1 Outline

In this lecture, we study

- proximal gradient descent,
- LASSO and FISTA,
- proximal point algorithm.

2 Proximal Gradient Descent

Let us 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. When h is given by the indicator function of a convex domain $\mathcal{X}$, the problem reduces to constrained smooth convex optimization. We also saw that the LASSO formulation can be viewed as an instance of the composite problem. In this section, we will learn and analyze the proximal gradient descent algorithm for solving the compositie optimization problem.

2.1 Proximal Operator

Recall that given a convex function $h : \mathbb{R}^d \to \mathbb{R}$ and $\eta > 0$, the proximal operator of h with parameter η is defined as

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

Let us prove the following useful lemma to understand the proximal operator.

Lemma 6.1. *$u = \text{prox}_{\eta h}(x)$ if and only if $x - u \in \eta\partial h(u)$.*

Proof. Note that $u = \text{prox}_{\eta h}(x)$ 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 $x - u \in \eta\partial h(u)$. □

We saw that the proximal operator for the indicator function of a convex set is the associated projector operator. We also consider the following two examples.

Example 6.2. Consider $h(x) = \|x\|_1$. Then

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

Let $u = \text{prox}_{\eta h}(x)$. Then, by Lemma 6.1,

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

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

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

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

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

For example,

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

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.

Example 6.3. Consider $h(x) = (1/2)x^\top Ax + b^\top x + c$ where A is positive semidefinite. Then

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

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

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

Therefore,

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

2.2 Analysis of Proximal Gradient Descent

The proximal gradient algorithm applies to this composite problem proceeds with the following update rule.

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

When f is smooth with parameter β, we set the step size $\eta = 1/\beta$ is in the smooth convex minimization setting. Proximal Gradient Descent applied to LASSO is referred to as Iterative

Algorithm 1 Proximal Gradient Descent

Initialize $x_1 \in \mathbb{R}^d$.
for $t = 1, \ldots, T$ **do**
  Update $x_{t+1} = \text{prox}_{\eta h}(x_t - (1/\beta)\nabla g(x_t))$ where β is the smoothness parameter of g.
end for
Return x_T.

Shrinkage-Thresholding Algorithm (ISTA). In this section, we will prove convergence of proximal gradient descent.

The gradient mapping is defined as

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

Here, $-\eta G_\eta(x)$ is equal to $\text{prox}_{\eta h}(x - \eta \nabla g(x)) - x$, 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 6.4. *$G_\eta(\hat{x}) = 0$ 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 6.5. *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

$$\begin{aligned} g\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) &\leq g(x) + \nabla g(x)^\top \left(\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) - x\right) + \frac{\beta}{2}\left\|\left(x - \frac{1}{\beta} G_{1/\beta}(x)\right) - x\right\|_2^2 \\ &= g(x) - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2 \\ &\leq g(z) - \nabla g(x)^\top (z - x) - \frac{\alpha}{2}\|z - x\|_2^2 - \frac{1}{\beta}\nabla g(x)^\top G_{1/\beta}(x) + \frac{1}{2\beta}\left\|G_{1/\beta}(x)\right\|_2^2 \end{aligned} \tag{6.1}$$

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

$$h\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq h(z) - \left(G_{1/\beta}(x) - \nabla g(x)\right)^\top \left(z - \left(x - \frac{1}{\beta}G_{1/\beta}(x)\right)\right). \tag{6.2}$$

Combining (6.1) and (6.2), 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 6.5 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

$$f\left(x - \frac{1}{\beta}G_{1/\beta}(x)\right) \leq f(x) - \frac{1}{2\beta}\|G_{1/\beta}(x)\|_2^2. \tag{6.3}$$

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 (6.3). Instead, we start from Lemma 6.5 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

$$f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) - \frac{\alpha}{2}\|x_t - x^*\|_2^2. \tag{6.4}$$

We can prove the following convergence result for Proximal Gradient Descent.

Theorem 6.6. *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 (6.4) for $t = 1, \ldots, T$ 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 (6.3), 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 $(\beta/2)\|x_1 - x^*\|_2^2$. □

Note that this recovers the convergence result for gradient descent for smooth functions by considering the case $h = 0$.

2.3 Proximal Gradient Descent with Momentum

In the last lecture, we explained to idea of momentum, which leads to a faster rate of convergence for gradient descent. In fact, the technique extends to proximal gradient descent.

Recall that the idea of momentum incorporates the direction $x_t - x_{t-1}$ that we took when moving from x_{t-1} to x_t to obtain the next iterate x_{t+1}. Let $\gamma_t > 0$ be a weight, and

$$y_t = x_t + \gamma_t(x_t - x_{t-1}).$$

Then we apply the primal gradient descent update on y_t to obtain the next point x_{t+1}, as follows.

$$x_{t+1} = \text{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right).$$

Algorithm 2 summarizes the accelerated version of proximal gradient descent.

Algorithm 2 Accelerated Proximal Gradient Descent

Initialize $x_1 \in \mathbb{R}^d$.
Set $x_0 = x_1$.
for $t = 1, \ldots, T$ **do**
  $y_t = x_t + \gamma_t(x_t - x_{t-1})$ for some $\gamma_t > 0$.
  $x_{t+1} = \text{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right)$.
end for
Return x_{T+1}.

To provide a convergence result of the accelerated proximal gradient descent method, we need the following lemma.

Lemma 6.7. *Let $u, v \in \mathbb{R}^d$. Then for all $z \in \mathbb{R}^d$,*

$$\frac{1}{\eta}(\text{prox}_{\eta h}(x) - x)^\top(z - \text{prox}_{\eta h}(x)) + h(z) \geq h(\text{prox}_{\eta h}(x)).$$

Proof. Note that

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

By the optimality condition, it follows that for some $g \in \partial h(\text{prox}_{\eta h}(x))$, for any $z \in \mathbb{R}^d$

$$\left(g + \frac{1}{\eta}\left(\text{prox}_{\eta h}(x) - x\right)\right)^\top (z - \text{prox}_{\eta h}(x)) \geq 0.$$

This implies that

$$\frac{1}{\eta}(\text{prox}_{\eta h}(x) - x)^\top(z - \text{prox}_{\eta h}(x)) + g^\top(z - \text{prox}_{\eta h}(x)) \geq 0.$$

Here, since h is convex, we have

$$h(z) \geq h(\text{prox}_{\eta h}(x)) + g^\top(z - \text{prox}_{\eta h}(x)).$$

Adding the two inequalities, we prove the desired bound of this lemma. □

Theorem 6.8. *Let $f = g + h$ where g is a β-smooth convex function in the ℓ_2 norm and h is convex. We set η and γ_t as*

$$\eta = \frac{1}{\beta}, \quad \gamma_t = \frac{t-2}{t+1}.$$

Then x_{T+1} returned by Accelerated Proximal Gradient Descent (Algorithm 2) satisfies

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

where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.

Proof. Note that Algorithm 2 is equivalent to

$$\begin{aligned} y_t &= (1-\lambda_t)x_t + \lambda_t v_t \\ x_{t+1} &= \mathrm{prox}_{h/\beta}\left(y_t - \frac{1}{\beta}\nabla g(y_t)\right) \\ v_{t+1} &= x_t + \frac{1}{\lambda_t}(x_{t+1} - x_t) \end{aligned}$$

where

$$\lambda_t = \frac{2}{t+1}.$$

This is because $y_t = x_t + \lambda_t(v_t - x_t)$ and

$$\lambda_t(v_t - x_t) = \lambda_t\left(\left(\frac{1}{\lambda_{t-1}} - 1\right)x_t + \left(1 - \frac{1}{\lambda_{t-1}}\right)x_{t-1}\right) = \frac{\lambda_t(1-\lambda_{t-1})}{\lambda_{t-1}}(x_t - x_{t-1}) = \gamma_t(x_t - x_{t-1}).$$

Moreover, we have $\lambda_1 = 1$, and for $t \geq 2$,

$$\frac{1-\lambda_t}{\lambda_t^2} \leq \frac{1}{\lambda_{t-1}^2}.$$

First, as g is β-smooth,

$$g(x_{t+1}) \leq g(y_t) + \nabla g(y_t)^\top(x_{t+1} - y_t) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2.$$

Next, Lemma 6.7 implies that for any $z \in \mathbb{R}^d$,

$$\begin{aligned} h(x_{t+1}) &\leq h(z) + \beta\left(x_{t+1} - y_t + \frac{1}{\beta}\nabla g(y_t)\right)^\top (z - x_{t+1}) \\ &= h(z) + \nabla g(y_t)^\top(z - x_{t+1}) + \beta\,(x_{t+1} - y_t)^\top (z - x_{t+1}). \end{aligned}$$

Adding these two inequalities, we deduce that

$$\begin{aligned} f(x_{t+1}) &\leq h(z) + g(y_t) + \nabla g(y_t)^\top(z - y_t) + \beta\,(x_{t+1} - y_t)^\top (z - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\ &\leq f(z) + \beta\,(x_{t+1} - y_t)^\top (z - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \end{aligned}$$

where the second inequality follows from convexity of g. By setting $z = x^*$ and $z = x_t$, we have

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq \beta\,(x_{t+1} - y_t)^\top (x^* - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\ f(x_{t+1}) - f(x_t) &\leq \beta\,(x_{t+1} - y_t)^\top (x_t - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2. \end{aligned}$$

Summing up the first inequality multiplied by λ_t and the second one multiplied by $(1-\lambda_t)$, we get

$$\begin{aligned}
&f(x_{t+1}) - f(x^*) - (1-\lambda_t)(f(x_t) - f(x^*)) \\
&\leq \beta\,(x_{t+1} - y_t)^\top\,(\lambda_t x^* + (1-\lambda_t)x_t - x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - y_t\|_2^2 \\
&= \frac{\beta}{2}(x_{t+1} - y_t)^\top(2\lambda_t x^* + 2(1-\lambda_t)x_t - x_{t+1} - y_t) \\
&= \frac{\beta}{2}\|y_t - (1-\lambda_t)x_t - \lambda_t x^*\|_2^2 - \frac{\beta}{2}\|x_{t+1} - (1-\lambda_t)x_t - \lambda_t x^*\|_2^2 \\
&= \frac{\beta\lambda_t^2}{2}\|v_t - x^*\|_2^2 - \frac{\beta\lambda_t^2}{2}\|v_{t+1} - x^*\|_2^2.
\end{aligned}$$

This implies that

$$\begin{aligned}
\frac{1}{\lambda_t^2}(f(x_{t+1}) - f(x^*)) + \frac{\beta}{2}\|v_{t+1} - x^*\|_2^2 &\leq \frac{1-\lambda_t}{\lambda_t^2}(f(x_t) - f(x^*)) + \frac{\beta}{2}\|v_t - x^*\|_2^2 \\
&\leq \frac{1}{\lambda_{t-1}^2}(f(x_t) - f(x^*)) + \frac{\beta}{2}\|v_t - x^*\|_2^2 \\
&\vdots \\
&\leq \frac{1}{\lambda_1^2}(f(x_2) - f(x^*)) + \frac{\beta}{2}\|v_2 - x^*\|_2^2 \\
&\leq \frac{1-\lambda_1}{\lambda_1^2}(f(x_1) - f(x^*)) + \frac{\beta}{2}\|v_1 - x^*\|_2^2 \\
&= \frac{\beta}{2}\|v_1 - x^*\|_2^2 \\
&= \frac{\beta}{2}\|x_1 - x^*\|_2^2.
\end{aligned}$$

Therefore, it follows that

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

as required. □

Hence, the convergence rate is $O(1/T^2)$, which matches the oracle lower bound. The number of required iterations to bound the error by ϵ is $O(1/\sqrt{\epsilon})$.

3 LASSO and FISTA

In this section, we consider linear regression. We assume that the relationship between the vector $x \in \mathbb{R}^d$ of features and the response variable $y \in \mathbb{R}$ is modeled using a linear equation given by

$$y = \theta_{\text{true}}^\top x + \epsilon$$

where:

- $\theta_{\text{true}} \in \mathbb{R}^d$ is the coefficient vector,
- $\epsilon \in \mathbb{R}$ is the noise term representing unexplained variation.

We infer the true coefficient vector θ_{true} using the method of least squares, which minimizes the average of squared differences between the observed and predicted values of y. Namely, given a set of n data $(x_1, y_1), \ldots, (x_n, y_n)$, we solve

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta^\top x_i\right)^2 \quad = \quad \min_{\theta} \quad \frac{1}{n}\|Y - X\theta\|_2^2. \tag{6.5}$$

Here, Y denotes the vector whose components are $y_1, \ldots, y_n$, and X denotes the matrix whose rows are $x_1^\top, \ldots, x_n^\top$. Note that

$$f(\theta) := \frac{1}{n}\|Y - X\theta\|_2^2 = \frac{1}{n}\theta^\top X^\top X\theta - \frac{2}{n}Y^\top X\theta + \frac{1}{n}Y^\top Y.$$

Since $X^\top X$ is positive semidefinite, it follows that the MSE loss $f(\theta)$ is convex.

Recall the formulation of LASSO, given by

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

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

As discussed in the last lecture, we know that the first part is smooth, and the other part is something whose subdifferential is well understood. Hence, we may apply proximal gradient descent with $h(\theta) = \lambda\|\theta\|_1$ whose associated prox operator is given by

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

The proximal gradient algorithm applies to this composite problem proceeds with the following update rule.

$$\theta_{t+1} = \text{prox}_{\eta h}(\theta_t - \eta\nabla g(\theta_t)).$$

Proximal Gradient Descent applied to LASSO is referred to as Iterative Shrinkage-Thresholding Algorithm (ISTA).

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 3 Proximal point algorithm

Initialize x_1.
for $t = 1, \dots, T$ **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

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

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

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 equvialent 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}.