CS2910 - Optimization Summer 2023

Lecture 6 — Proximal Gradient Descent

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

In this lecture, we focus on how to solve the optimization problem for much more families of convex functions. We will generalize a method of gradient descent named *the proximal gradient descent*. Similarly to how to generalize Wasserstein gradient flow, we will introduce a gradient-like term — sub-gradient — to illustrate how to well define the descent along the gradient. Based on the term we introduced, we will show the analysis of its rate of convergence.

2 Non-Smooth Convex Function and Sub-Gradient

Recall the optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

If $f \in C^1(\mathbb{R}^d)$ is a convex function, the gradient of f is well-defined and we safely employ gradient descent to get an approximate solution to it (without consider the rate of convergence). However, if f is still continuous and convex but not necessarily a C^1 function, gradient descent fails since it seems hard to say the existence of the gradients at every point of f. We need to introduce a new term to describe the structure of f.

Definition 1 (sub-gradient)**.** *Given a convex function $f : \mathbb{R}^d \cup \{\infty\}$, we say a vector $p \in \mathbb{R}^d$ is a sub-gradient of f at point $x \in \mathbb{R}^d$ if*

$$f(y) \geq f(x) + \langle y - x, p \rangle, \forall y \in \mathbb{R}^d.$$

The collection of sub-gradients at point x, denoted by $\partial f(x)$, is called the sub-differential of f at x.

Remark 1. When $f \in C^1(\mathbb{R}^d)$ is convex, for all $x \in \mathbb{R}^d$, $\partial f(x) = \{\nabla f(x)\}$; since f is a convex function, the optimal point $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d} \{f(x)\}$ is equivalent to $0 \in \partial f(x)$.

Examples: Consider $f(x) = \|x\|_2$. It is not hard to see f is a continuous convex function.

- When $x \neq 0$, f is a C_1 function. Then $\partial f(x) = \{\nabla f(x)\} = \{\text{sign}(x)\}$.
- When $x = 0$. We claim, $\partial f(0) = \left\{p \in \mathbb{R}^d : \|p\|_2 \leq 1\right\}$.

  Proof. For all $\|p\|_2 \leq 1$, it holds that

  $$\langle p, y\rangle \leq \|p\|_2 \cdot \|y\|_2 \leq \|y\|_2.$$

  This means $p \in \partial f(0)$. On the other hand, for all $p \in \partial f(0)$, it holds that for all $y \in \mathbb{R}^d$, $\|y\|_2 \geq \langle p, y\rangle$. We choose $y = p$, then $\|p\|_2^2 \leq \|p\|_2$, thus leading to $\|p\|_2 \leq 1$. □

3 Proximal Gradient Descent

Now we introduce how to solve the optimization problem (1) when f is a non-smooth convex function.

$$\min_{x\in\mathbb{R}^d} f(x) = g(x) + h(x) \tag{2}$$

where g is a function with some 'nice' properties, e.g., L-smoothness and convexity, and h is a function with some special structures, but might be non-differentiable.

Examples:

- Lasso function $f(\beta) = \frac{1}{2}\|X\beta - y\|_2^2 + \|\beta\|_1$.
- The convex constraints $f(x) = g(x) + \iota_D(x)$.

For this kind of optimization problem, it seems that gradient descent does not work, due to the function h.

Recall that, in each step of gradient descent, we choose x_{k+1} as

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|_2^2 \right\}.$$

Although f is non-differentiable, g is differentiable. This inspires us to just make quadratic approximation to g and leave h alone.

Now, consider

$$\begin{aligned} x_* &= \operatorname*{argmin}_{x\in\mathbb{R}^d} \left\{ g(x_k) + \langle \nabla g(x_k), x - x_k\rangle + \frac{1}{2\eta}\|x - x_k\|_2^2 + h(x) \right\} \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d} \left\{ \frac{1}{2\eta}\|x - (x_k - \eta\nabla g(x_k))\|_2^2 + h(x) \right\}. \end{aligned}$$

The first term means x_* must be located near the local minimum of g, and the second term means x_* should not make h large.

Definition 2. *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex function. We define the proximal mapping $\mathrm{prox}_h : \mathbb{R}^d \to \mathbb{R}^d$ as*

$$\mathrm{prox}_h(x) = \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \frac{1}{2}\|z - x\|_2^2 + h(z) \right\}.$$

Then we can describe the proximal gradient descent. Given a step size $\eta > 0$, at each step k, for x_k, we update

$$x_{k+1} = \mathrm{prox}_{\eta h}(x_k - \eta \nabla g(x_k)).$$

We would rewrite it as the following form:

$$x_{k+1} = x_k - \eta G_\eta(x_k)$$

where the function $G_\eta(x)$ is generalized by

$$G_\eta(x) = \frac{x - \mathrm{prox}_{\eta h}(x - \eta \nabla g(x))}{\eta}.$$

Remark 2. Since h is convex and the function $z \mapsto \frac{1}{2}\|z - x\|_2^2$ is strongly convex, the function $\frac{1}{2}\|z - x\|_2^2 + h(z)$ has the unique minimizer, which means the function prox_h is well-defined. Additionally, we would say that, under the assumption that h has some special structure, prox_h is easy to compute.

Now we give an example for proximal gradient descent. We would view projected gradient descent

$$\min_{z \in \mathbb{R}^d} \frac{1}{2}\|x - z\|_2^2 + \iota_D(z).$$

as a kind of proximal gradient descent. By definition, we compute

$$\begin{aligned} \mathrm{prox}_{\eta h}(x) &= \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \eta \iota_D(x) + \frac{1}{2}\|x - z\|_2^2 \right\} \\ &= \operatorname*{argmin}_{z \in D} \left\{ \frac{1}{2}\|x - z\|_2^2 \right\} = \Pi_D(x). \end{aligned}$$

This means the projection is some kind of proximal mapping.

3.1 Interpretation of proximal gradient descent

In this part, we give some interpretations of proximal gradient descent. Let $y_k \overset{\triangle}{=} x_k - \eta \nabla g(x_k)$. We write the update of proximal gradient descent as

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \frac{1}{2}\|z - y_k\| + \eta h(z) \right\} \\ &= \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \frac{1}{2\eta}\|z - y_k\| + h(z) \right\} \end{aligned}$$

Intuitively, proximal gradient descent minimizes g and h, instead of only minimizing g. Since when $z \neq y_k$, $\|z - y_k\| > 0$, by the optimality of x_{k+1}, it must hold that $h(x_{k+1}) < h(y)$. The larger $\|z - y_k\|$ is, the smaller $h(z) - h(y)$ is. This intuition can be seen as a generalization of the interpretation of each step of gradient descent (or mirror descent).

3.1.1 Proximal gradient descent and backward Euler method

If h is differentiable, we write

$$x_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ \frac{1}{2\eta} \|z - y_t\| + h(z) \right\}.$$

It is equivalent to

$$\frac{1}{\eta} (x_{k+1} - y_k) + \nabla h(x_{k+1}) = 0.$$

This means x_{k+1} is the solution to the following ODE

$$x_{k+1} = y_k - \eta \nabla h(x_{k+1}). \tag{3}$$

If $y_k = x_k$, then $h = f$ and (3) is exactly the *backward Euler method*. We compare it with the *Euler method*

$$x_{k+1} = x_k - \eta \nabla f(x_k). \tag{4}$$

Note that (4) is an explicit form of the update rule and (3) is an implicit form of update. Usually (3) has more precise approximation and faster convergence, but for implementation (4) is more common and useful.

3.2 Rate of convergence

Now we analyze the rate of the convergence of proximal gradient descent. The analysis is similar to what we do in gradient descent and mirror descent.

Lemma 3 (mirror descent lemma for the proximal step)**.** *Given $f = g + h$ where g is a differentiable convex function and h is a convex function, and a step size $\eta > 0$, let $\{x_k\}_{k \in \mathbb{N}}$ be the proximal gradient descent generated by f. For $k \in \mathbb{N}$, define $y_k = x_k - \eta \nabla g(x_k)$. Then for all $x \in \mathbb{R}^n$, it holds that*

$$h(x_{k+1}) \leq h(x) + \frac{1}{2\eta} \left(\|y_k - x\|_2^2 - \|x_{k+1} - x\|_2^2 - \|y_k - x_{k+1}\|_2^2 \right).$$

Proof. Since $x_{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^d} \left\{ \frac{1}{2} \|z - y_k\|_2^2 + \eta h(z) \right\}$, it holds that

$$\frac{y_k - x_{k+1}}{\eta} \in \partial h(x_{k+1})$$

By lower linear bound, for all $x \in \mathbb{R}^d$, it holds that

$$\begin{aligned}h(x_{k+1}) &\le h(x) - \frac{1}{\eta}\langle y_k - x_{k+1}, x_{k+1} - x\rangle \\ &= h(x) + \frac{1}{2\eta}\left(\|y_k - x\|^2 - \|x_{k+1} - x\|_2^2 - \|y_k - x_{k+1}\|^2\right)\end{aligned}$$

where the last equality holds by the law of cosines. □

Now we establish the convergence rate of proximal gradient descent.

Proposition 4 (convergence rate of proximal gradient descent)**.** *Given $f = g + h$ where g is a differentiable convex L-smooth function and h is a convex function, and a step size $\eta \le 1/L$, let $\{x_k\}_{k\in\mathbb{N}}$ be the proximal gradient descent generated by f. Let $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d}\{f(x)\}$. Then, we have*

$$\frac{1}{T}\sum_{k=1}^{T} f(x_k) \le f(x_*) + \frac{\|x_* - x_0\|^2}{2\eta T}.$$

Proof. Since g is an L-smooth convex function, it holds that

$$\begin{aligned}g(x_{k+1}) &\le g(x_k) + \langle \nabla g(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \\ &\le g(x_*) - \langle \nabla g(x_k), x_* - x_k\rangle + \langle \nabla g(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= g(x_*) + \langle \nabla g(x_k), x_{k+1} - x_*\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= g(x_*) + \frac{1}{\eta}\langle x_k - y_k, x_{k+1} - x_*\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.\end{aligned}$$

With Lemma 3 applied, we obtain

$$\begin{aligned}f(x_{k+1}) &\le f(x_*) - \frac{1}{\eta}\langle x_k - x_{k+1}, x_* - x_{k+1}\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_*) - \frac{1}{2\eta}\left(\|x_k - x_{k+1}\|^2 + \|x_* - x_{k+1}\|^2 - \|x_* - x_t\|^2\right) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_*) - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x_k - x_{k+1}\|^2 + \frac{1}{2\eta}\|x_* - x_{k+1}\|^2 - \frac{1}{2\eta}\|x_* - x_t\|^2.\end{aligned}$$

When $\eta \le 1/L$, it holds that

$$f(x_{k+1}) \le f(x_*) + \frac{1}{2\eta}\|x_* - x_{k+1}\|^2 - \frac{1}{2\eta}\|x_* - x_t\|^2.$$

Summing over both sides we prove the proposition. □