geq L_g$, where L_g is the lipschitz constant for the gradient of the function g.

Proof. Consider the fact that $m_x(\mathcal{P}x|L_f) \leq f(x)$ which gives:

$$\begin{aligned} m_x(\mathcal{P}x|L_f) &\leq f(x) \\ \implies m_x(\mathcal{P}x|L) &\leq f(x) \\ \implies h(\mathcal{P}x) + \langle \nabla g(x), \mathcal{P}x - x\rangle + \frac{L}{2}\|\mathcal{P}x - x\|^2 &\leq h(x) \\ h(\mathcal{P}x) - h(x) + \langle \nabla g(x) - \mathcal{P}x - x\rangle &\leq \frac{-L}{2}\|\mathcal{P}x - x\|^2, \qquad (\Delta) \end{aligned}$$

next we also consider the strong smooth property of g to obtain:

$$\begin{aligned} g(\mathcal{P}x) - g(x) - \langle \nabla g(x), \mathcal{P}x - x\rangle &\leq \frac{L_g}{2}\|\mathcal{P}x - x\|^2 \qquad (\nabla) \\ \implies h(\mathcal{P}x) + g(\mathcal{P}x) - g(x) - h(x) &\leq \left(\frac{L_g}{2} - \frac{L}{2}\right)\|\mathcal{P}x - x\|^2 \qquad (**) \\ f(\mathcal{P}x) - f(x) &\leq \left(\frac{L_g}{2} - \frac{L}{2}\right)\|\mathcal{P}x - x\|^2, \end{aligned}$$

where (**) is $(\nabla) + (\Delta)$. Observe that on the last line, if $L_g \leq L$, then the objective decrease is asserted. Additionally, using theorem 1, we have L^{-1} being the step sizes inside of the proximal gradient operator. □

Remark 3.3.1. The monotone decrease property of a step size is useful for engineering the back tracking routine for the proximal gradient method. More specifically, as long as the step size L^{-1} satisfies $m_x(\mathcal{P}x|L) \leq f(x)$, then it's an acceptabled step size.

3.4 Proximal Gradient Algorithm

Algorithm 1 Proximal Gradient With Fixed Step-sizes

Input: g, h, smooth and nonsmooth, L stepsize, $x^{(0)}$ an initial guess of solution.
for $k = 1, 2, \cdots, N$ **do**
  $x^{(k+1)} = \mathcal{P}_L^{g,h} x^{(k)}$
  if $x^{(k+1)}, x^{(k)}$ close enough **then**
    Break
  end if
end for

Remark 3.4.1. The Proximal Gradient With Fixed Step Size algorithm terminates either the iteration limit N is reached, or the fixed point iterations on the operator $\mathcal{P}$ has converged. Under some cases, the Lipschiz constant for g can be obtained, under some other cases it's not easy to obtain.

4 Convergence of Proximal Gradient

Here, we give analysis for the convergence behaviors of the algorithm in 1 with fixed stepsizes and assumption 1 is true.

4.1 Convergence Under the Convex Case

Before the proof, we state some of the quantities that are involved in the proof.

1. Recall from section 3.2 where $G_\beta(x) - \nabla g(x) \in \partial h(x^+)$ with $x^+ \in \mathcal{P}^{g,h}_{\beta^{-1}}(x)$, and this general condition is true for all values of x. We refers $G_\beta(x)$ as the residual of the proximal gradient algorithm. Finally, $G_\beta(x) = \beta(x - x^+)$
2. By choosing the stepsize $\beta^{-1} \leq L^{-1}$, we assert strict decrease of the value of the objective function, $f(x^+) \leq f(x)$.
3. We denote $\bar{f}$ to be $f(\bar{x})$ where $\bar{x}$ is one of the minimizer of f.

Theorem 3 (Convergence Under Convexity)**.** With assumption 1, execute the algorithm for N steps, we have:

$$f(x^{(N+1)}) - \bar{f} \leq \frac{\beta(\|x^{(0)} - \bar{x}\|^2 - \|x^{(N+1)} - \bar{x}\|^2)}{2(N+1)}.$$

Proof. This proof is standard and doesn't completely resemble the proof showed in [2, Aimir, Teboulle], nonetheless we will extract a lemma out of this proof and use that as the foundation for the proof in the Nesterov Accelerated case of the proximal gradient algorithm.

Firstly by the choice of step-size and the strong smoothness of the function g, we have the inequality:

$$g(x^+) \leq g(x) - \beta^{-1}\langle \nabla g(x), G_\beta(x)\rangle + \underbrace{\frac{L}{2\beta^2}\|G_\beta(x)\|^2}_{\leq \frac{1}{2\beta}\|G_\beta(x)\|^2}, \tag{*}$$

next, by the convexity of f, g we have inequalities:

$$\begin{aligned} g(x) &\leq g(z) - \langle \nabla g(x), x - z\rangle \\ h(x^+) &\leq h(z) + \langle \partial h(x^+), x^+ - z\rangle, \end{aligned}$$

where we abuse the notation $\partial h(x^+)$ to denote some vector in the subgradient of h at point x^+. Next we substitue the above results to into (*):

$$\begin{aligned} g(x^+) + h(x^+) &\leq g(x) + \beta^{-1}\langle \nabla g(x), G_\beta(x)\rangle + \frac{1}{2\beta}\|G_\beta(x)\|^2 + h(x^+) \\ &\leq g(z) + \underbrace{\langle \nabla g(x), x - z\rangle}_{[1]} - \underbrace{\beta^{-1}\langle \nabla g(x), G_\beta(x)\rangle}_{[2]} \\ &\quad + \frac{1}{2\beta}\|G_\beta(x)\|^2 + h(z) + \underbrace{\langle \partial h(x^+), x^+ - z\rangle}_{[4]}, \end{aligned} \tag{∇}$$

and we consider the summation for each of these numerically labeled term to obtain:

$$\begin{aligned} [3] &:= [1] + [2] \\ [3] &= \langle \nabla g(x), x - z - x + x^+\rangle = \langle \nabla g(x), x^+ - z\rangle \\ [3] + [4] &= \langle \nabla g(x), x^+ - z\rangle + \langle G_\beta(x) - \nabla g(x), x^+ - z\rangle \\ &= \langle G_\beta(x), x^+ - z\rangle \\ &= \langle G_\beta(x), x - z - (x - x^+)\rangle \\ &= \langle G_\beta, x - z\rangle - \langle G_\beta, \underbrace{x - x^+}_{=\beta^{-1}G_\beta(x)}\rangle \\ &= \langle G_\beta(x), x - z\rangle - \beta^{-1}\|G_\beta(x)\|^2, \end{aligned} \tag{**}$$

where at (*) we applied the substitution $G_\beta(x) - \nabla f(x) \in \partial h(x^+)$. Comtinued from ($\nabla$) we obtain

$$\begin{aligned} \underbrace{g(x^+) + h(x^+)}_{f(x^+)} &\leq \underbrace{g(z) + h(z)}_{f(x)} - \frac{1}{2\beta}\|G_\beta(x)\|^2 + \langle G_\beta, x - z\rangle \\ f(x^+) - f(z) &\leq \langle G_\beta(x), x - z\rangle - \frac{1}{2\beta}\|G_\beta(x)\|^2. \end{aligned} \tag{$\star$}$$

Next, we make the simplifications using algebra and get:

$$\begin{aligned} f(x^+) - f(\bar{x}) &\leq \frac{-1}{2\beta}\|G_\beta(x)\|^2 + \langle G_\beta, x - \bar{x}\rangle \\ &= -\frac{\beta}{2}(\|x - x^+\|^2 - 2\langle x - x^+, x - \bar{x}\rangle) \\ [5] \implies &= \frac{-\beta}{2}(\|x^+ - \bar{x}\|^2 - \|x - \bar{x}\|^2) \\ &= \frac{\beta}{2}(\|x - \bar{x}\|^2 - \|x^+ - \bar{x}\|^2), \end{aligned}$$

and since the step-size assert an non-decreasing sequence of number, we perform the tele-

scoping sum on one side and get:

$$
\begin{aligned}
f(x^{(k+1)}) - \bar{f} &\leq \frac{\beta}{2}(\|x^{(k)} - \bar{x}\|^2 - \|x^{(k+1)} - \bar{x}\|^2) \\
\implies \left(\sum_{i=0}^{N} f(x^{(i+1)}) - \bar{f}\right) &\leq \frac{\beta}{2}(\|x^{(0)} - \bar{x}\|^2 - \|x^{(N+1)} - \bar{x}\|^2) \\
f(x^{(N+1)}) - \bar{f} = \min_{i=0,\cdots,N} \left\{f(x^{(i+1)}) - \bar{f}\right\} &\leq \left(\frac{1}{N+1}\sum_{i=0}^{N} f(x^{(i+1)})\right) - \bar{f} \\
\implies f(x^{(N+1)}) - \bar{f} &\leq \frac{\beta(\|x^{(0)} - \bar{x}\|^2 - \|x^{(N+1)} - \bar{x}\|^2)}{2(N+1)} \\
&\leq \frac{\beta\|x^{(0)} - \bar{x}\|^2}{2(N+1)}.
\end{aligned}
$$

□

Remark 4.1.1. One important lemma that we can extract from this proof which will later be important for the proof for the accelerated case is the tagged expression $(\star)$ in the above derivation. We will refers this as the "Prox Step 2 Points" lemma. Expression $(\star)$ is equivalent to the lemma 2.3 in the FISTA paper[2].

Lemma 4.1.1 (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 $y^+ = \mathcal{P}^{g,h}_{\beta^{-1}}(y)$ we have for any $x \in \mathbb{E}$:

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

Proof. The proof is continuted from expression $(\star)$:

$$
\begin{aligned}
f(x^+) - f(z) &\leq \langle G_\beta(x), x - z\rangle - \frac{1}{2\beta}\|G_\beta(x)\|^2. \\
f(x^+) - f(z) &\leq \beta\langle x - x^+, x - z\rangle - \frac{1}{2\beta}\|\beta(x - x^+)\|^2 \\
f(z) - f(x^+) &\geq \frac{\beta}{2}\|x - x^+\|^2 + \beta\langle x^+ - x, x - z\rangle,
\end{aligned}
$$

and by substituting $x := y$ and $z := x$ in the last line, we completed the proof of the lemma. □

5 Accelerated Proximal Gradient

Here we state the FISTA algorithm in paper[2]. The convergence rate will also be stated but the proof with extra details that closedly follows what is in the paper will be put into the appendix.

5.1 Accelerated Proximal Gradient Algorithm

Algorithm 2 FISTA With Constant Step Size

Input: the step size β^{-1}, and $x^{(0)}$ the initial guess.
$y^{(1)} = x^{(0)}$
for $k = 1, \cdots, N$ **do**
  $x^{(k)} := \mathcal{P} y^{(k)}$
  if $y^{(k)} - x^{(k)}$ small enough **then**
    Break
  end if
  $t_{k+1} := \frac{1+\sqrt{1+4t_k^2}}{2}$
  $y^{(k+1)} := x^{(k)} + \frac{t_k - 1}{t_{k+1}}(x^{(k)} - x^{(k-1)})$
end for

5.2 Convergence Under the Convex Case

Theorem 4 (FISTA Convergence under Convexity)**.** If assumption 1 is satisfied, then the FISTA algorithm has convergence result of:

$$f(x^{(k)}) - f(\bar{x}) \leq \frac{2\beta^{-1}\|x^{(0)} - \bar{x}\|^2}{(k+1)^2},$$

where $\bar{x}$ is one of the optimizers and hence the rate of convergence is $\mathcal{O}(1/k^2)$.

Proof. For a proof see appendix A □

6 Numerical Experiments

Simple LASSO

As the name suggested, we consider the overuse example problem of:

$$\min_x \|Ax - b\|_2^2 + \lambda\|x\|_1$$

For a brief background, This optimization problem commonly appears in the contex of regression for generalized linear model where selection for sparse cofficients on the regression parameters is desired. Theoretically it corresponds to having a prior Laplace distribution for the regression parameters. The implementation is simple and the proximal gradient oracle for $\|\|_1$ is given as:

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

And one can interpret the sign function as a projection onto the interval $[-1, 1]$, and the $\max(|x| - t\lambda, 0)$ function as the distance of x to the set $[-t\lambda, t\lambda]$.

For this simple problem, we consider A to be 128 by 128, that is a diagonal matrix whose diagnals are equally space points from the interval $[0, 1]$. The right handside vector b is

the same as the diagonal of matrix A, but every odd index is replaced with a guassian random noise on the level of $1e-4$. The experiment is performed with using both ISTA and FISTA with $\lambda = 1e-1$ (This is used to prevent triggering the line search routine in the implementation), both uses a step size of 0.1 and an initial guess of a vector of all ones.

For the experiment we record and present the objective values f for each of the iterations and the norm of the proximal mapping $\|x^{(k+1)} - x^{(k)}\|_2$ for each iteration. See figure 1 for an illustration. Observe that the type of convergence for the proximal gradient operator in the FISTA case is fundamentally different compare to the ISTA case. In the case of ISTA, the norm of the proximal mapping on the log plot is close to a straight line, indicating a first order convergence of this quantity. In the case of FISTA, the rate is obviously slower from the plot, and there are non trivial amount of oscilations which is unlikely to be the results of numerical issues. In fact the convergence rate in general is linear when the smooth function

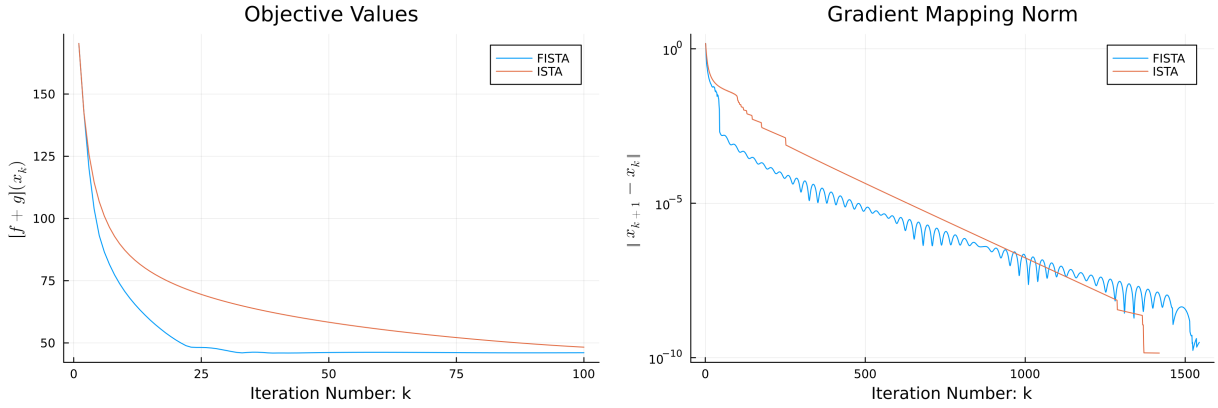

Figure 1: The left is the objective value of the function during all iterations and the right side is the norm of the gradient mapping for all the iteraitons.

g is strongly convex (**REFERENCES NEEDED**).

Image Deblurring

Here we reproduce the some of the experiment conducted in

My Ideas

A Proof for Convergence of FISTA

Here we prove theorem 4. The proof closely follows Beck and Teboulle's work [2], and here we prove everything with extra details.

A.1 The First Lemma

Recall from lemma 4.1.1 and assumption 1, instead of using β^{-1} as our step size we use L. Here we introduce the notation $\delta_k = f(x^{(k)}) - f(\bar{x})$ to denote the optimality gap at the k iteration of the algorithm where $\bar{x}$ denotes one of the minimizer of the function f.

Lemma A.1.1 (FISTA First Lemma)**.** The optimality gap generated via FISTA statisfies:

$$\begin{aligned}
&\frac{2}{L}t_k^2\Delta_k - \frac{2}{L}t_{k+1}^2\Delta_{k+1} \geq \|u^{(k+1)}\|^2 - \|u^{(k)}\|^2 \\
&\Delta_k := f(x^{(k)}) - f(\bar{x}) \\
&u^{(k)} := t_k x^{(k)} - (t_k - 1)x^{(k-1)} - \bar{x}.
\end{aligned}$$

Proof. We invoke the lemma 4.1.1 with $x = x^{(k)}, y = y^{(k+1)}$ to give:

$$\begin{aligned}
f(x^{(k)}) - f \circ \mathcal{P}y^{(k+1)} &\geq \frac{L}{2}\|\mathcal{P}y^{(k+1)} - y^{(k+1)}\|^2 + L\langle y^{(k+1)} - x^{(k)}, \mathcal{P}y^{(k+1)} - y^{(k+1)}\rangle \\
f(x^{(k)}) - f(x^{(k+1)}) &\geq \frac{L}{2}\|x^{(k+1)} - y^{(k+1)}\|^2 + L\langle y^{(k+1)} - x^{(k)}, x^{(k+1)} - y^{(k+1)}\rangle, \\
\implies 2L^{-1}(\Delta_k - \Delta_{k+1}) &\geq \|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle x^{(k+1)} - y^{(k+1)}, y^{(k+1)} - \bar{x}\rangle. \qquad (*)
\end{aligned}$$

Observe that from the first line to the second line we invoke the definition for the updates for $x^{(k)}$ in the FISTA algorithm (**HYPERLINK NEEDED**). We then invoke the lemma again with $x := \bar{x}, y = y^{(k+1)}$, and it gives us:

$$\begin{aligned}
f(\bar{x}) - f \circ \mathcal{P}y^{(k+1)} &\geq \frac{L}{2}\|\mathcal{P}y^{(k+1)} - y^{(k+1)}\|^2 + L\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)}\rangle \\
f(\bar{x}) - f(x^{(k+1)}) &\geq \frac{L}{2}\|x^{(k+1)} - y^{(k+1)}\|^2 + L\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)}\rangle \\
-2L^{-1}\Delta_{k+1} &\geq \|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)}\rangle. \qquad (\star)
\end{aligned}$$

Consider the expression $(t_k - 1)(*) + (\star)$, we obtain the LHS of that expression:

$$\begin{aligned}
&(t_{k+1} - 1)L^{-1}(\Delta_k - \Delta_{k+1}) - 2L^{-1}\Delta_{k+1} \\
&= 2L^{-1}((t_{k+1} + 1)\Delta_k - (t_{k+1} - 1)\Delta_{k+1} - \Delta_{k+1}) \\
&= 2L^{-1}((t_{k+1} - 1)\Delta_k - t_{k+1}\Delta_{k+1}), \qquad (\star*\text{LHS})
\end{aligned}$$

and then the RHS is:

$$\begin{aligned}
&(t_{k+1} - 1)\|x^{(k+1)} - y^{(k+1)}\|^2 + 2(t_{k+1} - 1)\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)}\rangle \\
&\quad + \|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle y^{(k+1)} - \bar{x}, x^{(k+1)} - y^{(k+1)}\rangle \\
&= t_{k+1}\|x^{(k+1)} - y^{(k+1)}\|^2 + \langle x^{(k+1)} - y^{(k+1)}, \underbrace{2(t_{k+1} - 1)(y^{(k+1)} - x^{(k)}) + 2(y^{(k+1)} - \bar{x})}_{=2(t_{k+1}y^{(k+1)} + (1-t_{k+1})x^{(k)} - \bar{x})}\rangle \\
&= t_{k+1}\|x^{(k+1)} - y^{(k+1)}\|^2 + 2\langle x^{(k+1)} - y^{(k+1)}, t_{k+1}y^{(k+1)} + (1 - t_{k+1})x^{(k)} - \bar{x}\rangle. \qquad (\star*\text{RHS})
\end{aligned}$$

The entirety of expression (3) is given by:

$$\begin{aligned}
2L^{-1}((t_{k+1} - 1)\Delta_k - t_{k+1}\Delta_{k+1}) &\geq t_{k+1}\|x^{(k+1)} - y^{(k+1)}\|^2 \\
&\quad + 2\langle x^{(k+1)} - y^{(k+1)}, t_{k+1}y^{(k+1)} + (1 - t_{k+1})x^{(k)} - \bar{x}\rangle. \qquad (\star*)
\end{aligned}$$

The next part of the proof shows some of the magics involves in changing the LHS,RHS of $(\star*)$ to be the similar to what is in the theorem statement. It's accomplished by the relations

for the sequence t_k, more specifically it's hinged on the relations $t_k^2 = t_{k+1}^2 - t_{k+1}$ which is asserted by the FISTA algorithm. Using this fact we proceed by multiplying t_{k+1} on both sides of $(\star\star)$ and obtain:

$$
\begin{aligned}
&2L^{-1}(\Delta_k t_k^2 - \Delta_{k+1} t_{k+t}^2) \\
&\quad \geq \|t_{k+1}(x^{(k+1)} - y^{(k+1)})\|^2 - 2\langle t_{k+1}(x^{(k+1)} - y^{(k+1)}), t_{k+1}y^{(k+1)} - (t_{k+1}-1)x^{(k)} - \bar{x}\rangle \\
&2L^{-1}(\Delta_k t_k^2 - \Delta_{k+1} t_{k+t}^2) \\
&\quad \geq \| \underbrace{t_{k+1}x^{(k+1)}}_{=:a} - \underbrace{t_{k+1}y^{(k+1)}}_{=:b} \|^2 - 2\langle \underbrace{t_{k+1}x^{(k+1)}}_{=:a} - t_{k+1}y^{(k+1)}, \underbrace{t_{k+1}y^{(k+1)}}_{=:b} - \underbrace{((t_{k+1}-1)x^{(k)} + \bar{x})}_{=:c} \rangle \\
&\quad \geq \|a-b\|^2 + 2\langle a-b, b-c\rangle \\
&\quad = \|a-b\|^2 + \|b-c\|^2 + 2\langle a-b, b-c\rangle - \|b-c\|^2 \\
&\quad = \|a-c\|^2 - \|b-c\|^2 \\
&\quad \geq \|t_{k+1}x^{(k+1)} - (t_{k+1}-1)x_k - \bar{x}\|^2 - \|(t_{k+1}-1)x^{(k)} - t_{k+1}y^{(k+1)} - \bar{x}\|^2,
\end{aligned}
$$

to prove the lemma, we need to match the form in the above 2 norm of the vector, we accomplish this by considering FISTA algorithm:

$$
\begin{aligned}
t_{k+1}y^{(k+1)} &= t_{k+1}x^{(k)} + (t_k - 1)(x^{(k)} - x^{(k-1)}) \\
t_{k+1}y^{(k+1)} - (t_{k+1}-1)x^{(k)} &= t_{k+1}x^{(k)} - (t_{k+1}-1)x^{(k)} + (t_k-1)(x^{(k)} - x^{(k-1)}) \\
&= x^{(k)} + (t_k-1)x^{(k)} - (t_k-1)x^{(k-1)} \\
&= t_k x^{(k)} - (t_k - 1)x^{(k-1)},
\end{aligned}
$$

cf from previously we have:

$$
\begin{aligned}
&2L^{-1}(\Delta_k t_k^2 - \Delta_{k+1} t_{k+t}^2) \\
&\geq \|t_{k+1}x^{(k+1)} + (1-t_{k+1})x^{(k)} - \bar{x}\|^2 - \|t_k x^{(k)} + (1-t_k)x^{(k-1)} - \bar{x}\|^2 \\
&\geq \|u^{(k+1)}\|^2 - \|u^{(k)}\|^2.
\end{aligned}
$$

And that completes the proof of the Second Lemma. □

Remark A.1.1. There should be some point, where we can infer the properties of the sequence t_k instead of taking the sequence from FISTA algorithm for granted, there should also be a way to make a different decision during the proof so that this becomes the proof for the algorithm without the accelerations technique.

A.2 The Second Lemma

Lemma A.2.1 (FISTA Second Lemma)**.** Let $\{a, b\}$ be positive real numbers sequence satisfying: $a_k - a_{k-1} \geq b_{k+1} - b_k, \forall k \geq 1$, with $a_1 + b_1 \leq c, c > 0$, and then it would mean that $a_{k+1} \leq c$.

Proof. The base case of the proof is obvious by the fact that b_1 is positive, hence $a_1 \leq c$ is true. the relation automatically holds true for all $k \geq 1$ because $a_k - a_{k+1} \geq b_{k+1} - b_k \implies a_k + b_k \leq c$, then $a_k - a_{k+1} \geq b_{k+1} - b_k \implies a_k + b_k \leq a_{k+1} + b_{k+1} \implies a_{k+1} + b_{k+1} \leq c \implies a_{k+1} \leq c$. □

A.3 The Third Lemma

Lemma A.3.1 (FISTA Third Lemma)**.** The FISTA asserts $t_k \geq (k+1)/2, \forall k \geq 1$.

$$
\begin{aligned}
t_k &\geq \frac{k+1}{2} \\
4t_k^2 &\geq 4\left(\frac{k+1}{2}\right)^2 \\
\implies t_k &= \frac{1}{2}\left(1+\sqrt{1+4t_k^2}\right) \\
&= \frac{1}{2} + \frac{\sqrt{1+(k+1)^2}}{2} \\
&\geq \frac{1+k}{2},
\end{aligned}
$$

A.4 Convergence Proof

Firstly we define the quantities:

1. $a_k := (2/L)t_k^2\Delta_k$.
2. $b_k := \|u^{(k)}\|^2$.
3. $c := \|x^{(0)} - \bar{x}\|^2 = \|y^{(1)} - \bar{x}\|^2$.

Recall from the first lemma, and we can represents it using the quantities listed above:

$$
\begin{aligned}
2L^{-1}(\Delta_k t_k^2 - \Delta_{t+1}t_{k+1}^2) &\geq \|u^{(k+1)}\|^2 - \|u^{(k)}\|^2 \\
a_k - a_{k+1} &\geq b_{k+1} - b_k,
\end{aligned}
$$

To demonstrate how the base case hold up for the second lamma, we consider lemma lemma 4.1.1, substituting $x^{(1)}$ for x and $\bar{x}$ for y, implicitly using the fact that $\bar{x}$ is a fixed point of $\mathcal{P}$:

$$
\begin{aligned}
\Delta_1 &\geq \frac{L}{2}\|\mathcal{P}y^{(1)} - y^{(1)}\|^2 + L\langle y^{(1)} - \bar{x}, \mathcal{P}y^{(1)} - y^{(1)}\rangle \\
&= \frac{L}{2}\|\mathcal{P}x^{(1)} - y^{(1)}\|^2 + L\langle x^{(1)} - \bar{x}, x^{(1)} - y^{(1)}\rangle \\
&= \frac{L}{2}(\|x^{(1)} - \bar{x}\|^2 - \|y^{(1)} - \bar{x}\|^2) \\
\implies 2L^{-1}\Delta_1 &\leq \underbrace{\|y^{(1)} - \bar{x}\|^2}_{=c} - \|x^{(1)} - \bar{x}\|^2 \\
\implies a_1 + b_1 &\leq c,
\end{aligned}
$$

using the fact that $t_1 = 1$ we have $u^{(1)} = x^{(1)} - \bar{x} \implies b_1 = \|x^{(1)} - \bar{x}\|^2$, please observe that the above expression simplifies to $a_1 + b_1 \leq c$. Invoking the second lemma, we have the claim that $a_{k+1} \leq c$, which is stated as:

$$\begin{aligned} 2L^{-1}t_{k+1}^2\Delta_{k+1} &\leq \|x^{(0)} - \bar{x}\|^2 \\ \implies \Delta_{k+1} &\leq \frac{L\|x^{(0)} - \bar{x}\|^2}{2t_k^2} \\ &\leq \frac{L\|x^{(0)} - \bar{x}\|^2}{2 \times 2^{-2}(k+1)^2} = \frac{2L\|x^{(0)} - \bar{x}\|^2}{(k+1)^2}, \end{aligned}$$

and the proof is now complete.

References

[1] Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems, 2009.