Convex Duality

1 Introduction

In this lecture, we consider the concept of duality in convex analysis, which is important for understanding various optimization algorithms. We will then discuss the duality for composite convex optimization problem, and present a primal-dual algorithm which is closely related to dual averaging.

2 Conjugate Function

Given a function $f(x)$, its conjugate function (or dual) is defined as

$$f^*(x) = \sup_y [y^\top x - f(y)].$$

Since $f^*(x)$ is the sup over a family of linear (thus convex) functions of x indexed by y, we know that $f^*(x)$ is convex (even if $f(x)$ is not convex). It is also a closed convex function (that is, its epi-graph is a closed set).

By definition, we always have the following inequality, which is referred to as the Fenchel's inequality:

$$f^*(x) + f(y) \geq y^\top x. \tag{1}$$

We also have the following theorem.

Theorem 1 *If $f(x)$ is a closed convex function, then*

$$f^{**}(x) = f(x).$$

Moreover, for any pair (x, y), the following conditions are equivalent:

- $x \in \partial f(y)$
- $y \in \partial f^*(x)$
- (x, y) *satisfies the equality*

$$f^*(x) + f(y) = y^\top x.$$

Proof Given y. Let $x \in \partial f(y)$, then we know that y achieves the optimal of

$$\sup_y [y^\top x - f(y)].$$

Thus

$$f^*(x) = y^\top x - f(y).$$

This means that for any x', we have

$$f^*(x') - f^*(x) \geq (y^\top x' - f(y)) - (y^\top x - f(y)) = y^\top (x' - x).$$

Therefore $y \in \partial f^*(x)$ by definition. This implies that x achieves the optimal of

$$\sup_x [y^\top x - f^*(x)],$$

and thus

$$f^{**}(y) = y^\top x - f^*(x) = y^\top x - [y^\top x - f(y)] = f(y).$$

Other directions of equivalence relations can be similarly obtained. ■

Note that if we impose a constraint on $f(x)$, restricted to $x \in C$, then we may replace it by the unconstrained convex function $f_C(x) = f(x) + I_C(x)$, where I_C is the indicator function of C (which takes value 0 in C and value $+\infty$ out of C). The function $f_C(x)$ has domain C, and its subdifferential of at $x \in C$ is $\partial f(x) + \mathcal{N}_C(x)$, where $\mathcal{N}_C(x) = \{g : g^\top (y - x) \leq 0 \; \forall y \in C\}$. If C is defined by the equality constraints $\{x : Ax - b = 0\}$, where $A \in \mathbb{R}^{m \times d}$, then

$$\mathcal{N}_C(x) = \{A^\top \lambda : \lambda \in \mathbb{R}^m\}.$$

Therefore the subdifferential of $f_C(x)$ at x is $\{\partial f(x) + A^\top \lambda : \lambda \in \mathbb{R}^m\}$. This is equivalent to the Lagrangian multiplier method. Inequality constraints can be handled similarly.

Given a norm $\|\cdot\|$, we define its dual norm as

$$\|y\|_* = \sup_{\|x\| \leq} y^\top x.$$

This leads to the following inequality:

$$\|x\| \|y\|_* \geq y^\top x$$

for all x and y.

Given a norm $\|\cdot\|$, we may define the convex function

$$f(x) = \frac{1}{2} \|x\|^2.$$

Then its convex conjugate is

$$f^*(y) = \sup_x [y^\top x - \frac{1}{2} \|x\|^2] = \frac{1}{2} \|y\|_*^2.$$

This is because if x achieves the optimality, with $\|x\| = r$, then we must have $y^\top x = \|y\|_* \|x\|$ over $\|x\| = r$, and the optimal r is $r = \|y\|_*$.

3 Examples

Example 1 *Let* $f(x) = -\ln x$*, defined on* R^+*, then*

$$f^*(y) = \sup_x [xy + \ln x],$$

The first order condition is

$$y = \nabla f(x) = -1/x.$$

Therefore $x = -1/y$*, and*

$$f^*(y) = -1 - \ln(-y),$$

defined on R^-*.*

Example 2 *Let* $f(x) = \frac{1}{2} x^\top Ax$*, where* A *is a positive definite matrix. Then*

$$f^*(y) = \sup_x \left[x^\top y - \frac{1}{2} x^\top Ax \right].$$

The optimal solution is at

$$y = \nabla f(x) = Ax,$$

which gives $x = A^{-1}y$*. This implies that*

$$f^*(y) = (A^{-1}y)^\top y - \frac{1}{2}(A^{-1}y)^\top A(A^{-1}y) = \frac{1}{2} y^\top A^{-1} y.$$

Example 3 *Let*

$$f(x) = \frac{1}{p}\|x\|_p^p,$$

then

$$f^*(y) = \sup_x \left[x^\top y - \frac{1}{p}\|x\|_p^p \right].$$

The optimal is achieved at $y = \nabla f(x)$*, which is*

$$y_j = [|x_j|^{p-1} \mathrm{sign}(x_j)].$$

That is,

$$x_j = [|y_j|^{q-1} \mathrm{sign}(y_j)],$$

where $1/p + 1/q = 1$*. This gives*

$$f^*(y) = \frac{1}{q}\|y\|_q^q.$$

4 Smoothness and Strong Convexity

We may generalize the smoothness and strong convex with respect to the $\|\cdot\|_2$ norm to an arbitrary norm as follows.

Definition 1 *A function is L-smooth with respect to a norm $\|\cdot\|$ if*

$$f(x') \leq f(x) + \nabla f(x)^\top (x' - x) + \frac{L}{2}\|x' - x\|^2$$

for all x and x'.

A function is λ-strongly convex with respect to a norm $\|\cdot\|$ if

$$f(x') \geq f(x) + \nabla f(x)^\top (x' - x) + \frac{\lambda}{2}\|x' - x\|^2$$

for all x and x'.

We have the following result concerning the properties of conjugate function.

Theorem 2 *Consider a norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. If $f(x)$ is L-smooth with respect to $\|\cdot\|$, then $f^*(y)$ is L^{-1} strongly convex with respect to $\|\cdot\|_*$.*

Similarly, if $f(x)$ is λ-strongly convex with respect to $\|\cdot\|$, then $f^*(y)$ is λ^{-1} smooth with respect to $\|\cdot\|_*$.

Proof We prove the first statement. Consider y and y'. Let $x \in \partial f^*(y)$. This implies that $y \in \partial f(x)$, and thus for all x':

$$f(x') \leq f(x) + y^\top (x' - x) + \frac{L}{2}\|x' - x\|^2 = -f^*(y) + y^\top x' + \frac{L}{2}\|x' - x\|^2. \tag{2}$$

Therefore

$$\begin{aligned}
f^*(y') =& \sup_{x'} [(y')^\top x' - f(x')] \\
\geq& \sup_{x'} \left[(y')^\top x' + f^*(y) - y^\top x' - \frac{L}{2}\|x' - x\|^2\right] \\
=& f^*(y) + x^\top (y' - y) + \sup_{x'} \left[(y' - y)^\top (x' - x) - \frac{L}{2}\|x' - x\|^2\right] \\
=& f^*(y) + x^\top (y' - y) + \frac{1}{2L}\|y' - y\|_*^2.
\end{aligned}$$

The inequality follows from (2). The last equality is due to the fact that the conjugate of $0.5\|\cdot\|^2$ is $0.5\|\cdot\|_*^2$. This proves the first statement. The second statement is similar. ■

5 Bregman Divergence

Given convex function f, and let

$$D_f(x', x) = f(x') - f(x) - y^\top (x' - x)$$

be its Bregman divergence, where $y \in \partial f(x)$. Note that y may not be uniquely defined, and we may pick any subgradient y to define a Bregman divergence that depends on y. We have the following result.

Theorem 3 *Let $y \in \partial f(x)$ and $y' \in \partial f(x')$. Then*

$$\begin{aligned} D_f(x', x) =& f(x') - f(x) - y^\top (x' - x) \\ =& f^*(y) - f^*(y') - (x')^\top (y - y') = D_{f^*}(y, y'). \end{aligned}$$

Proof We have

$$f(x') = (x')^\top y' - f^*(y'), \qquad f(x) = (x)^\top y - f^*(y).$$

Therefore

$$\begin{aligned} f(x') - f(x) - y^\top (x' - x) =& [(x')^\top y' - f^*(y')] - [(x)^\top y - f^*(y)] - y^\top (x' - x) \\ =& f^*(y) - f^*(y') - (x')^\top (y - y'). \end{aligned}$$

This proves the desired result. ∎

6 Moreau's Identity

Given $g(x)$, we denote the proximal mapping by

$$\mathrm{prox}_g(x) = \arg\min_z \left[g(z) + \frac{1}{2} \|z - x\|_2^2 \right].$$

Then we have the following result

Theorem 4

$$\mathrm{prox}_g(x) + \mathrm{prox}_{g^*}(x) = x.$$

Proof Let $z = \mathrm{prox}_g(x)$, then there exists $y \in \partial g(z)$ such that

$$y + z - x = 0.$$

This implies that $z \in \partial g^*(y)$, and thus

$$y = \arg\min_{y'} \left[g^*(y') + \frac{1}{2} \|y' - x\|_2^2 \right].$$

That is $y = \mathrm{prox}_{g^*}(x)$. This proves the result. ∎

7 Fenchel's duality

Consider the composite optimization problem

$$\phi(x) = f(x) + g(x),$$

and let x_* be its solution.

We may rewrite the composite optimization problem as:

$$\phi(x) = f(x) + g(x') \qquad x = x'.$$

It follows that we may write the Lagrangian as

$$f(x) + g(x') + \alpha^\top (x - x').$$

Given any α

$$f(x_*) + g(x_*) \geq \min_{x,x'} [f(x) + g(x') + \alpha^\top (x - x')] = -f^*(-\alpha) - g^*(\alpha).$$

The problem

$$\phi_D(\alpha) = -f^*(-\alpha) - g^*(\alpha)$$

is called the dual problem, and $\phi(x)$ is called the primal problem.

Theorem 5 *Given $\alpha_* \in \arg\max_\alpha \phi_D(\alpha)$, there exists $x_* \in \arg\min_x \phi(x)$ such that*

$$x_* \in \partial g^*(\alpha_*), \qquad x_* \in \partial f^*(-\alpha_*),$$

and

$$\phi(x_*) = \phi_D(\alpha_*).$$

Proof We know that α_* satisfies the first order condition

$$0 \in \partial[f^*(-\alpha_*) + g^*(\alpha_*)].$$

Therefore there exists $x_* \in \partial g^*(\alpha_*)$ such that $x_* \in \partial f^*(-\alpha_*)$. It follows from the property of convex conjugate function that

$$\min_{x,x'} [f(x) + g(x') + \alpha_*^\top (x - x')]$$

is achieved at $x = x' = x_*$. Therefore

$$\phi(x_*) = \min_{x,x'} [f(x) + g(x') + \alpha_*^\top (x - x')] = \phi_D(\alpha_*).$$

■

One can design primal dual method based on the dual formulation, as in Algorithm 1. The method tries to find the optimal solution of the dual problem, and is closely related to dual averaging.

Algorithm 1: Primal Dual Ascent Method

Input: $f(\cdot)$, $g(\cdot)$, x_0, $\eta_0, \eta_1, \eta_2, \ldots$ and α_0

Output: x_T

for $t = 1, 2, \ldots, T$ **do**

  Let $\alpha_t = (1 - \eta_{t-1})\alpha_{t-1} - \eta_{t-1}\nabla f(x_{t-1})$

  Let $x_t = \arg\min_x \left[-\alpha_t^\top x + g(x)\right] = \nabla g^*(\alpha_t)$

Return: x_T

Theorem 6 *Consider Algorithm 1, and assume that $f(x)$ is an L-smooth convex function, and $g(x)$ is λ-strongly convex. Let $w_* = \arg\min_w \phi(w)$ and $\alpha_* = \arg\max_\alpha \phi_D(\alpha)$. If we take $\eta \leq \lambda/(\lambda + L)$, then*

$$\phi_D(\alpha_*) - \phi_D(\alpha_t) \leq (1 - \eta)^t [\phi_D(\alpha_*) - \phi_D(\alpha_0)],$$

and

$$\phi(w_{t-1}) \leq \phi(w_*) + \frac{(1 - \eta)^t}{\eta} [\phi_D(\alpha_*) - \phi_D(\alpha_0)].$$

Proof Since $f^*(\cdot)$ is $1/L$-strongly convex, we have:

$$\begin{aligned} -f^*(-\alpha_t) \geq & - (1 - \eta) f^*(-\alpha_{t-1}) - \eta f^*(\nabla f(x_{t-1})) + \frac{(1 - \eta)\eta}{2L} \|\alpha_{t-1} + \nabla f(x_{t-1})\|_2^2 \\ = & - (1 - \eta) f^*(-\alpha_{t-1}) - \eta x_{t-1}^\top \nabla f(x_{t-1}) + \eta f(x_{t-1}) + \frac{(1 - \eta)\eta}{2L} \|\alpha_{t-1} + \nabla f(x_{t-1})\|_2^2. \end{aligned}$$

Since $g^*(\cdot)$ is $1/\lambda$ smooth, and $x_{t-1} = \nabla g^*(\alpha_{t-1})$, we have

$$\begin{aligned} -g^*(\alpha_t) \geq & - g^*(\alpha_{t-1}) - \eta \nabla g^*(\alpha_{t-1})^\top (-\nabla f(x_{t-1}) - \alpha_{t-1}) - \frac{\eta^2}{2\lambda} \|\alpha_{t-1} + \nabla f(x_{t-1})\|_2^2 \\ \geq & - (1 - \eta) g^*(\alpha_{t-1}) + \eta g(x_{t-1}) + \eta x_{t-1}^\top \nabla f(x_{t-1}) - \frac{\eta^2}{2\lambda} \|\alpha_{t-1} + \nabla f(x_{t-1})\|_2^2. \end{aligned}$$

By adding the two terms, we have

$$\phi_D(\alpha_t) \geq (1 - \eta)\phi_D(\alpha_{t-1}) + \eta \phi(w_{t-1}).$$

Since $\phi_D(\alpha_*) \leq \phi(w_{t-1})$, this implies that

$$\phi_D(\alpha_*) - \phi_D(\alpha_t) \leq (1 - \eta)[\phi_D(\alpha_*) - \phi_D(\alpha_{t-1})] \leq \cdots (1 - \eta)^t [\phi_D(\alpha_*) - \phi_D(\alpha_0)].$$

Moreover, from $\phi_D(\alpha_t) \leq \phi(w_*) = \phi_D(\alpha_*)$, we have

$$\eta[\phi(w_{t-1}) - \phi(w_*)] \leq (1 - \eta)[\phi_D(\alpha_*) - \phi_D(\alpha_{t-1})] \leq (1 - \eta)^t [\phi_D(\alpha_*) - \phi_D(\alpha_0)].$$

This leads to the desired bound. ■