Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- convergence of proximal gradient descent.
- ISTA and FISTA for LASSO.
- Proximal point algorithm.

2 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. For constrained minimixation, we take $h(x) = I_C(x)$ where C is the convex domain. Then the associated prox operator is equivalent to the projection operator. For LASSO, we take $h(\beta) = \lambda\|\beta\|_1$ whose associated prox operator is given by

$$\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(\beta) = \left(\underbrace{\max\{0, |\beta_i| - \eta\lambda\}}_{\text{shirinkage operator}} \cdot \mathrm{sign}(\beta_i)\right)_{i\in[d]}$$

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 β 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 16.1. *$G_\eta(\hat{x}) = 0$ if and only if $\hat{x} \in \text{argmin}_{x\in\mathbb{R}^d}\, g(x) + h(x)$.*

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

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

Note that $\hat{x} = \text{prox}_{\eta h}(\hat{x} - \eta \nabla g(\hat{x}))$ is equivalent to

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

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 16.2. *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

$$
\begin{aligned}
g\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) &\leq g(x) + \nabla g(x)^\top \left(\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) - x\right) + \frac{\beta}{2}\left\|\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) - x\right\|_2^2 \\
&= g(x) - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2 \\
&\leq g(z) - \nabla g(x)^\top (z - x) - \frac{\alpha}{2}\|z - x\|_2^2 - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2
\end{aligned}
\tag{16.1}
$$

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. Note that

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

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

$$h\left(x-\frac{1}{\beta}G_{1/\beta}(x)\right)\leq h(z)-\left(G_{1/\beta}(x)-\nabla g(x)\right)^\top\left(z-\left(x-\frac{1}{\beta}G_{1/\beta}(x)\right)\right). \tag{16.2}$$

Combining (16.1) and (16.2), 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 16.2 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

$$f\left(x-\frac{1}{\beta}G_{1/\beta}(x)\right)\leq f(x)-\frac{1}{2\beta}\|G_{1/\beta}(x)\|_2^2. \tag{16.3}$$

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 (16.3). Instead, we start from Lemma 16.2 by plugging in $z=x^*$ and $x=x_t$. Then

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

This implies that

$$f(x_{t+1})-f(x^*)\leq\frac{\beta}{2}\left(\|x_t-x^*\|_2^2-\|x_{t+1}-x^*\|_2^2\right)-\frac{\alpha}{2}\|x_t-x^*\|_2^2. \tag{16.4}$$

Theorem 16.3. *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*

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

Proof. First, sum up (16.4) for $t=1,\dots,T$ and then divide each side by T. Then we obtain

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

By (16.3), we know that $f(x_{T+1})\leq f(x_T)\leq\cdots\leq f(x_2)$. Moreover, $\|x_t-x^*\|_2\geq 0$. 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 16.4. *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*

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

Proof. Note that the left-hand side of (16.4) is greater than or equal to 0, and so is the right-hand side. Then it follows that

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

as required. □

3 ISTA and FISTA for LASSO

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

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

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.

4 Proximal point algorithm

Remember that the proximal gradient method works for the following composite minimization problem.

$$\text{minimize} \quad f(x) = g(x) + h(x).$$

The proximal gradient method proceeds with the update rule

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

In this section, we discuss the proximal point method, which is a special case of proximal gradient, and its application to the dual problem. Note that minimizing a closed convex function f can be written as a (trivial) composite minimization as follows.

$$\text{minimize} \quad f(x) = 0 + f(x).$$

Here, the first part is $g = 0$, which is trivially smooth, and the second part is $h = f$. Then the corresponding proximal gradient update is given by

$$x_{t+1} = \text{prox}_{\eta f}(x_t).$$

The algorithm with this update rule is referred to as the proximal point method. As $g = 0$ is smooth, the proximal point algorithm converges with a rate of $O(1/T)$.

Algorithm 2 Proximal point algorithm

Initialize x_1.
for $t = 1, \ldots, T$ **do**
  Update $x_{t+1} = \text{prox}_{\eta f}(x_t)$.
end for
Return x_{T+1}.

Theoretically, we can use any function h_t to run the proximal point algorithm, even if the objective is not h_t, in which case, the update rule corresponds to

$$x_{t+1} = \text{prox}_{\eta h_t}(x_t).$$

Hence, at each time step t, we may use a different function h_t hypothetically. Let us consider the first-order approximation of the objective function f at $x = x_t$.

$$h_t(x) = f(x_t) + \nabla f(x_t)^\top (x - x_t).$$

We know that $f(x) \geq h_t(x)$ for all x by convexity. Then what is the proximal point update with h_t? Note that

$$\begin{aligned}\text{prox}_{\eta h_t}(x_t) &= \operatorname*{argmin}_u \left\{ f(x_t) + \nabla f(x_t)^\top (u - x_t) + \frac{1}{2\eta}\|u - x_t\|_2^2 \right\} \\ &= x_t - \eta \nabla f(x_t).\end{aligned}$$

Therefore, the proximal point algorithm with the first-order approximation of f is precisely gradient descent. Hence, one can interpret gradient descent as an instance of the proximal point algorithm.

Let us now compare the proximal point algorithm with the objective f and gradient descent.

Lemma 16.5. $\text{prox}_{\eta f}(x) = (I + \eta \partial f)^{-1}(x)$.

Proof. Let $u = \text{prox}_{\eta f}(x)$. Remember that $u = \text{prox}_{\eta f}(x)$ if and only if $x - u \in \eta \partial f(u)$. Note that $x - u \in \eta \partial f(u)$ is equivalent to $x \in (I + \eta \partial f)(u)$, which is equivalent to $u \in (I + \eta \partial f)^{-1}(x)$. In summary,

$$u = \text{prox}_{\eta f}(x) \quad \leftrightarrow \quad u \in (I + \eta \partial f)^{-1}(x).$$

Since u is unique, it follows that $u = (I + \eta \partial f)^{-1}(x)$. □

By this lemma, the proximal point update rule can be written as

$$x_{t+1} = \text{prox}_{\eta f}(x_t) = (I + \eta \partial f)^{-1}(x_t).$$

This is equivalent to $x_t = (I + \eta \partial f)(x_{t+1}) = x_{t+1} + \eta \nabla f(x_{t+1})$, which is

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

In contrast to gradient descent that proceeds with $x_{t+1} = x_t - \eta \nabla f(x_t)$, we use the gradient at x_{t+1}.