Proximal Gradient: Convergence, Implementations and Applications

Hongda Li

UBC Okanagan

November 23, 2022

ToC

Sum of 2 Functions

$$\min_{x} g(x) + h(x) \tag{1}$$

Through out the presentation we assume that the objective of some kind of function f can be interpreted as the sum of 2 functions. The paper we will be focusing on: FISTA (Fast Iterative-Shrinkage Algorithm) by Beck and Teboulle.

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.
3. BIG Numerical Experiments!

Stuff to Go Over

What is FISTA

Simply speaking, the FISTA algorithm is the non-smooth analogy of gradient descend 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 arised during the proof for the proximal gradient method that is later useful for the proof for the FISTA.
3. Derive the FISTA algorithm's convergence rate and construct the sequence of the Nesterov Momentum during the proof using a template algorithm.

Proximal Operator Definition

Definition

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

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

Prox is the Resolvant of Subgradient

Lemma (Resolvant 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.

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

□

An Example of Prox

Definition (Soft Thresholding)

For some $x \in \mathbb{R}$, the proximal operator of it's absolute value is given as:

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

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 $[-t\lambda, t\lambda]$.

Strong Smoothness

Definition (Strong Smoothness)

A differentiable function g is called strongly smooth with a constant α then it satisfies:

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

Remark

The absolute value sign can be removed and replaced with $0 \le$ 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 α if and only if the gradient ∇g is globally Lipschitz continuous with a parameter of α and g is closed and convex.

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

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* $ri \circ dom(g) \cap ri \circ dom(h) \neq \emptyset$. *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 an upper bounding functions at the point x evaluated at y for the function f and it's 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 $\mathit{prox}_{h,\beta^{-1}}(x + \beta^{-1}\nabla g(x))$, 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 $m_x(y|\beta)$.

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

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 simplicty 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 $\mathcal{P}^{g,h}_{\beta^{-1}}(x)$ when there is no ambiguity we simply use $\mathcal{P}x$.

The Proximal Gradient Method

Algorithm 1 Proximal Gradient With Fixed Step-sizes

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

The proximal Gradient Method

1. Converges Monotonically for stepsize $L \geq L_g$.
2. 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 satisfiying assumption 1.

References