Introductory Lectures on Optimization

Beyond The Black-box Model (2)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

December 16, 2024

Outline

1. Duality Principle
   - Duality Principle
2. Lagrangian Duality and Algorithms
   - Lagrangian Duality
   - KKT
   - Algorithms using Lagrangian duality
3. Fenchel conjugate and algorithms
   - Fenchel conjugate
   - Properties
   - Fenchel duality
4. Reference

Part I
Duality Principle and Algorithms

Duality Principle

In mathematical optimization theory, duality or the duality principle is the principle that optimization problems may be viewed from either of two perspectives, the primal problem or the dual problem.

1. **Solve primal problem by solving dual problem:** The solution to the dual problem provides a lower bound to the solution of the primal (minimization) problem.

2. **Duality Gap:** However in general the optimal values of the primal and dual problems need not be equal. Their difference is called the duality gap.

Example: Duality in Linear Programs

Given $\boldsymbol{c} \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $G \in \mathbb{R}^{r\times n}$, $\boldsymbol{h} \in \mathbb{R}^r$:

Primal LP:

$$\begin{aligned} &\min_{\boldsymbol{x}} c^\top \boldsymbol{x} \\ \text{s.t.}\quad &A\boldsymbol{x} = \boldsymbol{b}, \\ &G\boldsymbol{x} \le \boldsymbol{h}. \end{aligned}$$

Dual LP:

$$\begin{aligned} &\max_{\boldsymbol{u},\boldsymbol{v}} -\boldsymbol{b}^\top \boldsymbol{u} - \boldsymbol{h}^\top \boldsymbol{v} \\ \text{s.t.}\quad &-A^\top \boldsymbol{u} - G^\top \boldsymbol{v} = \boldsymbol{c}, \\ &\boldsymbol{v} \ge 0. \end{aligned}$$

Explanation: for any $\boldsymbol{u}$ and $\boldsymbol{v} \ge 0$, and $\boldsymbol{x}$ primal feasible,

$$\begin{aligned} &\boldsymbol{u}^\top(\boldsymbol{b} - A\boldsymbol{x}) + \boldsymbol{v}^\top(\boldsymbol{h} - G\boldsymbol{x}) \ge 0, \\ \Leftrightarrow\ &(-A^\top \boldsymbol{u} - G^\top \boldsymbol{v})^\top \boldsymbol{x} \ge -\boldsymbol{b}^\top \boldsymbol{u} - \boldsymbol{h}^\top \boldsymbol{v}. \end{aligned}$$

So if $\boldsymbol{c} = -A^\top \boldsymbol{u} - G^\top \boldsymbol{v}$, we get a bound on primal optimal value.

Example: Duality in Linear Programs

We can also establish an alternative explanation: for any u and $v \geq 0$, and x primal feasible

$$\boldsymbol{c}^\top \boldsymbol{x} \geq \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{u}^\top (A\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{v}^\top (G\boldsymbol{x} - \boldsymbol{h}) \equiv L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}).$$

So if C denotes primal feasible set, f^* primal optimal value, then for any $\boldsymbol{u}$ and $\boldsymbol{v} \geq 0$,

$$f^* \geq \min_{\boldsymbol{x} \in C} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \geq \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \equiv g(\boldsymbol{u}, \boldsymbol{v})$$

In other words, $g(\boldsymbol{u}, \boldsymbol{v})$ is a lower bound on f^* for any $\boldsymbol{u}$ and $\boldsymbol{v} \geq 0$. Note that

$$g(\boldsymbol{u}, \boldsymbol{v}) = \begin{cases} -\boldsymbol{b}^\top \boldsymbol{u} - \boldsymbol{h}^\top \boldsymbol{v}, & \text{if } \boldsymbol{c} = -A^\top \boldsymbol{u} - G^\top \boldsymbol{v}, \\ -\infty, & \text{otherwise.} \end{cases}$$

This second explanation reproduces the same dual, but is actually completely general and applies to arbitrary optimization problems (even nonconvex ones).

Duality Problem

For minimization problem with inequality constraints,

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x}), \ \text{s.t. } g_j(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m,$$

We have, for example,

Lagrangian dual problem:

$$\max_{\boldsymbol{u}} \ \inf_{\boldsymbol{x}}(f(\boldsymbol{x}) + \sum_{j=1}^{m} \boldsymbol{u}_j g_j(\boldsymbol{x}))$$

$$\text{s.t.} \quad \boldsymbol{u}_i \geq 0, \ i = 1, \ldots, m,$$

and Wolfe dual problem:

$$\max_{\boldsymbol{u},\boldsymbol{x}} \ (f(\boldsymbol{x}) + \sum_{j=1}^{m} \boldsymbol{u}_j g_j(\boldsymbol{x}))$$

$$\text{s.t.} \quad \nabla f(\boldsymbol{x}) + \sum_{j=1}^{m} \boldsymbol{u}_j \nabla g_j(\boldsymbol{x}) = 0$$

$$\boldsymbol{u}_i \geq 0, \ i = 1, \ldots, m.$$

Remark: There is also the Fenchel Duality problem.

Duality Problem

Given an optimization problem, we do not obtain a dual problem until we specify how to perturb the optimization problem. This is why equivalent formulations of an optimization problem can lead to different dual problems. By reformulating it we have in fact specified a different way to perturb it.

As is typical in math, the ideas become clear when we work at an appropriate level of generality. Assume that our optimization problem is

$$\min_{\boldsymbol{x}} \left\{ \phi(\boldsymbol{x}, 0) \equiv f(\boldsymbol{x}) \right\},$$

where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex. Standard convex optimization problems can be written in this form with an appropriate choice of ϕ.

Duality Problem

The perturbed problems are

$$\min_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{y}).$$

Let $h(\boldsymbol{y}) = \inf_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{y})$. The primal optimization problem is simply to evaluate $h(0)$.

From our knowledge of conjugate functions, we know that

$$h(0) \geq h^{**}(0)$$

and that typically we have equality. For example, if h is subdifferentiable at 0 (which is typical for a convex function) then $h(0) = h^{**}(0)$. The dual problme is simply to evaluate $h^{**}(0)$.

Lagrangian

Consider general minimization problem

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{s.t.} \quad &h_i(\boldsymbol{x}) \le 0, \\ &\ell_j(\boldsymbol{x}) = 0. \end{aligned}$$

f need not be convex, but of course we will pay special attention to convex case. We define the Lagrangian as

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{u}_i h_i(\boldsymbol{x}) + \sum_{j=1}^{r} \boldsymbol{v}_j \ell_j(\boldsymbol{x}).$$

New variables $\boldsymbol{u} \in \mathbb{R}^m$, $\boldsymbol{v} \in \mathbb{R}^r$, with $\boldsymbol{u} \ge 0$ (implicitly, we define $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = -\infty$ for $\boldsymbol{u} < 0$)

Lagrangian

Important Property: for any $\boldsymbol{u} \geq 0$ and v,

$$f(\boldsymbol{x}) \geq L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \text{ at each feasible } \boldsymbol{x}.$$

Why? For feasible x,

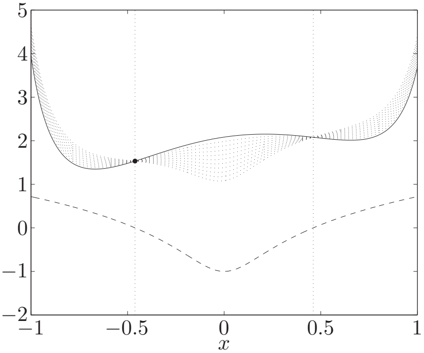
$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{u}_i \underbrace{h_i(\boldsymbol{x})}_{\leq 0} + \sum_{j=1}^{r} \boldsymbol{v}_j \underbrace{\ell_j(\boldsymbol{x})}_{=0}.$$

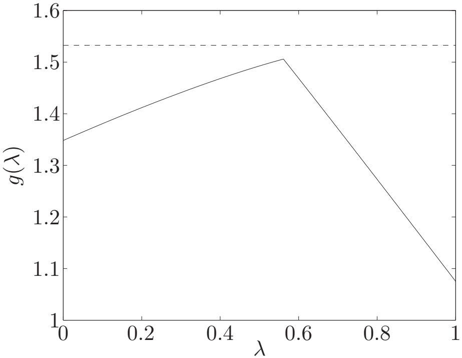
Let C denote primal feasible set, f^* denote primal optimal value. Minimizing $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ over all $\boldsymbol{x}$ gives a lower bound:

$$f^* \geq \min_{\boldsymbol{x} \in C} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \geq \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \equiv g(\boldsymbol{u}, \boldsymbol{v})$$

We call $g(\boldsymbol{u}, \boldsymbol{v})$ the Lagrange dual function.

Lagrangian


5
4
3
2
1
0
−1
−2
−1
−0.5
0
0.5
1
x


1.6
1.5
1.4
1.3
1.2
1.1
1
$g(\lambda)$
0
0.2
0.4
0.6
0.8
1
λ

Lagrange dual problem

Our constructed dual function $g(\boldsymbol{u}, \boldsymbol{v})$ satisfies $f^* \geq g(\boldsymbol{u}, \boldsymbol{v})$ for all $\boldsymbol{u} \geq 0$ and $\boldsymbol{v}$. Hence best lower bound is given by maximizing $g(\boldsymbol{u}, \boldsymbol{v})$ over all dual feasible $\boldsymbol{u}, \boldsymbol{v}$, yielding Lagrange dual problem:

$$\begin{aligned} &\max_{\boldsymbol{u}, \boldsymbol{v}} g(\boldsymbol{u}, \boldsymbol{v}) \\ \text{s.t.} \quad &\boldsymbol{u} \geq 0. \end{aligned}$$

Key property, called weak duality: if dual optimal value is g^*, then

$$f^* \geq g^*.$$

Note that this always holds (even if primal problem is nonconvex).

Lagrange dual problem

Another key property: the dual problem is a convex optimization problem (as written, it is a concave maximization problem).

Again, this is always true (even when primal problem is not convex). By definition:

$$g(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x}} \{ f(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{u}_i h_i(\boldsymbol{x}) + \sum_{j=1}^{r} \boldsymbol{v}_j \ell_j(\boldsymbol{x}) \} \tag{5}$$

$$= \underbrace{-\max_{\boldsymbol{x}} \{ -f(\boldsymbol{x}) - \sum_{i=1}^{m} \boldsymbol{u}_i h_i(\boldsymbol{x}) - \sum_{j=1}^{r} \boldsymbol{v}_j \ell_j(\boldsymbol{x}) \}}_{\text{pointwise maximum of convex functions in } (\boldsymbol{u},\boldsymbol{v})} \tag{6}$$

i.e., g is concave in $(\boldsymbol{u}, \boldsymbol{v})$, and $\boldsymbol{u} \geq 0$ is a convex constraint, hence dual problem is a concave maximization problem.

Lagrange dual problem

Recall that we always have $f^* \geq g^*$ (weak duality). On the other hand, in some problems we have observed that actually

$$f^* = g^*$$

which is called strong duality.

Slater's condition: if the primal is convex problem (i.e., f and $h_1, \ldots, h_m$ are convex, $\ell_1, \ldots, \ell_m$ are affine), and there exists at least one strictly feasible $x \in \mathbb{R}^n$, meaning

$$h_1(\boldsymbol{x}) < 0, \ldots, h_m(\boldsymbol{x}) < 0, \quad \text{and } \ell_1(0) = 0, \ldots, \ell_r(\boldsymbol{x}) = 0$$

then strong duality holds.

KKT Theorem

Theorem 16 (Kuhn-Tucker)

Let f_i be differentiable convex functions, $i = 0 \dots m$. Suppose that there exists a point $\bar{\boldsymbol{x}}$ such that $f_i(\bar{\boldsymbol{x}}) < 0$ for all $i = 1 \dots m$. (Slater condition.) A point $\boldsymbol{x}^*$ is a solution to the problem

$$\min\{f_0(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) \le 0, i = 1 \dots m\} \tag{7}$$

if and only if it is feasible and there exist non-negative number $\lambda_i, i = 1 \dots m$, such that

$$\nabla f_0(\boldsymbol{x}^*) + \sum_{i \in I^*} \lambda_i \nabla f_i(\boldsymbol{x}^*) = 0,$$

where $I^* = \{i \in [1, m] : f_i(\boldsymbol{x}^*) = 0\}$.

KKT Theorem

Consider the related parametric max-type function of (7)

$$f(t;\boldsymbol{x}) = \max\{f_0(\boldsymbol{x}) - t; f_i(\boldsymbol{x}), i = 1, \ldots, m\}, t \in \mathbb{R}^1, \boldsymbol{x} \in Q.$$

Let us introduce the function $f^*(t) = \min_{\boldsymbol{x} \in Q} f(t;\boldsymbol{x})$.

Lemma 17 (Lemma 2.3.4 of Nesterov [2013])

let t^* be an optimal value of problem (7). Then

$$f^*(t) \leq 0 \text{ for all } t \geq t^*,$$

$$f^*(t) > 0 \text{ for all } t < t^*.$$

Thus, the smallest root of function $f^*(t)$ corresponds to the optimal value of the problem (7).

KKT Theorem

Proof.

In view of Lemma 2.3.4, $\boldsymbol{x}^*$ is a solution to (7) if and ony if it is a global minimizer of the function

$$\phi(\boldsymbol{x}) = \max\{f_0(\boldsymbol{x}) - f^*; f_i(\boldsymbol{x}), i = 1 \ldots m\}$$

In view of Theorem 3.1.15, this is the case if and only if $0 \in \partial\phi(\boldsymbol{x}^*)$. Further, in view of Lemma 3.1.10, this is true if and only if there exist nonnegative $\bar{\lambda}_i$, such that

$$\bar{\lambda}_0 \nabla f_0(\boldsymbol{x}^*) + \sum_{i \in I^*} \bar{\lambda}_i \nabla f_i(\boldsymbol{x}^*) = 0, \quad \bar{\lambda}_0 + \sum_{i \in I^*} \bar{\lambda}_i = 1.$$

Remark.

$$\partial f(\boldsymbol{x}) = \mathrm{Conv}\{\partial f_i(\boldsymbol{x}) | i \in I(\boldsymbol{x})\},$$

where $I(\boldsymbol{x}) = \{i : f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}$.

KKT Theorem

Proof. (Continued.)

Thus, we need to prove only that $\bar{\lambda}_0 > 0$. Indeed, if $\bar{\lambda}_0 = 0$, then for any $\bar{\boldsymbol{x}}$, since

$$\sum_{i \in I^*} \bar{\lambda}_i \nabla f_i(\boldsymbol{x}^*) = 0 \text{ and } f_i(\boldsymbol{x}^*) = \phi(\boldsymbol{x}^*) = 0, i \in I^*,$$

we have

$$\sum_{i \in I^*} \bar{\lambda}_i f_i(\bar{\boldsymbol{x}}) \geq \sum_{i \in I^*} \bar{\lambda}_i [f_i(\boldsymbol{x}^*) + \langle \nabla f_i(\boldsymbol{x}^*),\ \bar{\boldsymbol{x}} - \boldsymbol{x}^* \rangle] = 0.$$

This contradicts the Slater condition (exists a point $\bar{\boldsymbol{x}}$ such that $f_i(\bar{\boldsymbol{x}}) < 0$ for all $i = 1 \ldots m$). Therefore $\bar{\lambda}_0 > 0$ and we can take

$$\lambda_i = \bar{\lambda}_i / \bar{\lambda}_0, i \in I^*.$$

KKT Theorem

Lemma 18 (Lemma 3.1.12)

Let $A \succ 0$. Then

$$\max_{\boldsymbol{x}} \{ \langle \boldsymbol{c}, \ \boldsymbol{x} \rangle : \langle A\boldsymbol{x}, \ \boldsymbol{x} \rangle \le 1 \} = \langle A^{-1}\boldsymbol{c}, \ \boldsymbol{c} \rangle^{1/2}.$$

Proof. Note that all conditions of Theorem 16 are satisfied and the solution $\boldsymbol{x}^*$ of the above problem is attained at the boundary of the feasible set. Therefore, in acordance with Theorem 16 we have to solve the following equations:

$$\boldsymbol{c} = \lambda A\boldsymbol{x}^*, \quad \langle A\boldsymbol{x}^*, \ \boldsymbol{x}^* \rangle = 1.$$

Thus, $\boldsymbol{x}^* = \frac{1}{\lambda} A^{-1}\boldsymbol{c}$, and we also have $\langle \lambda A\boldsymbol{x}^*, \ \boldsymbol{x}^* \rangle = \lambda$, it implies $\lambda = \langle A^{-1}\boldsymbol{c}, \ \boldsymbol{c} \rangle^{1/2}$. □

Algorithms: Dual gradient ascent

Consider the following optimization problem:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \text{s.t. } A\boldsymbol{x} = b.$$

Convavity of the dual function $g(\lambda)$ ensures existence of subgradients, so the subgradient method can be applied to optimize g(λ). The dual gradient(subgradient) algorithm is as follows:

start from initial λ_0. For all $t \geq 0$:

$$\boldsymbol{x}_k = \arg \inf_{\boldsymbol{x} \in C} L(\boldsymbol{x}, \lambda_k)$$
$$\lambda_{k+1} = \lambda_k + \eta(A\boldsymbol{x}_k - b)$$

This yields the $O(1/\sqrt{t})$ convergence rate obtained by the subgradient method.

Algorithms: Dual gradient ascent

Remark. Let $l(\boldsymbol{x}) = A\boldsymbol{x} - b$.

$$\begin{aligned} g(\lambda) &= \left\{ \min_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) \triangleq f(\boldsymbol{x}) + \lambda l(\boldsymbol{x}) \right\} \\ &\leq f(\boldsymbol{x}_k) + \lambda^\top l(\boldsymbol{x}_k) \\ &= f(\boldsymbol{x}_k) + \lambda_k^\top l(\boldsymbol{x}_k) + (\lambda - \lambda_k)^\top l(\boldsymbol{x}_k) \\ &= g(\lambda_k) + (\lambda - \lambda_k)^\top l(\boldsymbol{x}_k) \end{aligned}$$

Last step follows since $\boldsymbol{x}_k$ is a mimimizer of $L(\boldsymbol{x}, \lambda_k)$.
Thus, we know $l(\lambda_k)$ is a subgradient of $g(\lambda)$.
So we first compute $\boldsymbol{x}_k$ from the λ_k, then do the gradient (subgradient accent) to obtain λ_{k+1}.

Algorithms: Augmented Lagrangian method

Whereas dual gradient ascent updates λ_{k+1} by taking a step in a (sub)gradient direction, a method known as the dual proximal method can be motivated by starting with using the proximal operator as an update rule for iteratively optimizing λ:

$$\lambda_{k+1} = \mathrm{prox}_{\eta\ g}(\lambda_k) = \arg\sup_{\lambda} \underbrace{\underbrace{\inf_{\boldsymbol{x}\in C} f(\boldsymbol{x}) + \lambda^\top(A\boldsymbol{x} - \boldsymbol{b})}_{g(\lambda)} - \underbrace{\frac{1}{2\eta_k}\left\|\lambda - \lambda_k\right\|^2}_{\text{proximal term}}}_{h(\lambda)}.$$

Notice that this expression includes a proximal term which make $h(\lambda)$ strongly convex.

Algorithms: Augmented Lagrangian method

However, this update rule is not always directly useful since it requires optimizing $h(\lambda)$ over λ, which may not be available in closed form. Instead, notice that if we can interchange inf and sup (e.g. strong duality, Sion's theorem applied when C is compact) then we can rewrite

$$\begin{aligned} &\sup_{\lambda} \left\{ \inf_{\boldsymbol{x}\in C} f(\boldsymbol{x}) + \lambda^\top (A\boldsymbol{x} - \boldsymbol{b}) - \frac{1}{2\eta_k} \|\lambda - \lambda_k\|^2 \right\} \\ &= \inf_{\boldsymbol{x}\in C} \sup_{\lambda} \left\{ f(\boldsymbol{x}) + \lambda^\top (A\boldsymbol{x} - \boldsymbol{b}) - \frac{1}{2\eta_k} \|\lambda - \lambda_k\|^2 \right\} \\ &= \inf_{\boldsymbol{x}\in C} \left\{ f(\boldsymbol{x}) + \lambda_k^\top (A\boldsymbol{x} - \boldsymbol{b}) + \frac{\eta_k}{2} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 \right\}, \end{aligned}$$

where the inner sup is optimized in closed-form by $\lambda = \lambda_k + \eta_k(A\boldsymbol{x} + \boldsymbol{b})$.

Algorithms: Augmented Lagrangian method

To isolate the remaining optimization over $\boldsymbol{x}$, we make the following definition.

Definition 19 (Augmented Lagrangian)

The augmented Lagrangian is

$$L_{\eta_k}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda^\top (A\boldsymbol{x} - \boldsymbol{b}) + \frac{\eta_k}{2} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 .$$

The augmented Lagrangian method (aka Method of Multipliers) is defined by teh following iteratons:

$$\boldsymbol{x}_k = \arg \inf_{\boldsymbol{x} \in C} L_{\eta_k}(\boldsymbol{x}, \lambda_k)$$

$$\lambda_{k+1} = \lambda_k + \eta_k (A\boldsymbol{x}_k - \boldsymbol{b})$$

Algorithms: Dual decomposition

A major advantage of dual decomposition that it can lead to update rules which are trivially parallelizable. Suppose we can partition the primal problem into N blocks of size $(n_i)_{i=1}^N$.

$$X^\top = ((\boldsymbol{x}^{[1]})^\top, \ldots, (\boldsymbol{x}^{[N]})^\top), \ \boldsymbol{x}^{(i)} \in \mathbb{R}^{n_i} \text{ and } \sum_{i=1}^N n_i = n.$$

$$A = [A_1, \cdots, A_N], \ A \in \mathbb{R}^{m \times n}, A_i \in \mathbb{R}^{m \times n_i}.$$

$$A\boldsymbol{x} = \sum_{i=1}^N A_i \boldsymbol{x}^{[i]}. \quad f(\boldsymbol{x}) = \sum_{i=1}^N f_i(\boldsymbol{x}^{[i]}).$$

Then the Lagrangian is also separable in $\boldsymbol{x}$:

$$L(\boldsymbol{x}, \lambda) = \sum_{i=1}^N \left\{ L_i(\boldsymbol{x}^{[i]}, \lambda) \equiv f_i(\boldsymbol{x}^{[i]}) + \lambda^\top A_i \boldsymbol{x}^{[i]} - \frac{1}{N}\lambda^\top \boldsymbol{b} \right\}$$

Fenchel conjugate

The Fenchel conjugate of a function is a generalization of the Legendre transformation. It is also known as Legendre-Fenchel transformation or Fenchel transformation (after Adrien-Marie Legendre and Werner Fenchel).

It is used to transform an optimization problem into its corresponding dual problem, which can often be simpler to solve. The Fenchel conjugate is defined as follows.

Definition 20 (Fenchel conjugate)

The Fenchel conjugate of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$f^*(\boldsymbol{p}) = \sup_{\boldsymbol{x}} \langle \boldsymbol{p},\ \boldsymbol{x} \rangle - f(\boldsymbol{x}).$$

Important properties

1. convex. f^* is convex. Indeed f^* is the supremum of affine function and therefore convex. Thus, the Fenchel conjugate of f is also known as its convex conjugate.
2. Inverse. $(f^*)^*(\boldsymbol{x}) = f$ if f is convex (and lower semi-continuous and proper). In other words, the Fenchel conjugate is its own inverse for convex function.
3. subdifferential The subdifferential of f^* at $\boldsymbol{p}$ is $\partial f^*(\boldsymbol{p}) = \{\boldsymbol{x} : \boldsymbol{p} \in \partial f(\boldsymbol{x})\}$.
4. If f is u-strongly convex then f^* is continuously differentiable and $\frac{1}{u}$-Lipschitz smooth.

Example for Fenchel conjugate

Example[ERM]. In empirical risk minimization, we often want to minimize a function of the following form:

$$p(\boldsymbol{w}) = \sum_{i=1}^{m} \phi_i(\langle \boldsymbol{w},\ \boldsymbol{x}_i \rangle) + R(\boldsymbol{w}).$$

We can think of $\boldsymbol{w} \in \mathbb{R}^n$ as the model parameter that we want to optimize over (in this case it corresponds to picking a hyperplane), and $\boldsymbol{x}_i$ as the features of the i-th example in the training set. $\phi_i(\cdot, \boldsymbol{x}_i)$ is the loss function for the i-th training example and may depend on its label. $R(\boldsymbol{w})$ is the regularizer, and we typically choose it to be of the form $R(\boldsymbol{w}) = \frac{\lambda}{2} \|\boldsymbol{w}\|^2$. The primal problem, $\min_{\boldsymbol{w} \in \mathbb{R}^n} P(\boldsymbol{w})$, can be equivalently written as follows:

$$\min_{\boldsymbol{w},\boldsymbol{z}} \sum_{i=1}^{m} \phi_i(\boldsymbol{z}_i) + R(\boldsymbol{w}),\ \text{subject to } X^\top \boldsymbol{w} = \boldsymbol{z}.$$

Example for Fenchel conjugate

Example[ERM]. (continued) By Lagrangian duality, we know that the dual problem is the following:

$$\max_{\alpha\in\mathbb{R}^m} \min_{\boldsymbol{z},\boldsymbol{w}} \sum_{i=1}^{m} \phi_i(\boldsymbol{z}_i) + R(\boldsymbol{w}) - \alpha^\top(X^\top \boldsymbol{w} - \boldsymbol{z}) \tag{8}$$

$$= \max_{\alpha\in\mathbb{R}^m} \min_{\boldsymbol{z},\boldsymbol{w}} \sum_{i=1}^{m} \phi_i(\boldsymbol{z}_i) + \alpha_i \boldsymbol{z}_i + R(\boldsymbol{w}) - \alpha^\top X^\top \boldsymbol{w} \tag{9}$$

$$= \max_{\alpha\in\mathbb{R}^m} \left\{ -\max_{\boldsymbol{z},\boldsymbol{w}} \left\{ -\left\{ \sum_{i=1}^{m} \phi_i(\boldsymbol{z}_i) + \alpha_i \boldsymbol{z}_i \right\} + (X\alpha^\top)^\top \boldsymbol{w} - R(\boldsymbol{w}) \right\} \right\} \tag{10}$$

$$= \max_{\alpha\in\mathbb{R}^m} - \left\{ \sum_{i=1}^{m} \max_{\boldsymbol{z}_i}(-\phi_i(\boldsymbol{z}_i) - \alpha_i \boldsymbol{z}_i) + \max_{\boldsymbol{w}}(X\alpha)^\top \boldsymbol{w} - R(\boldsymbol{w}) \right\} \tag{11}$$

$$= \max_{\alpha\in\mathbb{R}^m} - \sum \phi_i^*(-\alpha) - R^*(X\alpha) \tag{12}$$

Fenchel duality

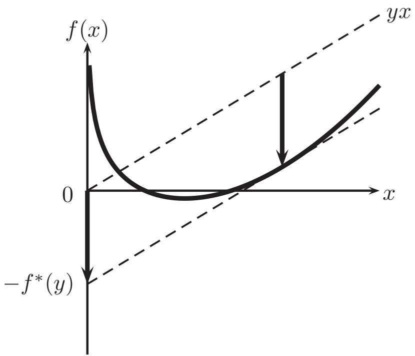
Theorem 21 (Fenchel duality)

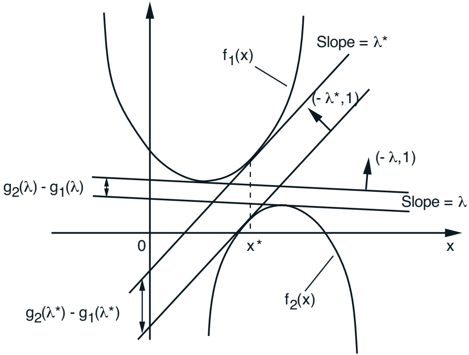
Suppose we have f proper convex, and g proper concave. Then

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) - g(\boldsymbol{x}) = \max_{p} g^*(\boldsymbol{p}) - f^*(\boldsymbol{p}).$$

In the one-dimensional case, we can illustrate Fenchel duality with the following figure.

Fenchel duality





Fenchel duality

In the minimization problem, we want to find $\boldsymbol{x}$ such that the vertical distance between f and g at $\boldsymbol{x}$ is as small as possible.

In the (dual) maximization problem, we draw tangents to the graphs of f and g such that the tangnt lines have the same slope $\boldsymbol{p}$, and we want to find $\boldsymbol{p}$ such that the vertical distance between the tangent lines is as large as possible.

The duality theorem above states that strong duality holds, that is, the two problems have the same solution.

Example using Fenchel duality

Example. Consider optimization problem as follows.

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{x}_0 \right\|^2, \text{ subject to } \left\| \boldsymbol{x} \right\|_\infty \leq 1.$$

It can be rewritten as

$$\min_{\boldsymbol{x}} \left\{ \{ f(\boldsymbol{x}) \equiv \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{x}_0 \right\|^2 \} - \{ g(\boldsymbol{x}) \equiv -I_C(\boldsymbol{x}) \} \right\},$$

where $C = \{ \boldsymbol{x} : \left\| \boldsymbol{x} \right\|_\infty \leq 1 \}$. In view of Theorem 21, we have the Fenchel duality problem as follows.

$$\max_{\boldsymbol{p}} g^*(\boldsymbol{p}) - f^*(\boldsymbol{p}) = \max_{\boldsymbol{p}} - \left\| \boldsymbol{p} \right\|_1 - \left(\frac{1}{2} \left\| \boldsymbol{p} \right\|^2 + \langle \boldsymbol{p}, \ \boldsymbol{x}_0 \rangle \right),$$

which is much easier to solve than the primal problem.

References I

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Martin Jaggi. Convex optimization without projection steps. *arXiv preprint arXiv:1108.1170*, 2011.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

References II

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Xinhua Zhang. Bregman divergence and mirror descent. URL `https://www2.cs.uic.edu/~zhangx/teaching/bregman.pdf`.

Thank You!

Email:qianhui@zju.edu.cn