Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 14: Convex Duality

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.

Duality Theorem

Theorem

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

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

and thus

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

Other directions of equivalence relations can be similarly obtained.

Norm

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

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

Then its convex conjugate is

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

Example

Example

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

Example

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

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

Example

Let

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

then

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

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

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

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

for all x and x'.
A function is λ-strongly convex with respect to a norm $\|\cdot\|$ if

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

for all x and x'.

Theorem

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

Theorem

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':

$$f(x') \leq f(x) + y^\top (x' - x) + \frac{L}{2}\|x' - x\|^2 = -f^*(y) + y^\top x' + \frac{L}{2}\|x' - x\|^2. \quad (1)$$

Therefore

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

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 $y \in \partial f$. We have the following result.

Theorem

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

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

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

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

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.

Strong Duality

Theorem

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

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

Dual Algorithm

One can design primal dual method based on the dual formulation, as in Algorithm 1. The method 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

Example: $L_1 - L_2$ Regularization

Solving

$$f(x) + \frac{\lambda}{2}\|x\|_2^2 + \mu\|x\|_1 .$$

Proximal Mapping:

$$\begin{aligned}\mathrm{prox}_\eta(x) =& \arg\min_z \left[\frac{1}{2\eta}\|z - x\|_2^2 + \frac{\lambda}{2}\|z\|_2^2 + \mu\|z\|_1\right] \\ =&(1+\lambda\eta)^{-1}[\mathrm{sign}(x_j)(|x_j| - \mu\eta)_+]_{j=1,\ldots,d}\end{aligned}$$

Proximal Gradient

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

RDA

- $\alpha_t = \alpha_{t-1} - \eta\nabla f(x_{t-1})$
- $x_t = \mathrm{prox}_{\eta t}(\alpha_t)$

Primal Dual Ascent: $\eta' = \eta\lambda$

- $\alpha_t = (1-\eta')\alpha_{t-1} - \eta'\nabla f(x_{t-1})$
- $x_t = (\lambda)^{-1}[\mathrm{sign}([\alpha_t]_j)(|[\alpha_t]_j| - \mu)_+]_{j=1,\ldots,d}$

Convergence Theorem

Theorem

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 $\eta \le \lambda/(\lambda + L)$, then

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

and

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

Summary

Convex Duality

- $x \in \nabla f(y)$ equivalent to $y \in \nabla f^*(x)$.
- Duality between smooth and strong convexity
- Duality and Bregman divergence
- Duality and Proximal mapping

Composite Optimization

- Primal formulation: $\phi(x)$
- Dual formulation: $\phi_D(\alpha)$
- Dual algorithm