对偶理论(续2)

王尧

西安交通大学智能决策与机器学习中心
(Email: yao.s.wang@gmail.com)

2022. 5

回顾：Slater条件

strong duality holds for a convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

if it is strictly feasible, *i.e.*,

$$\exists x \in \mathbf{int}\, \mathcal{D}: \qquad f_i(x) < 0, \quad i = 1, \ldots, m, \qquad Ax = b$$

- also guarantees that the dual optimum is attained (if $p^\star > -\infty$)

Refinement: actually only need strict inequalities for non-affine f_i

通常称为weak slater条件，更多介绍参见Boyd版5.2.3节与5.3.2节

回顾：不等式约束的线性规划

primal problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

dual function

$$g(\lambda) = \inf_x \left((c + A^T\lambda)^T x - b^T\lambda\right) = \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

dual problem

$$\begin{array}{ll} \text{maximize} & -b^T\lambda \\ \text{subject to} & A^T\lambda + c = 0, \quad \lambda \succeq 0 \end{array}$$

- from Slater's condition: $p^\star = d^\star$ if $A\tilde{x} \prec b$ for some $\tilde{x}$
- in fact, $p^\star = d^\star$ except when primal and dual are infeasible (由weak slater条件可知)

回顾：KKT条件

Given general problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0,\ i = 1, \dots, m \\ & \ell_j(x) = 0,\ j = 1, \dots, r \end{aligned}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial_x \left(f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x) \right)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \le 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \ge 0$ for all i (dual feasibility)

注意：上述条件没有任何凸假设

回顾：最优解与KKT条件的关系

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$x^\star$ and $u^\star, v^\star$ are primal and dual solutions

$\iff$ $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions

注意：上述等价性关系是针对原问题是凸问题的情形。若原问题非凸，满足KKT条件的解未必是最优解，此时stationarity条件不能保证下述等式成立：

$$g(u^\star, v^\star) = f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star)$$

一些注记

- 若能直接求解出凸优化问题的KKT对，则其就是对应问题的最优解.
- Slater条件的意义在于当问题最优解存在时，其相应KKT条件也会得到满足.
- 当Slater条件不满足时，即使原始问题存在全局极小值点，也可能不存在(x^*, λ^*) 满足KKT条件.

基追踪(Basis pursuit)问题

$$\begin{array}{ll}\min\limits_{x\in\mathbb{R}^n} & \|x\|_1,\\ \text{s.t.} & Ax=b.\end{array}\iff\begin{array}{ll}\min & \sum\limits_i x_i^+ + x_i^-,\\ \text{s.t.} & Ax^+ - Ax^- = b,\\ & x^+, x^- \geq 0.\end{array}\iff\begin{array}{ll}\min\limits_{y\in\mathbb{R}^{2n}} & \mathbf{1}^{\mathrm{T}} y,\\ \text{s.t.} & [A,-A]y = b,\\ & y\geq 0.\end{array}$$

- 拉格朗日函数: $L(x,\nu) = \|x\|_1 + \nu^{\mathrm{T}}(Ax-b)$.
- x^* 为全局最优解当且仅当存在$\nu^* \in \mathbb{R}^m$ 使得

$$\begin{cases} 0 \in \partial\|x^*\|_1 + A^{\mathrm{T}}\nu^*, \\ b = Ax^*. \end{cases}$$

仿射空间的投影问题

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \frac{1}{2}\|x-y\|_2^2, \\ \text{s.t.} \quad & Ax=b, \end{aligned}$$

其中$A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$ 以及$y\in\mathbb{R}^n$ 为给定的矩阵和向量且A 满秩.

- 拉格朗日函数: $L(x,\lambda)=\frac{1}{2}\|x-y\|^2+\lambda^{\mathrm{T}}(Ax-b)$.
- Slater条件成立, x^* 为一个全局最优解当且仅当存在$\lambda^*\in\mathbb{R}^m$ 使得

$$\begin{cases} x^*-y+A^{\mathrm{T}}\lambda^*=0, \\ Ax^*=b. \end{cases}$$

- 由上述KKT 条件第一式, 等号左右两边同时左乘A 可得

$$Ax^*-Ay+AA^{\mathrm{T}}\lambda=0\Rightarrow\lambda^*=(AA^{\mathrm{T}})^{-1}(Ay-b).$$

- 将λ^* 代回KKT 条件第一式可知

$$x^*=y-A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}(Ay-b).$$

Uses of Duality

Two key uses of duality:

- For x primal feasible and u, v dual feasible,

$$f(x) - f(x^\star) \leq f(x) - g(u, v)$$

  can be used as a stopping criterion in algorithms

- Under strong duality, given dual optimal $u^\star, v^\star$, any primal solution $x^\star$ solves

$$\min_x \; L(x, u^\star, v^\star)$$

  (i.e., satisfies the stationarity condition). This can be used to characterize or compute primal solutions from dual solution

原问题的对偶解法

An important consequence of stationarity: under strong duality, given a dual solution $u^\star, v^\star$, any primal solution $x^\star$ solves

$$\min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x)$$

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit characterization of primal solutions from dual solutions

Furthermore, suppose the solution of this problem is unique; then it must be the primal solution $x^\star$

This can be very helpful when the dual is easier to solve than the primal

原问题的对偶解法

For example, consider:

$$\min_x \sum_{i=1}^n f_i(x_i) \quad \text{subject to} \quad a^T x = b$$

where each $f_i(x_i) = \frac{1}{2} c_i x_i^2$ (smooth and strictly convex). Dual function:

$$\begin{aligned} g(v) &= \min_x \sum_{i=1}^n f_i(x_i) + v(b - a^T x) \\ &= bv + \sum_{i=1}^n \min_{x_i} \left\{ f_i(x_i) - a_i v x_i \right\} \\ &= bv - \sum_{i=1}^n f_i^*(a_i v), \end{aligned}$$

where each $f_i^*(y) = \frac{1}{2c_i} y^2$, called the conjugate of f_i

Therefore the dual problem is

$$\max_v \; bv - \sum_{i=1}^n f_i^*(a_i v) \quad \Longleftrightarrow \quad \min_v \sum_{i=1}^n f_i^*(a_i v) - bv$$

This is a convex minimization problem with scalar variable—much easier to solve than primal

Given $v^\star$, the primal solution $x^\star$ solves

$$\min_x \sum_{i=1}^n \left(f_i(x_i) - a_i v^\star x_i \right)$$

Strict convexity of each f_i implies that this has a unique solution, namely $x^\star$, which we compute by solving $f_i'(x_i) = a_i v^\star$ for each i. This gives $x_i^\star = a_i v^\star / c_i$

对偶范数的对偶

Let $\|x\|$ be a norm, e.g.,

- ℓ_p norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, for $p \geq 1$
- Trace norm: $\|X\|_{\text{tr}} = \sum_{i=1}^r \sigma_i(X)$

We define its dual norm $\|x\|_*$ as

$$\|x\|_* = \max_{\|z\| \leq 1} \; z^T x$$

Dual norm of dual norm: can show that $\|x\|_{**} = \|x\|$

Proof: consider the (trivial-looking) problem

$$\min_y \ \|y\| \ \text{ subject to } \ y = x$$

whose optimal value is $\|x\|$. Lagrangian:

$$L(y,u) = \|y\| + u^T(x-y) = \|y\| - y^T u + x^T u$$

Using definition of $\|\cdot\|_*$,

- If $\|u\|_* > 1$, then $\min_y \{\|y\| - y^T u\} = -\infty$
- If $\|u\|_* \leq 1$, then $\min_y \{\|y\| - y^T u\} = 0$

Therefore Lagrange dual problem is

$$\max_u \ u^T x \ \text{ subject to } \ \|u\|_* \leq 1$$

By strong duality $f^\star = g^\star$, i.e., $\|x\| = \|x\|_{**}$

Lagrange对偶与共轭函数(续)

Conjugates appear frequently in derivation of dual problems, via

$$-f^*(u) = \min_x \; f(x) - u^T x$$

in minimization of the Lagrangian. For example, consider

$$\min_x \; f(x) + g(x)$$

Equivalently: $\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; x = z$. Dual function:

$$g(u) = \min_x \; f(x) + g(z) + u^T(z - x) = -f^*(u) - g^*(-u)$$

Hence dual problem is

$$\max_u \; -f^*(u) - g^*(-u)$$

Lagrange对偶与共轭函数(续)

- Norms: the dual of

$$\text{Primal}: \quad \min_x \ f(x) + \|x\|$$

$$\text{Dual}: \quad \max_u \ -f^*(u) \ \text{ subject to } \ \|u\|_* \leq 1$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

两点注记

- Often, we will transform the dual into an equivalent problem and still call this the dual. Under strong duality, we can use solutions of the (transformed) dual problem to characterize or compute primal solutions

  Warning: the optimal value of this transformed dual problem is not necessarily the optimal primal value

- A common trick in deriving duals for unconstrained problems is to first transform the primal by adding a dummy variable and an equality constraint

  Usually there is ambiguity in how to do this. Different choices can lead to different dual problems!

牛顿法

Given unconstrained, smooth convex optimization

$$\min_x \ f(x)$$

where f is convex, twice differentable, and $\mathrm{dom}(f) = \mathbb{R}^n$. Recall that gradient descent chooses initial $x^{(0)} \in \mathbb{R}^n$, and repeats

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

In comparison, Newton's method repeats

$$x^{(k)} = x^{(k-1)} - \left(\nabla^2 f(x^{(k-1)})\right)^{-1} \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Here $\nabla^2 f(x^{(k-1)})$ is the Hessian matrix of f at $x^{(k-1)}$

牛顿法

Recall the motivation for gradient descent step at x: we minimize the quadratic approximation

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

over y, and this yields the update $x^+ = x - t\nabla f(x)$

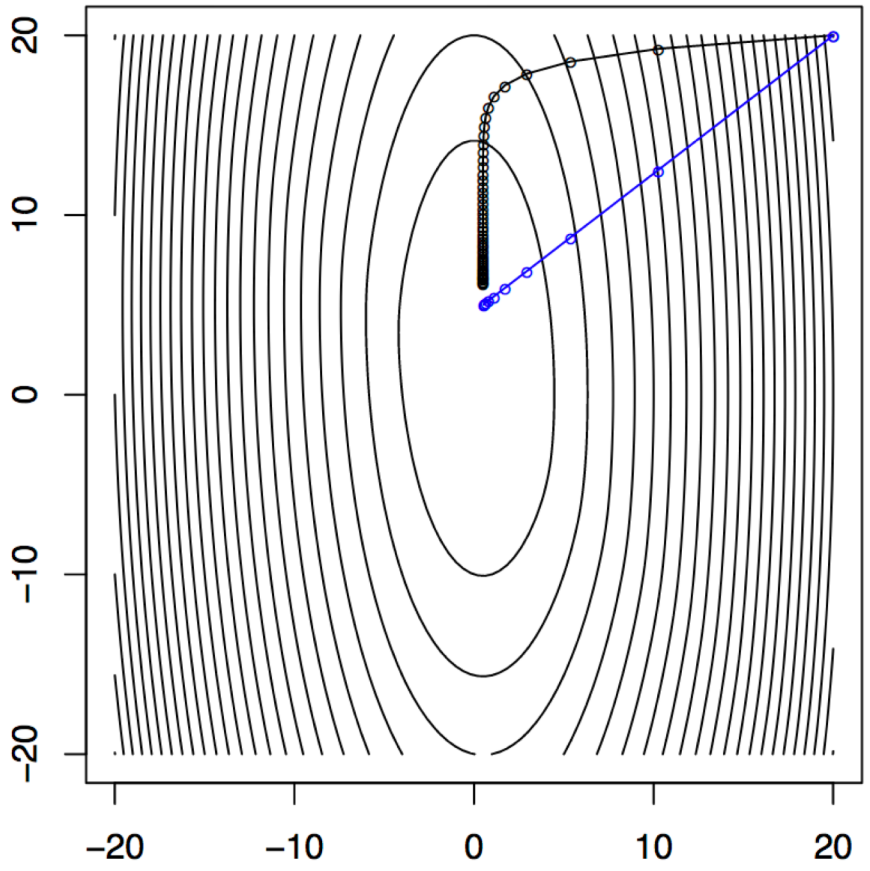
Newton's method uses in a sense a better quadratic approximation

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

and minimizes over y to yield $x^+ = x - (\nabla^2 f(x))^{-1} \nabla f(x)$

Consider minimizing $f(x) = (10x_1^2 + x_2^2)/2 + 5\log(1 + e^{-x_1 - x_2})$

We compare gradient descent (black) to Newton's method (blue), where both take steps of roughly same length



Thank you for your attentions !