Linear convergence of proximal gradient descent

Robin Francis
Student
ECE, IISc
robinfrancis@iisc.ac.in

Abstract

Gradient descent algorithms are not applicable for solving minimization of composite functions in which some functions are not differentiable. Proximal gradient methods provide a suitable methodology to solve the composite functions which may not be completely differentiable. When the non differentiable part is the indicator function the proximal operator reduces to the projection operator. The proximal gradient methods exhibit sublinear convergence for smooth convex functions. In this report, we investigate the convergence for smooth and strongly convex functions. We present proofs of linear convergence and experimentally verify the convergence results.

1 Introduction

Consider a constrained optimization problems of the form

$$\underset{\mathbf{x}}{\text{minimize}} \quad F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \tag{1}$$

where $f : \mathbf{dom}(f) \to \mathrm{R}$ is a L-smooth and closed proper μ-strongly convex function and $g : \mathbf{dom}(g) \to \mathrm{R}$ is a closed proper convex function that need not be differentiable. The optimal set of the problem (1) is nonempty and denoted by $\mathcal{X}^*$. If the function $F(\cdot)$ was differentiable then gradient descent (GD) algorithms would suffice. The proximal gradient methods solves the above optimization problems were the composite function need not be differentiable. Proximal algorithm is an algorithm for solving a convex optimization problem that uses proximal operators of the objective terms. We split the composite function $F(\cdot)$ into two terms, one of which is differentiable. We perform GD on the differentiable term as

$$\mathbf{y}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) \tag{2}$$

where $\mathbf{y}_{k+1}$ is the intermediate variable, $\mathbf{x}_k$ denote the iterate at kth iterate, η the learning rate and $\nabla f(\mathbf{x_k})$ the gradient evaluated at the current iterate. To obtain the next iterate we do a proximal map of the intermediate variable $\mathbf{y}_{k+1}$. The next iterate is given as

$$\mathbf{x}_{k+1} = \mathit{prox}_g(\mathbf{y}_{k+1}) = \mathit{prox}_g\left(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)\right) \tag{3}$$

The proximal mapping is an operator defined for a function $g : \mathbf{dom}(g) \to \mathrm{R}$ is given by

$$\mathit{prox}_g(\mathbf{x}) = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ g(\mathbf{u}) + \frac{1}{2} ||\mathbf{u} - \mathbf{x}||^2 \right\}. \tag{4}$$

The properties of the proximal map obtained from the proximal operator is given in the next lemma.

Lemma 1. *[Theorem 10.16 [1]] (uniqueness). Let* $g : \mathbf{dom}(g) \to \mathrm{R}$ *is a closed proper convex function. Then* $\mathit{prox}_g(\mathbf{x})$ *is a singleton for any* $x \in \mathbf{dom}(g)$.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Proof. The $\mathit{prox}_g(\cdot)$ can be written as

$$\mathit{prox}_g(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \quad \tilde{g}(\mathbf{u}, \mathbf{x})$$

where $\tilde{g}(\mathbf{u}, \mathbf{x}) = g(\mathbf{u}) + \frac{1}{2}||\mathbf{u} - \mathbf{x}||^2$. The function $\tilde{g}(\mathbf{u}, \mathbf{x})$ is a closed and strongly convex function. As the function $\tilde{g}(\mathbf{u}, \mathbf{x})$ is closed and strongly convex there exists a unique minimizer for $\mathit{prox}_g(\mathbf{x})$. □

From Lemma: 1, if the function g is proper closed and convex, then $\mathit{prox}_g(\mathbf{x})$ is always a singleton i.e., the proximal map exists and is unique.

If the function $g(\cdot)$ is the indicator function then the proximal operator is the projection operator. The indicator function is given as

$$\mathbf{I}_{\mathcal{C}} := \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{C}, \\ \infty & \text{else }, \end{cases} \tag{5}$$

The projection operator $\Pi_{\mathcal{C}}(\mathbf{x})$ is given as

$$\Pi_{\mathcal{C}}(\mathbf{x}) = \underset{\mathbf{u} \in \mathcal{C}}{\text{argmin}} \quad \frac{1}{2}||\mathbf{u} - \mathbf{x}||^2.$$

In this next section we prove the convergence of the proximal operator.

2 Convergence of proximal GD – strongly convex case

In this section, we discuss the convergence of proximal gradient descent algorithms, where the composite function $f(\cdot)$ is assumed to be strongly convex. The definitions for L smoothness and μ strong convexity is defined below.

Definition 1. *(L smoothness) The function $f : \mathcal{R}^d \to \mathcal{R}$ is L smooth if,*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||_2^2, \quad \forall \mathbf{y}, \mathbf{x} \in \mathcal{R}^d.$$

Definition 2. *(Convexity and μ strong convexity) The function $f : \mathcal{R}^d \to \mathcal{R}$ is said to be convex if,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y}, \mathbf{x} \in \mathcal{R}^d.$$

and μ strong convex if,

$$\text{(i)} \quad f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||_2^2, \quad \forall \mathbf{y}, \mathbf{x} \in \mathcal{R}^d.$$

$$\text{(ii)} \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)\mu}{2}\|x - y\|^2, \alpha \in [0, 1].$$

Now, we present the proximal gradient descent for solving the optimization problem in Equation (1). We define the

$$P(\mathbf{x}) := \text{prox}_{L^{-1}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right). \tag{6}$$

If the function $g(\cdot)$ is the indicator function then the operator $P(\mathbf{x}) := \Pi_{\mathcal{C}}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)$ is the projection operator. The pseudo code for proximal GD is given in Algorithm: 1.

Algorithm 1 Proximal gradient descent

```
Initialize $\mathbf{x}_0$ and $\eta$
for $k = 0, 1, \ldots$ do
    $\mathbf{y}_{k+1} \leftarrow \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k)$        ▷ Gradient descent update
    $\mathrm{x}_{k+1} \leftarrow P(\mathbf{x}_k)$        ▷ Proximal map
```

2.1 Analysis 1

The analysis of Proximal GD is based on the analysis from [1]. The analysis is based on an inequality as defined in the following lemma.

Lemma 2. *[Theorem 10.16] Consider any composite function $F(\cdot)$ as defined in (1). For any $\mathbf{x} \in \mathbf{dom}(F), \mathbf{y} \in (\mathbf{dom}(f))$ and $L > 0$ satisfying the smoothness definition*

$$f\left(P(\mathbf{y})\right) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), P(\mathbf{y}) - \mathbf{y} \rangle + \frac{L}{2} \left\| P(\mathbf{y}) - \mathbf{y} \right\|^2$$

it holds that

$$F(\mathbf{x}) - F\left(P(\mathbf{y})\right) \geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2 - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y})$$

where,

$$\ell_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Proof. For the analysis of the theorem consider the function defined as

$$\varphi(\mathbf{u}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{u} - \mathbf{y} \rangle + g(\mathbf{u}) + \frac{L}{2} \|\mathbf{u} - \mathbf{y}\|^2 \tag{7}$$

where, the functions f and g are defined in the formulation (1). Since the function φ is L-strongly convex function and the proximal map $P(\mathbf{y}) = \underset{\mathbf{u} \in \mathbb{E}}{\operatorname{argmin}} \varphi(\mathbf{u})$. As the function $\varphi(\cdot)$ is L strongly convex we deduce the following expression

$$\varphi(\mathbf{x}) - \varphi\left(P(\mathbf{y})\right) \geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2.$$

Now from the definition of $\varphi(\cdot)$ as in (7),

$$\begin{aligned}
\varphi\left(P(\mathbf{y})\right) &= f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), P(\mathbf{y}) - \mathbf{y} \rangle + \frac{L}{2} \left\| P(\mathbf{y}) - \mathbf{y} \right\|^2 + g\left(P(\mathbf{y})\right) \\
&\overset{(a)}{\geq} f\left(P(\mathbf{y})\right) + g\left(P(\mathbf{y})\right) \\
&\overset{(b)}{=} F\left(P(\mathbf{y})\right).
\end{aligned}$$

The expression (a), is obtained from the definition of L smoothness for the function f and (b) follows from the definition of the composite function $F(\cdot)$. Thus for any $\mathrm{x} \in \mathbf{dom}(f)$,

$$\varphi(\mathbf{x}) - F\left(P(\mathbf{y})\right) \geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2.$$

Plugging the expression for $\varphi(\mathrm{x})$ in (7), into the above inequality, we obtain the expression

$$\begin{aligned}
f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + g(\mathbf{x}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 - F\left(P(\mathbf{y})\right) &\geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2 \\
f(\mathbf{x}) - f(\mathbf{x}) + g(\mathbf{x}) + f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 - F\left(P(\mathbf{y})\right) &\geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2,
\end{aligned}$$

the above expression is obtained by introducing the term $f(\mathbf{x})$. On rearranging the expression we obtain the desired expression

$$F(\mathbf{x}) - F\left(P(\mathbf{y})\right) \geq \frac{L}{2} \left\| \mathbf{x} - P(\mathbf{y}) \right\|^2 - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 + f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

□

Lemma 3. *The function defined as*

$$\varphi(\mathbf{u}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{u} - \mathbf{y} \rangle + g(\mathbf{u}) + \frac{L}{2} \|\mathbf{u} - \mathbf{y}\|^2,$$

is L strongly convex.

Proof. Consider the term $\varphi(\alpha x + (1-\alpha)z) + \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x}-\mathbf{z}\|^2$ for some $\alpha \in [0,1]$

$$= f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \alpha\mathbf{x} + (1-\alpha)\mathbf{z} - \mathbf{y}\rangle + g(\alpha\mathbf{x} + (1-\alpha)\mathbf{z}) + \frac{L}{2}\|\alpha\mathbf{x} + (1-\alpha)\mathbf{z} - \mathbf{y}\|^2 + \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x}-\mathbf{z}\|^2$$

$$\overset{(a)}{\leq} \alpha\left(f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + g(\mathbf{x})\right) + (1-\alpha)\left(f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{z}-\mathbf{y}\rangle + g(\mathbf{z})\right)$$

$$\alpha^2\frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 + (1-\alpha)^2\frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2 + \alpha(1-\alpha)L\langle \mathbf{x}-\mathbf{y}, \mathbf{z}-\mathbf{y}\rangle + \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x}-\mathbf{z}\|^2$$

$$= \alpha\varphi(\mathbf{x}) + (1-\alpha)\varphi(\mathbf{z}) - \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x}-\mathbf{y}\|^2 - \frac{\alpha(1-\alpha)L}{2}\|\mathbf{z}-\mathbf{y}\|^2$$

$$+ \alpha(1-\alpha)L\langle \mathbf{x}-\mathbf{y}, \mathbf{z}-\mathbf{y}\rangle + \frac{\alpha(1-\alpha)L}{2}\|\mathbf{x}-\mathbf{z}\|^2$$

$$= \alpha\varphi(\mathbf{x}) + (1-\alpha)\varphi(\mathbf{z}).$$

The expression (a) is obtained from the assumption that the function $g(\cdot)$ is convex. From the definition of the strong convexity we can deduce that the function $\varphi(\cdot)$ is L strongly convex. □

Lemma 4. *Given a μ strongly convex function $f(\cdot)$ the suboptimality gap, $f(\mathbf{x}_k) - f^*$ can be bounded as*

$$f(\mathbf{x}_k) - f^* \geq \frac{\mu}{2}\left\|\mathbf{x}_k - \mathbf{x}^*\right\|^2. \tag{8}$$

Proof. The bound directly follows from the definition of strong convexity with the fact that $\nabla f(\mathbf{x}^*) = 0$. □

Corollary 1. *From Lemma: 2, the function value for the proximal GD is a monotonically decreasing sequence.*

Proof. From Lemma: 2, with $\mathbf{x} = \mathbf{y} = \mathbf{x}_k$

$$F(\mathbf{x}_k) - F\left(\mathbf{x}_{k+1}\right) \geq \frac{L}{2}\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|^2 - \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}_k\|^2 + \ell_f(\mathbf{x}_k, \mathbf{x}_k)$$

$$F(\mathbf{x}_k) - F\left(\mathbf{x}_{k+1}\right) \geq \frac{L}{2}\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|^2$$

$$F(\mathbf{x}_k) - F\left(\mathbf{x}_{k+1}\right) \geq 0.$$

□

Equipped with the inequality from Lemma: 2, we derive the linear convergence of the proximal GD for strongly convex case.

Theorem 1 (Theorem 10.29)**.** *Consider function $F(\cdot)$ as defined above and that in addition f is μ strongly convex ($\mu > 0$). Let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method. Then for any $k \geq 0$,*

$$(a)\ \left\|\mathrm{x}_k - \mathrm{x}^*\right\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \left\|\mathrm{x}^0 - \mathrm{x}^*\right\|^2, \tag{9}$$

$$(b) F\left(\mathbf{x}^{k+1}\right) - F^* \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{k+1}\left\|\mathbf{x}^0 - \mathbf{x}^*\right\|^2. \tag{10}$$

Proof. To prove the above we make use of the Lemma: 2 with $\mathbf{x} = \mathbf{x}^*$, and $\mathbf{y} = \mathbf{x}_k$, we obtain

$$F\left(\mathbf{x}^*\right) - F\left(\mathbf{x}_{k+1}\right) \geq \frac{L}{2}\left\|\mathbf{x}^* - \mathbf{x}_{k+1}\right\|^2 - \frac{L}{2}\left\|\mathbf{x}^* - \mathbf{x}_k\right\|^2 + \ell_f\left(\mathbf{x}^*, \mathbf{x}_k\right),$$

the above expression follows as $\mathbf{x}_{k+1} = P(\mathbf{x}_k)$. Since f is μ strongly convex, we bound the linear term $\ell_f\left(\mathbf{x}^*, \mathbf{x}_k\right)$ as

$$\ell_f\left(\mathbf{x}^*, \mathbf{x}_k\right) = f\left(\mathbf{x}^*\right) - f\left(\mathbf{x}_k\right) - \left\langle \nabla f\left(\mathbf{x}_k\right), \mathbf{x}^* - \mathrm{x}_k\right\rangle \geq \frac{\mu}{2}\left\|\mathbf{x}_k - \mathbf{x}^*\right\|^2$$

Using the above expression for the linear term, the above expression reduces to

$$
\begin{aligned}
F\left(\mathbf{x}^*\right)-F\left(\mathbf{x}_{k+1}\right) &\geq \frac{L}{2}\left\|\mathbf{x}^*-\mathbf{x}_{k+1}\right\|^2-\frac{L}{2}\left\|\mathbf{x}^*-\mathbf{x}_k\right\|^2+\frac{\mu}{2}\left\|\mathbf{x}^*-\mathbf{x}_k\right\|^2 \\
F\left(\mathbf{x}^*\right)-F\left(\mathbf{x}_{k+1}\right) &\geq \frac{L}{2}\left\|\mathbf{x}^*-\mathbf{x}_{k+1}\right\|^2-\frac{L-\mu}{2}\left\|\mathbf{x}^*-\mathbf{x}_k\right\|^2 \qquad (11)
\end{aligned}
$$

Since x^* is a minimizer of F, we employ the inequality $F\left(\mathrm{x}^*\right)-F\left(\mathrm{x}^{k+1}\right) \leq 0$ to obtain

$$
\left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\|^2 \leq\left(1-\frac{\mu}{L}\right)\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2 .
$$

The part (a) follows from the above equation by unrolling as

$$
\left\|\mathrm{x}_k-\mathrm{x}^*\right\|^2 \leq\left(1-\frac{\mu}{L}\right)^k\left\|\mathrm{x}^0-\mathrm{x}^*\right\|^2 .
$$

From the above expression, we can infer that the iterates generated from the proximal gradient algorithm has linear convergence. Now to obtain the convergence of the functional value we use the Equation 11 as

$$
\begin{aligned}
F\left(\mathbf{x}_{k+1}\right)-F^* &\leq \frac{L-\mu}{2}\left\|\mathbf{x}_k-\mathbf{x}^*\right\|^2-\frac{L}{2}\left\|\mathbf{x}_{k+1}-\mathbf{x}^*\right\|^2 \\
&\overset{(a)}{\leq} \frac{L-\mu}{2}\left\|\mathbf{x}^k-\mathbf{x}^*\right\|^2 \\
&= \frac{L}{2}\left(1-\frac{\mu}{L}\right)\left\|\mathbf{x}^k-\mathbf{x}^*\right\|^2 \\
&\overset{(b)}{\leq} \frac{L}{2}\left(1-\frac{\mu}{L}\right)^{k+1}\left\|\mathbf{x}^0-\mathbf{x}^*\right\|^2
\end{aligned}
$$

where the expression in (a) is obtained from non negativity of the norm and the expression (b) is obtained by employing the iterate convergence as given above. □

2.1.1 For projected GD

The analysis for projected gradient method follows the same procedure with $g=\mathbf{I}_{\mathcal{C}}$ the proximal operator reduces to the projection operator $P=\Pi_{\mathcal{C}}$. The inequality for the projected GD modifies as

Lemma 5. *Consider any composite function* $F(\cdot)$ *as defined in* (1)*. For any* $\mathbf{x} \in \mathcal{C}, \mathbf{y} \in(\mathbf{dom}(f))$ *and* $L>0$ *satisfies*

$$
f(\mathbf{x})-f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \geq \frac{L}{2}\left\|\mathbf{x}-\Pi_{\mathcal{C}}(\mathbf{y})\right\|^2-\frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2+\ell_f(\mathbf{x}, \mathbf{y})
$$

where,

$$
\ell_f(\mathbf{x}, \mathbf{y})=f(\mathbf{x})-f(\mathbf{y})-\langle\nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle
$$

Proof. For the analysis of the theorem consider the function defined as

$$
\varphi(\mathbf{u})=f(\mathbf{y})+\langle\nabla f(\mathbf{y}), \mathbf{u}-\mathbf{y}\rangle+\mathbf{I}_{\mathcal{C}}+\frac{L}{2}\|\mathbf{u}-\mathbf{y}\|^2 \qquad (12)
$$

Since the function φ is L-strongly convex function and projection $\Pi_{\mathcal{C}}(\mathbf{y})=\underset{\mathbf{u} \in \mathcal{C}}{\operatorname{argmin}} \varphi(\mathbf{u})$. As the function $\varphi(\cdot)$ is L strongly convex we deduce the following expression

$$
\varphi(\mathbf{x})-\varphi\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \geq \frac{L}{2}\left\|\mathbf{x}-\Pi_{\mathcal{C}}(\mathbf{y})\right\|^2 .
$$

Now from the definition of $\varphi(\cdot)$ as in (7),

$$
\begin{aligned}
\varphi\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) &= f(\mathbf{y})+\left\langle\nabla f(\mathbf{y}), \Pi_{\mathcal{C}}(\mathbf{y})-\mathbf{y}\right\rangle+\frac{L}{2}\left\|\Pi_{\mathcal{C}}(\mathbf{y})-\mathbf{y}\right\|^2+\mathbf{I}_{\mathcal{C}}(P(\mathbf{y})) \\
&\overset{(a)}{\geq} f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right)+\mathbf{I}_{\mathcal{C}}\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \\
&\overset{(b)}{=} f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) .
\end{aligned}
$$

The expression (a), is obtained from the definition of L smoothness for the function f and (b) follows as the projected vector is in the constraint set $\mathcal{C}$. Thus for any $\mathrm{x} \in \mathbf{dom}(f)$,

$$\varphi(\mathbf{x}) - f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \geq \frac{L}{2}\left\|\mathbf{x} - P(\mathbf{y})\right\|^2 .$$

Plugging the expression for $\varphi(\mathrm{x})$ in (7), into the above inequality, we obtain the expression

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \mathbf{I}_{\mathcal{C}}(\mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 - f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \geq \frac{L}{2}\left\|\mathbf{x} - P(\mathbf{y})\right\|^2$$

the final expression is obtained by introducing the term $f(\mathbf{x})$. On rearranging the expression we obtain the desired expression

$$f(\mathbf{x}) - f\left(\Pi_{\mathcal{C}}(\mathbf{y})\right) \geq \frac{L}{2}\left\|\mathbf{x} - \Pi_{\mathcal{C}}(\mathbf{y})\right\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y})$$

□

The Linear convergence proof follows similar to the proximal GD analysis.

2.2 Analysis 2

The analysis is from the lecture series [2]. Let $x^* \in \mathcal{X}^*$ denote the optimal solution of (1). The solution x^* is unique owing to strong convexity of $F(\cdot)$. We know that the set of minimizers is same as the set of fixed points of the proximal operator i.e., $\mathcal{X}^* = \operatorname{fix} P$.

The update of the proximal is given as $P(\mathbf{x}) = \operatorname{prox}_{L^{-1}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)$. Observe that

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^*\| &= \|P(\mathbf{x}_k) - P(\mathbf{x}^*)\| \\
&= \left\|\operatorname{prox}_{L^{-1}g}\left(\mathbf{x}_k - \frac{1}{L}\nabla f\left(\mathbf{x}_k\right)\right) - \operatorname{prox}_{L^{-1}g}\left(\mathbf{x}^* - \frac{1}{L}\nabla f\left(\mathbf{x}^*\right)\right)\right\| \\
&\leq \left\|\mathbf{x}_k - \frac{1}{L}\nabla f\left(\mathbf{x}_k\right) - \mathbf{x}^* + \frac{1}{L}\nabla f\left(\mathbf{x}^*\right)\right\| \qquad (13)
\end{aligned}$$

The last inequality follows from the non expansiveness of the proximal operator.

Since the function f is strongly convex and has a Lipschitz continuous gradient, it follows that for all vectors $\mathbf{x}$ and $\mathbf{y}$ and all positive scalars η

$$\begin{aligned}
\|\mathbf{x} - \eta\nabla f(\mathbf{x}) - (\mathbf{y} - \eta\nabla f(\mathbf{y}))\| &\leq \| \int_0^1 \left(I - \eta\nabla^2 f(\mathbf{x} + \eta(\mathbf{y} - \mathbf{x})\right)^T(\mathbf{y} - \mathbf{x}) d\eta \| \\
&\overset{(a)}{\leq} \int_0^1 \left\|I - \eta\nabla^2 f(\mathbf{x} + \eta(\mathbf{y} - \mathbf{x}))^T(\mathbf{y} - \mathbf{x})\right\| d\eta \\
&\overset{(b)}{\leq} \int_0^1 \left\|I - \eta\nabla^2 f(\mathbf{x} + \eta(\mathbf{y} - \mathbf{x}))\right\| \|(\mathbf{y} - \mathbf{x})\| \, d\eta \\
&\overset{(c)}{\leq} \sup_{\mathbf{z}} \left\|I - \eta\nabla^2 f(\mathbf{z})\right\| \|\mathbf{y} - \mathbf{z}\|.
\end{aligned}$$

Here, the expression (a) follows from triangular inequality of the norm, (b) follows from Cauchy-Schwartz inequality and (c) is obtained by taking supreme over the range of η. Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least μ and the maximum eigenvalue is at most L. Therefore the eigenvalues of $I - \eta\nabla^2 f(z)$ are at most $\max(1 - \eta L, 1 - \eta\mu)$ and at least $\min(1 - \eta L, 1 - \eta\mu)$. Therefore, $\left\|I - \eta\nabla^2 f(z)\right\| \leq \max(|1 - \eta L|, |1 - \eta\mu|)$. In particular, using this upper bound in (13), we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \max\left\{\left|1 - \eta L\right|, \left|1 - \eta\mu\right|\right\} \|\mathbf{x} - \mathbf{y}\|.$$

$\eta = \frac{2}{L+\mu}$ minimizes the right hand side for all k. Setting η to this value, we obtain

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \left(\frac{L-\mu}{L+\mu}\right) \|\mathbf{x}_k - \mathbf{x}^*\|,$$

denoting $\kappa = \frac{L}{\mu}$ and $D_0 = \|\mathbf{x}_0 - \mathbf{x}^*\|$,

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k D_0$$

. Hence, for strongly convex f and arbitrary P, the proximal gradient algorithm converges at a linear rate under a constant step-size policy. To see the convergence of the suboptimality gap, we employ the L-smoothness with $\mathbf{y} = \mathbf{x}_k$ and $\mathbf{x} = \mathbf{x}^*$, we get

$$\begin{aligned} F(\mathbf{x}_k) - F(\mathbf{x}^*) &\leq \frac{L}{2}||\mathbf{x}_k - \mathbf{x}^*||^2 \\ &\leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^k D_0. \end{aligned}$$

The above expression is obtained from the convergence of the iterates of the proximal GD. The same analysis holds for projected GD.

3 Numerical experiments

In this section we perform numerical analysis to verify the performance of the proposed algorithm on real datasets. Here we consider LASSO regression using proximal GD to analysis the performance on `California housing` and the performance of projected GD on `California housing` and `MovieLens10k` datasets.

3.1 Lasso

Lasso performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting model. The optimization problem is least squares with a l_1 regularizer that ensures that the model is sparse and is given by,

$$\frac{1}{2}||\mathbf{Ax} - \mathbf{y}||_\mathbf{2}^\mathbf{2} \quad \text{s.to} \quad ||\mathbf{x}||_1 \leq t.$$

Here, $\mathbf{A}$ represents the observation matrix and $\mathbf{y}$ is the dependent variables that depend on $\mathbf{A}$. The performance of the proximal GD is observed for Lasso formulation on `California housing` and the optimal objective function value is calculated using lasso function from `scikit-learn` [3]. The proximal operator is the soft thresholding operator and is given as

$$\text{prox}_g(\mathbf{x})_i = \begin{cases} \mathbf{x}_i + \lambda & \mathbf{x}_i < -\lambda \\ 0 & -\lambda \leq \mathbf{x}_i \leq \lambda \\ \mathbf{x}_i - \lambda & \mathbf{x}_i > \lambda \end{cases}$$

3.1.1 Synthetic data

Here we consider synthetic data with $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{y} \in \mathbb{R}^n$. We considered data with $n = 10$, and the performance of the proximal GD with step size $\eta = 0.0003$ is given in Fig: 1. The suboptimality gap and the iterate shows linear convergence but the rate of decrease is slow for iterates as compared to the suboptimality gap.

3.1.2 Real data

The convergence of the proximal GD is evaluated on the `California housing` dataset with $m = 3000$ instances and $n = 8$ features. The performance of the proximal GD with step size $\eta = 0.385$ is given in Fig: 2. The suboptimality gap and the iterate shows linear convergence but the rate of decrease is slow for iterates as compared to the suboptimality gap.

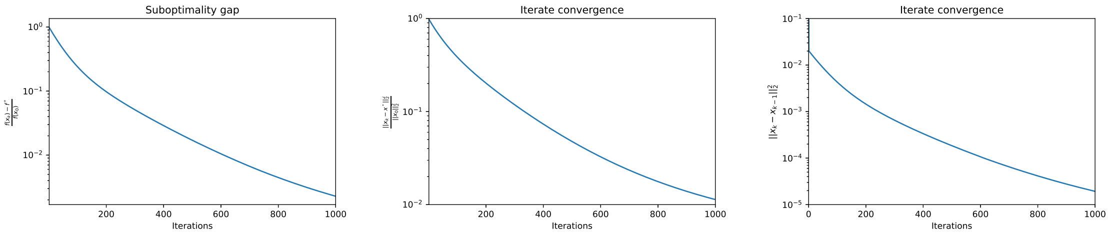

Figure 1: Performance of the proximal GD algorithm on synthetic data (a) Suboptimality gap (b) convergence of iterate to optimal solution (c) error between consecutive iterates.

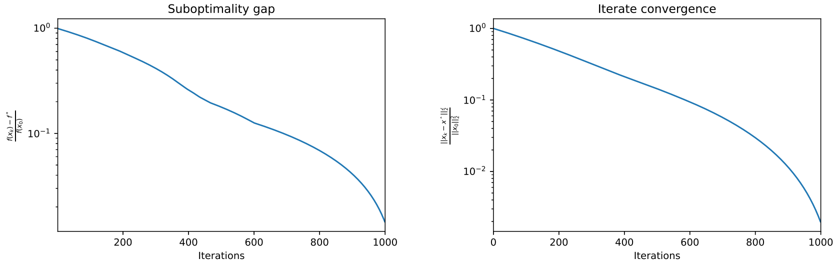

Figure 2: Performance of the proximal GD algorithm on real dataset (a) Suboptimality gap (b) convergence of iterate to optimal solution.

3.2 Ridge regression

Next, we consider ridge regression using a Least squared loss function with an l_2-norm constraint [4]:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & ||\mathbf{A}\mathbf{x} - \mathbf{y}||_2^2 \\ \text{s. to} \quad & ||\mathbf{x}||_2 \leq t, \end{aligned}$$

Here, $\mathbf{A}$ represents the observation matrix and $\mathbf{y}$ is the dependent variables that depend on $\mathbf{A}$. We apply ridge regression on `California housing` dataset with $m = 3000$ observations and each observations has $d = 8$ features.

3.2.1 Synthetic data

Here we consider synthetic data with $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{y} \in \mathbb{R}^n$. We considered data with $n = 10$, and the performance of the projected GD with $t = 15$, and step size $\eta = 0.0003$ is given in Fig: 3. The suboptimality gap and the iterate shows linear convergence.

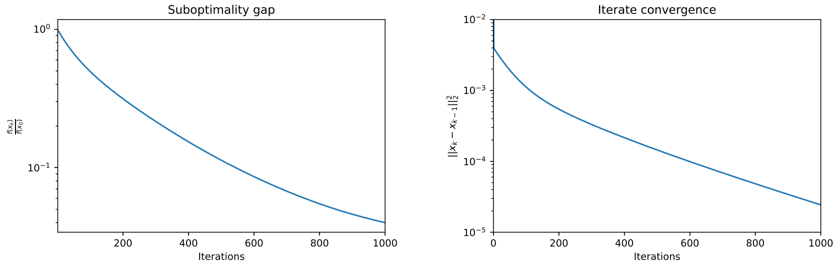

Figure 3: Performance of the projected GD algorithm on synthetic data (a) Suboptimality gap (b) convergence of iterate to optimal solution.

3.2.2 Real data

The convergence of the proximal GD is evaluated on the `California housing` dataset with $m = 3000$ instances and $n = 8$ features. The performance of the projected GD with $t = 15$, and step size $\eta = 0.385$ is given in Fig: 4. The suboptimality gap and the iterate shows linear convergence.

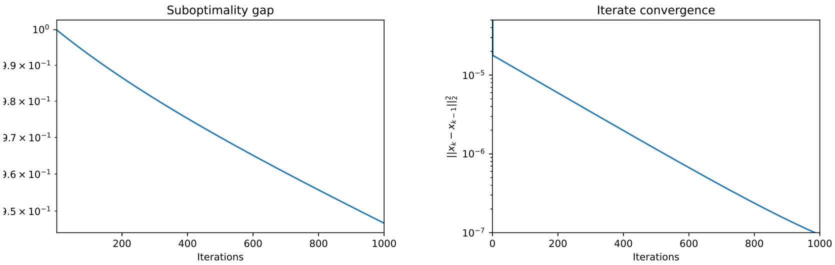

Figure 4: Performance of the projected GD algorithm on real dataset (a) Suboptimality gap (b) convergence of iterate to optimal solution.

3.3 Matrix completion

The matrix completion problem is to predict the unobserved entries from a fraction of observed entries which can be also corrupted. There are many ways to the same problem and we will be fill the missing entries with the assumption that the final matrix is low rank. The above optimization can be formulated as,

$$\frac{1}{N}||\mathbf{X} - \mathbf{A}||_F^2 \quad \text{s.to} \quad ||\mathbf{X}||_* \leq R,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the partially observed matrix with the entries A_{ij} available for $(i, j) \in \Omega$. The sampling mask Ω is know and the nuclear norm constraint $||\mathbf{X}||_* \leq R$ is employed to ensure that the solution is low rank. We apply robust matrix completion on the *Movielens 100K* [5] dataset with $m = 1682$ movies rated by $n = 943$ users with 6.30% percent ratings observed.. The performance of the algorithms is evaluated with respect to $\frac{1}{N}||\mathbf{X}_k - \mathbf{A}||_F^2$, where $\mathbf{X}_k$ represents the output of the algorithm at kth iterate. The projection operator onto observed set:

$$[P_\Omega(B)]_{ij} = \begin{cases} B_{ij} & (i, j) \in \Omega \\ 0 & (i, j) \notin \Omega \end{cases}$$

The performance of the projected gradient descent is shown in Fig: 5. The convergence is almost linear barring some jitters but the rate of convergence is slow as compared to the theoretical results.

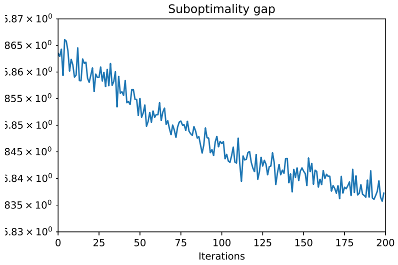

Figure 5: Performance of the projected GD algorithm on *Movielens 100K* dataset

4 Conclusion

The proximal gradient methods is seen to have linear convergence for smooth and strongly convex objectives. The experimental results also shows linear convergence. The same proofs of the proximal gradient methods holds for projected gradient methods and the experimental results shows similar performance.

References

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

[2] B. Recht, “Projected gradient methods,” 2012. [Online]. Available: https://pages.cs.wisc.edu/~brecht/cs726docs/ProjectedGradientMethods.pdf

[3] F. Pedregosa, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] A. B. Owen, “A robust hybrid of lasso and ridge regression,” *Contemporary Mathematics*, vol. 443, no. 7, pp. 59–72, 2007.

[5] “Movielens 100k dataset,” GroupLens Research, https://grouplens.org/datasets/movielens/100k/, accessed on 09/10/2021.