1 Outline

In this lecture, we study

- Proximal point algorithm,
- KKT conditions,
- Lagrangian duality.

2 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 1 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

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

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

Proof. Let $u = \mathrm{prox}_{\eta f}(x)$. Remember that $u = \mathrm{prox}_{\eta f}(x)$ 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,

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

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

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

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}.

3 Lagrangian Duality

We consider problems of the following structure.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\ & h_j(x) = 0 \quad \text{for } j = 1, \ldots, \ell. \end{aligned} \tag{16.1}$$

We consider the most general setting for which we do not impose the condition that the objective and constraint functions are convex. We may define vector-valued functions $g : \mathbb{R}^d \to \mathbb{R}^m$ and $h : \mathbb{R}^d \to \mathbb{R}^\ell$ such that

- $g(x) = (g_1(x), \ldots, g_m(x))^\top$,
- $h(x) = (h_1(x), \ldots, h_\ell(x))^\top$.

Then (16.1) can be written as

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0. \end{aligned} \tag{16.2}$$

3.1 Lagrangian Dual Problem

The *Lagrangian function* of (16.1) is given by

$$\begin{aligned}\mathcal{L}(x,\lambda,\mu) &= f(x) + \lambda^\top g(x) + \mu^\top h(x) \\ &= f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{\ell} \mu_j h_j(x).\end{aligned}$$

When the objective function f is convex, constraint functions $g_1, \ldots, g_m$ are convex, constraint functions $h_1, \ldots, h_\ell$ are affine, and the multiplier $\lambda \geq 0$, the Lagrangian function is convex in x for any fixed λ and μ. Moreover, the Lagrangian function is affine in λ and μ for any fixed x.

The *Lagrangian dual function* of (16.1) is

$$q(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) = \inf_x \left\{ f(x) + \lambda^\top g(x) + \mu^\top h(x) \right\}.$$

Notice that the Lagrangian dual function is concave in (λ,μ), regardless of f, $g_1, \ldots, g_m$, and $h_1, \ldots, h_\ell$. This is because $\mathcal{L}(x,\lambda,\mu)$ is affine in λ and μ for any fixed x, and $q(\lambda,\mu)$ is a point-wise minimum of affine functions.

Proposition 16.2. *Let x be a feasible solution to (16.1), and $\lambda \geq 0$. Then*

$$f(x) \geq q(\lambda,\mu).$$

Proof. Since x is feasible, $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_j(x) = 0$ for $j = 1, \ldots, \ell$. Then for any $\lambda \geq 0$, we have

$$\sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{\ell} \mu_j h_j(x) \leq 0.$$

This implies that

$$f(x) \geq \mathcal{L}(x,\lambda,\mu).$$

Note that

$$q(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) \leq \mathcal{L}(x,\lambda,\mu).$$

Therefore, $f(x) \geq q(\lambda,\mu)$. □

By Proposition 16.2, if (16.1) is unbounded below, the Lagrangian dual function $q(\lambda,\mu) = -\infty$ for any $\lambda \geq 0$.

With the Lagrangian dual function, we can provide a lower bound on the problem (16.1). The *Lagrangian dual problem* is defined as

$$\begin{aligned} \text{maximize} \quad & q(\lambda,\mu) \\ \text{subject to} \quad & \lambda \geq 0. \end{aligned} \tag{16.3}$$

We often call (16.1) as *primal* and (16.3) as the *associated (Lagrangian) dual*. The following result states that the optimal value of the primal is lower bounded by the optimal value of the dual.

Theorem 16.3 (Weak duality)**.** *Consider the problem (16.1) and the associated Lagrangian dual problem (16.3). Then the following statement holds.*

$$\min_{x \in C} f(x) \geq \max_{\lambda \geq 0} q(\lambda,\mu)$$

where $C = \{x : g_i(x) \leq 0$ for $i = 1, \ldots, m,\ h_j(x) = 0$ for $j = 1, \ldots, \ell\}$.

Proof. By proposition 16.2, we know that $f(x) \geq q(\lambda, \mu)$ for any $x \in C$ and $\lambda \geq 0$. Then taking the minimum of $f(x)$ over $x \in C$, it follows that $\min_{x \in C} f(x) \geq q(\lambda, \mu)$. Then taking the maximum of $q(\lambda, \mu)$ over $\lambda \geq 0$, we obtain the desired inequality. □

Theorem 16.3 holds regardless of whether the objective and constraint functions are convex or not. Then our next question is whether the equality holds. To answer this, we define the notion of *Slater's condition*.

Definition 16.4 (Slater's condition)**.** Suppose that $g_1, \ldots, g_k$ are affine and $g_{k+1}, \ldots, g_m$ are convex functions that are not affine. Then we say that the problem (16.1) satisfies Slater's condition if there exists a solution $\bar{x}$ such that

$$g_i(\bar{x}) \leq 0 \text{ for } i = 1, \ldots, k, \quad g_i(\bar{x}) < 0 \text{ for } i = k+1, \ldots, m, \quad h_j(\bar{x}) = 0 \text{ for } j = 1, \ldots, \ell.$$

If we assume that the objective f is convex and the constraint functions satisfy Slater's condition, then the inequality given in Theorem 16.3 holds with equality.

Theorem 16.5 (Strong duality)**.** *Consider the primal problem* (16.1) *and the associated Lagrangian dual problem* (16.3)*. Assume that the objective function f and the constraint functions $g_1, \ldots, g_m$ are convex, and $h_1, \ldots, h_\ell$ are affine. If the primal problem* (16.1) *has a finite optimal value and Slater's condition, given in Definition 16.7, is satisfied, then there exist $\lambda^* \geq 0$ and μ^* such that*

$$\min_{x \in C} f(x) = q(\lambda^*, \mu^*) = \max_{\lambda \geq 0} q(\lambda, \mu)$$

where $C = \{x : g_i(x) \leq 0$ for $i = 1, \ldots, m,\ h_j(x) = 0$ for $j = 1, \ldots, \ell\}$.

3.2 Examples

Consider the following linear program in standard form.

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0. \end{aligned} \tag{16.4}$$

Then the Lagrangian dual function is given by

$$\begin{aligned} q(\lambda, \mu) &= \inf_x \mathcal{L}(x, \lambda, \mu) \\ &= \inf_x \left\{ c^\top x - \lambda^\top x + \mu^\top (Ax - b) \right\} \\ &= -b^\top \mu + \inf_x \left\{ (c - \lambda + A^\top \mu)^\top x \right\}. \end{aligned}$$

Note that $\inf_x \left\{ (c - \lambda + A^\top \mu)^\top x \right\} = -\infty$ unless $c - \lambda + A^\top \mu = 0$. Hence, to maximize the Lagrangian dual function $q(\lambda, \mu)$, it is sufficient to consider (λ, μ) satisfying $c - \lambda + A^\top \mu = 0$. Therefore, the associated Lagrangian dual problem is equivalent to

$$\begin{aligned} \text{maximize} \quad & -b^\top \mu \\ \text{subject to} \quad & c - \lambda + A^\top \mu = 0, \\ & \lambda \geq 0. \end{aligned} \tag{16.5}$$

In fact, we can eliminate the variable from the constraints $c + A^\top \mu \geq \lambda$ and $\lambda \geq 0$, and they can be equivalently written as $c + A^\top \mu \geq 0$. Moreover, the variables μ are unrestricted, so we can replace μ by $-\mu$. Then (16.5) is equivalent to

$$\begin{aligned} \text{maximize} \quad & b^\top \mu \\ \text{subject to} \quad & A^\top \mu \leq c, \end{aligned} \tag{16.6}$$

which is the dual linear program for (16.4).

Next we consider the following quadratic program.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^\top Q x + p^\top x \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{16.7}$$

where Q is positive definite and thus is invertible. The corresponding Lagrangian function is given by

$$\begin{aligned} \mathcal{L}(x, \mu) &= \frac{1}{2} x^\top Q x + p^\top x + \mu^\top (Ax - b) \\ &= -b^\top \mu + \left(\frac{1}{2} x^\top Q x + (p + A^\top \mu)^\top x \right). \end{aligned}$$

Then

$$\nabla_x \mathcal{L}(x, \mu) = Qx + (p + A^\top \mu),$$

and therefore, $\nabla_x \mathcal{L}(x, \mu) = 0$ if and only if $x = -Q^{-1}(p + A^\top \mu)$. This in turn implies that the Lagrangian dual function is given by

$$\begin{aligned} q(\mu) &= \inf_x \mathcal{L}(x, \mu) \\ &= \mathcal{L}\left(-Q^{-1}(p + A^\top \mu), \mu \right) \\ &= -b^\top \mu - \frac{1}{2} (p + A^\top \mu)^\top Q^{-1} (p + A^\top \mu) \\ &= -\frac{1}{2} \mu^\top A Q^{-1} A^\top \mu - (b + A Q^{-1} p)^\top \mu - \frac{1}{2} p^\top p. \end{aligned}$$

Hence, the Lagrangian dual problem is

$$\max_\mu \left\{ -\frac{1}{2} \mu^\top A Q^{-1} A^\top \mu - (b + A Q^{-1} p)^\top \mu \right\}.$$

3.3 Lagrangian dual for conic programming

Consider the following conic programming problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq_{K_i} 0 \quad \text{for } i = 1, \ldots, m \\ & h_j(x) = 0 \quad \text{for } j = 1, \ldots, \ell \end{aligned} \tag{16.8}$$

where $K_1, \ldots, K_m$ are proper cones. Remember that $g_i(x) \leq_{K_i} 0$ means $-g_i(x) \in K_i$. Moreover, recall that the dual cone of a cone K is given by

$$K^* = \{ y : \ y^\top x \geq 0 \ \forall x \in K \}.$$

As we picked a nonnegative multiplier $\lambda \geq 0$ to define the Lagrangian function, we pick a multiplier λ from the dual cone K^*. The Lagrangian function of (16.8) is given by

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i^\top g_i(x) + \sum_{j=1}^{\ell} \mu_j h_j(x)$$

where $\lambda_i \in K_i^*$ is now a vector from the dual cone of K_i for each i. Then the Lagrangian dual function is similarly defined as $q(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu)$. The Lagrangian dual problem is given by

$$\begin{array}{ll} \text{maximize} & q(\lambda, \mu) \\ \text{subject to} & \lambda_i \geq_{K_i^*} 0 \quad \text{for } i = 1, \ldots, m. \end{array} \tag{16.9}$$

As an example, we consider the following semidefinite program.

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \displaystyle\sum_{i=1}^{d} x_i A_i \geq_{S_+^m} B \end{array} \tag{16.10}$$

where S_+^m denotes the PSD cone containing all $m \times m$ PSD matrices. We learned that the PSD cone is self-dual, so the dual of S_+^m is itself. Let $Y \in S_+^m$, and consider the associated Lagrangian dual function.

$$q(Y) = \inf_x \mathcal{L}(x, Y) = \inf_x \left\{ c^\top x - \sum_{i=1}^{d} x_i \mathrm{tr}(Y^\top A_i) + \mathrm{tr}(Y^\top B) \right\}.$$

Note that the Lagrangian dual function $q(Y)$ has a finite value if and only if $c_i = \mathrm{tr}(Y^\top A_i)$ for every $i \in [d]$. Then the Lagrangian dual problem is given by

$$\begin{array}{ll} \text{maximize} & \mathrm{tr}(Y^\top B) \\ \text{subject to} & \mathrm{tr}(Y^\top A_i) = c_i \quad \text{for } i = 1, \ldots, d. \\ & Y \in S_+^m \end{array} \tag{16.11}$$

4 Karush-Kuhn-Tucker conditions

Remember that x^* is an optimal solution to

$$\min_{x \in C} \quad f(x)$$

where C is a convex set and f is differentiable if and only if

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall x \in C.$$

However, the structure of C may be arbitrary, which makes the condition difficult to verify. In this section, we present another way of verifying optimality. Namely, Karu-Kuhn-Tucker conditions, often referred to as KKT conditions.

4.1 Linear constraints

We consider problems of the following structure.

$$\begin{array}{rl} \text{minimize} & f(x) \\ \text{subject to} & Ax \leq b \\ & Cx = d \end{array} \tag{16.12}$$

where

- $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$,
- $C \in \mathbb{R}^{\ell \times d}$ and $d \in \mathbb{R}^\ell$.

Theorem 16.6 (KKT conditions for linearly constrained problems)**.** *The linearly constrained problem as in* (16.12) *satisfies the following.*

1. *(Necessity) If x^* is a feasible solution to* (16.12) *and $f(x^*)$ is a local minimum, then there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ such that*

$$\nabla f(x^*)^\top + \lambda^{*\top} A + \mu^{*\top} C = 0 \quad \& \quad \lambda^{*\top}(Ax - b) = 0. \tag{$\star$}$$

2. *(Sufficiency) If f is convex, x^* is a feasible solution to* (16.12)*, and there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ satisfying* ($\star$)*, then x^* is an optimal solution to* (16.12)*.*

4.2 General convex constraints

We consider problems of the following structure.

$$\begin{array}{rll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & \text{for } i = 1, \ldots, m \\ & h_j(x) = 0 & \text{for } j = 1, \ldots, \ell \end{array} \tag{16.13}$$

where

- f is convex,
- $g_1, \ldots, g_m$ are convex,
- $h_1, \ldots, h_\ell$ are affine.

Definition 16.7 (Slater's condition)**.** Suppose that $g_1, \ldots, g_k$ are affine and $g_{k+1}, \ldots, g_m$ are convex functions that are not affine. Then we say that the problem (16.13) satisfies Slater's condition if there exists a solution $\bar{x}$ such that

$$g_i(\bar{x}) \leq 0 \text{ for } i = 1, \ldots, k, \quad g_i(\bar{x}) < 0 \text{ for } i = k+1, \ldots, m, \quad h_j(\bar{x}) = 0 \text{ for } j = 1, \ldots, \ell.$$

Theorem 16.8 (KKT conditions for convex constrained problems)**.** *The convex programming problem as in* (16.13) *satisfies the following.*

1. *(Necessity) Assume that Slater's condition is satisfied. If x^* is a feasible optimal solution to (16.13), then there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ such that*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{\ell} \mu_j^* \nabla h_j(x^*) = 0 \quad \& \quad \lambda_i^* g_i(x^*) = 0 \text{ for all } i = 1, \ldots, m. \quad (\star\star)$$

2. *(Sufficiency) If x^* is a feasible solution to (16.13) and there exist $\lambda^* \in \mathbb{R}^m_+$ and $\mu^* \in \mathbb{R}^\ell$ satisfying ($\star\star$), then x^* is an optimal solution to (16.13).*