IE 539: Convex Optimization **KAIST, Fall 2022**
Lecture #17: Proximal gradient descent November 3, 2022
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Proximal operator and proximal gradient descent.
- Convergence of proximal gradient descent.

2 Proximal gradient descent

Recall the formulation of LASSO, given by

$$\min_{\beta} \quad \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Here, the objective function is non-differentiable because of the ℓ_1-regularization term $\lambda\|\beta\|_1$, and therefore, it is non-smooth. On the other hand, the objective is convex, and we have a characterization of the subdifferential of $\|\beta\|_1$, so we can simply apply the subgradient method. To bound the additive error by ϵ, the subgradient method requires $O(1/\epsilon^2)$ iterations.

If you take a closer look at the objective, it consists of two part. One part is smooth, and the other part is something whose subdifferential is well understood. Can we use this structure to obtain a better algorithm? The main subject of this section is developing an algorithm that converges to an ϵ-approximate solution after $O(1/\epsilon)$ iterations.

2.1 Projection and proximal operator

We studied the projected gradient descent method, where at each step, we take a projection to the constraint set. When the constraint set is given by C, the projection operator is given by

$$\mathrm{proj}_C(x) = \operatorname*{argmin}_{u\in C} \frac{1}{2}\|u - x\|_2^2 = \operatorname*{argmin}_{u\in\mathbb{R}^d} \left\{ I_C(u) + \frac{1}{2}\|u - x\|_2^2 \right\}$$

where $I_C(u)$ is the indicator function of C. This definition is proper as there is a unique minimizer for the optimization problem. Hence, the projection operator is defined by the indicator function and the proximity term $(1/2)\|u - x\|_2^2$. The proximal operator is a generalization of the projection operator replacing the indicator function by other general functions.

The proximal operator with respect to a convex function h is defined as follows.

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

Again the definition is proper because the objective of the optimization problem is strongly convex. Hence, for any $\eta > 0$,

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

As projected gradient descent proceeds with the update rule

$$x_{t+1} = \mathrm{proj}_C \left\{ x_t - \eta \nabla f(x_t) \right\},$$

we can defined the proximal gradient method with the update rule

$$x_{t+1} = \mathrm{prox}_{\eta h}(x_t - \eta \nabla f(x_t)).$$

In particular, when we take the indicator function I_C for h, the proximal gradient method reduces to the projeced gradient descent method.

Lemma 17.1. *$u = prox_{\eta h}(x)$ if and only if $x - u \in \eta \partial h(u)$.*

Proof. Note that $u = \mathrm{prox}_{\eta h}(x)$ means that u minimizes $h(u) + (1/2\eta)\|u - x\|_2^2$. By the optimality condition, it is equivalent to $0 \in \partial h(u) + \{(1/\eta)(u - x)\}$, and this is equivalent to $x - u \in \eta \partial h(u)$. □

2.2 Example: ℓ_1 regularization

Consider $h(x) = \|x\|_1$. Then

$$\mathrm{prox}_{\eta h}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ \|u\|_1 + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

Let $u = \mathrm{prox}_{\eta h}(x)$. Then, by Lemma 17.1,

$$x - u \in \eta \partial \|u\|_1.$$

Recall that $g \in \partial \|u\|_1$ if and only if

$$g_i = \begin{cases} \mathrm{sign}(u_i), & \text{if } u_i \neq 0, \\ \text{a value in } [-1, 1], & \text{if } u_i = 0. \end{cases}$$

Based on this, we can argue that $x - u \in \eta \partial \|u\|_1$ if and only if

$$u_i = \begin{cases} x_i - \eta, & \text{if } x_i \geq \eta, \\ 0, & \text{if } -\eta \leq x_i \leq \eta. \\ x_i + \eta, & \text{if } x_i \leq -\eta. \end{cases}$$

Moreover, $x - u \in \eta \partial \|u\|_1$ if and only if

$$u_i = \max\{0, |x_i| - \eta\} \cdot \mathrm{sign}(x_i).$$

For example,

$$\mathrm{prox}_h((3, 1, -2)^\top) = (2, 0, -1)^\top.$$

Note that when $h = \|x\|_1$, the corresponding proximal operator "shrinks" the vector. For this reason, the operator is called the self-thresholding operator or the shrinkage operator.

2.3 Example: quadratic function

Consider $h(x) = (1/2)x^\top Ax + b^\top x + c$ where A is positive semidefinite. Then

$$\mathrm{prox}_{\eta h}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ \frac{1}{2}u^\top Au + b^\top u + c + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

Setting $v = \mathrm{prox}_{\eta h}(x)$, it follows from the optimality condition that

$$0 = Av + b + \frac{1}{\eta}(v - x).$$

Therefore,

$$\mathrm{prox}_{\eta h}(x) = v = (I + \eta A)^{-1}(x - \eta b).$$

2.4 Convergence of proximal gradient descent

We consider the following composite convex optimization problem.

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

where we assume that g is a smooth convex function and h is convex. The proximal gradient algorithm applies to this composite problem proceeds with the following update rule.

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

Algorithm 1 Proximal gradient descent

```
Initialize $x_1 \in C$.
for $t = 1, \dots, T$ do
    Update $x_{t+1} = \mathrm{prox}_{\eta h}(x_t - (1/\beta)\nabla g(x_t))$ where $\beta$ is the smoothness parameter of $g$.
end for
Return $x_{T+1}$.
```

The gradient mapping is defined as

$$G_\eta(x) = \frac{1}{\eta}\left(x - \mathrm{prox}_{\eta h}(x - \eta\nabla g(x))\right).$$

Here, $-\eta G_\eta(x)$ is equal to $\mathrm{prox}_{\eta h}(x - \eta\nabla g(x)) - x$, which is the difference between the current point x and the one obtained after the proximal gradient update applied to x. Then

$$x_{t+1} = x_t - \eta G_\eta(x_t).$$

Note that when h is the indicator function of $\mathbb{R}^d$, the gradient mapping is simply $\nabla g(x)$. Hence, the gradient mapping operator is similar in spirit to the gradient operator. In fact, we can derive the following optimality condition in terms of the gradient mapping.

Lemma 17.2. *$G_\eta(\hat{x}) = 0$ if and only if $\hat{x} \in \mathit{argmin}_{x\in\mathbb{R}^d}\, g(x) + h(x)$.*

Proof. By the optimality condition, $\hat{x}$ minimizes $g + h$ if and only if

$$\begin{aligned} 0 \in \{\nabla g(\hat{x})\} + \partial h(\hat{x}) \quad &\leftrightarrow \quad -\nabla g(\hat{x}) \in \partial h(\hat{x}) \\ &\leftrightarrow \quad (\hat{x} - \eta\nabla g(\hat{x})) - \hat{x} \in \eta\partial h(\hat{x}) \\ &\leftrightarrow \quad \hat{x} = \mathrm{prox}_{\eta h}(\hat{x} - \eta\nabla g(\hat{x})) \end{aligned}$$

where the last equivalence comes from Lemma 17.1. Note that $\hat{x} = \mathrm{prox}_{\eta h}(\hat{x} - \eta\nabla g(\hat{x}))$ is equivalent to

$$G_\eta(\hat{x}) = \frac{1}{\eta}\left(\hat{x} - \mathrm{prox}_{\eta h}(\hat{x} - \eta\nabla g(\hat{x}))\right) = 0$$

Therefore, $\hat{x}$ is a minimizer of $g + h$ if and only if $G_\eta(\hat{x}) = 0$. □

To analyze the convergence of proximal gradient descent, we need the following lemma.

Lemma 17.3. *Consider $f = g + h$ where g is β-smooth and α-strongly convex in the ℓ_2 norm and h is convex. Assume that $\beta > 0$ and $\alpha \geq 0$. Then for any x, z,*

$$f\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq f(z) + G_{1/\beta}(x)^\top (x - z) - \frac{1}{2\beta}\|G_{1/\beta}(x)\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2.$$

Proof. As $f = g + h$, we upper bound g and h separately, thereby bounding f. Note that

$$\begin{aligned} g\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) &\leq g(x) + \nabla g(x)^\top \left(\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) - x\right) + \frac{\beta}{2}\left\|\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) - x\right\|_2^2 \\ &= g(x) - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2 \\ &\leq g(z) - \nabla g(x)^\top (z - x) - \frac{\alpha}{2}\|z - x\|_2^2 - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2 \end{aligned} \tag{17.1}$$

where the first inequality is due to the β-smoothness of g and the second inequality is due to the α-strong convexity of g.

Next we consider the h part. By Lemma 17.1,

$$u = \text{prox}_{(1/\beta)h}(x - (1/\beta)\nabla g(x)) = x - \frac{1}{\beta}G_{1/\beta}(x)$$

if and only if

$$\left(x - \frac{1}{\beta}\nabla g(x)\right) - \left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \in \frac{1}{\beta}\partial h\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right).$$

Multiplying each side by β, it is equivalent to

$$G_{1/\beta}(x) - \nabla g(x) \in \partial h\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right).$$

Then it follows from the convexity of h that

$$h\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq h(z) - \left(G_{1/\beta}(x) - \nabla g(x)\right)^\top \left(z - \left(x - \frac{1}{\beta}G_{1/\beta}(x)\right)\right). \tag{17.2}$$

Combining (17.1) and (17.2), we get

$$f\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq f(z) - G_{1/\beta}(x)^\top (z - x) - \frac{1}{2\beta}\|G_{1/\beta}(x)\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2,$$

as required. □

One would find that Lemma 17.3 is analogous to the lemma stating that the gradient descent with step size $1/\beta$ always improves for a β-smooth function. In fact, plugging in $z = x$, we obtain

$$f\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq f(x) - \frac{1}{2\beta}\|G_{1/\beta}(x)\|_2^2. \tag{17.3}$$

The next step we took for smooth functions was to use $f(x) \leq f(x^*) - \nabla f(x)^\top (x^* - x)$. However, as $\nabla f(x) \neq G_{1/\beta}(x)$, we cannot directly use (17.3). Instead, we start from Lemma 17.3 by plugging in $z = x^*$ and $x = x_t$. Then

$$\begin{aligned}
f(x_{t+1}) &\le f(x^*) + G_{1/\beta}(x)^\top (x_t - x^*) - \frac{1}{2\beta}\|G_{1/\beta}(x_t)\|_2^2 - \frac{\alpha}{2}\|x_t - x^*\|_2^2 \\
&= f(x^*) + \frac{\beta}{2}\left(\|x_t - x^*\|_2^2 - \left\|x_t - x^* - \frac{1}{\beta}G_{1/\beta}(x_t)\right\|_2^2\right) - \frac{\alpha}{2}\|x_t - x^*\|_2^2 \\
&= f(x^*) + \frac{\beta}{2}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) - \frac{\alpha}{2}\|x_t - x^*\|_2^2.
\end{aligned}$$

This implies that

$$f(x_{t+1}) - f(x^*) \le \frac{\beta}{2}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) - \frac{\alpha}{2}\|x_t - x^*\|_2^2. \tag{17.4}$$

Theorem 17.4. *Let $f = g + h$ where g is a β-smooth convex function in the ℓ_2 norm and h is convex. Then x_{T+1} returned by Proximal Gradient Descent (Algorithm 1) satisfies*

$$f(x_{T+1}) - f(x^*) \le \frac{\beta\|x_1 - x^*\|_2^2}{2}.$$

Proof. First, sum up (17.4) for $t = 1, \dots, T$ and then divide each side by T. Then we obtain

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}) - f(x^*) \le \frac{\beta}{2}\left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\right) - \frac{\alpha}{2}\sum_{t=1}^{T}\|x_t - x^*\|_2^2.$$

By (17.3), we know that $f(x_{T+1}) \le f(x_T) \le \cdots \le f(x_2)$. Moreover, $\|x_t - x^*\|_2 \ge 0$. Thus the left-hand side is greater than or equal to $f(x_{T+1}) - f(x^*)$ and the right-hand side is at most $(\beta/2)\|x_1 - x^*\|_2^2$. □

Furthermore, when α is strictly positive, in which case, g is strongly convex, we deduce the following convergence result.

Theorem 17.5. *Let $f = g + h$ where g is β-smooth and α-strongly convex in the ℓ_2 norm and h is convex. Then x_{T+1} returned by Proximal Gradient Descent (Algorithm 1) satisfies*

$$\|x_{T+1} - x^*\|_2^2 \le \left(1 - \frac{\alpha}{\beta}\right)^T \|x_1 - x^*\|_2^2.$$

Proof. Note that the left-hand side of (17.4) is greater than or equal to 0, and so is the right-hand side. Then it follows that

$$\|x_{t+1} - x^*\|_2^2 \le \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|_2^2,$$

as required. □

2.5 ISTA and FISTA

In the last section, we discussed proximal gradient descent and its convergence. Next we apply proximal gradient descent to solve LASSO. We consider

$$\min_{\beta} \quad f(\beta) = g(\beta) + h(\beta)$$

where

$$g(\beta) = \frac{1}{n}\|y - X\beta\|_2^2 \quad \text{and} \quad h(\beta) = \lambda\|\beta\|_1.$$

Iterative Shrinkage-Thresholding Algorithm (ISTA) is basically proximal gradient descent applied to LASSO. The first part g is smooth with smoothness parameter

$$\frac{1}{\eta} = \frac{2}{n}\|X\|_2.$$

We observed that

$$\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(x) = (\max\{0, |x_i| - \eta\lambda\} \cdot \mathrm{sign}(x_i))_{i\in[d]}.$$

Basically, if any component x_i is greater than $\eta\lambda$ or less than $-\eta\lambda$, we shrink $|x_i|$ to $\eta\lambda$ where

$$\eta\lambda = \frac{n\lambda}{2\|X\|_2}.$$

FISTA stands for Fast ISTA, that is an accelerated version of ISTA.

ISTA requires $O(1/\epsilon)$ iterations, while FISTA needs $O(1/\sqrt{\epsilon})$ iterations to converge to an ϵ-approximate solution.