对偶理论

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：强凸情形的梯度法收敛速率

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$

Assuming Lipschitz gradient as before, and also strong convexity:

Theorem: Gradient descent with fixed step size $t \leq 2/(m+L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^\star \leq \gamma^k \frac{L}{2}\|x^{(0)} - x^\star\|_2^2$$

where $0 < \gamma < 1$

Rate under strong convexity is $O(\gamma^k)$, exponentially fast! That is, it finds ϵ-suboptimal point in $O(\log(1/\epsilon))$ iterations

回顾：凸情形的梯度法收敛速率

Assume that f convex and differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally that ∇f is Lipschitz continuous with constant $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y$$

Theorem: Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

and same result holds for backtracking, with t replaced by β/L

We say gradient descent has convergence rate $O(1/k)$. That is, it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations

回顾：凸情形的次梯度法收敛速率

- 假设$\left\|x^0 - x^*\right\| \leqslant R$，并且总迭代步数$k$是给定的，在固定步长下，

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2} \leq \frac{R^2}{2kt} + \frac{G^2 t}{2}.$$

- 由平均值不等式知当t满足$\frac{R^2}{2kt} = \frac{G^2 t}{2}$，即$t = \frac{R}{G\sqrt{k}}$时，右端达到最小.
- k步后得到的上界是

$$\hat{f}^k - f^* \leqslant \frac{GR}{\sqrt{k}}$$

- 这表明在$k = O(1/\epsilon^2)$步迭代后可以得到$\hat{f}^k - f^* \leq \epsilon$的精度

上述收敛速率远小于梯度法的收敛速率

课后选做题：试推导强凸情形的次梯度法的收敛速率？

回顾：近似点梯度法的收敛性

For criterion $f(x) = g(x) + h(x)$, we assume:

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$, and ∇g is Lipschitz continuous with constant $L > 0$
- h is convex, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated

Theorem: Proximal gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

Proximal gradient descent has convergence rate $O(1/k)$ or $O(1/\epsilon)$. Matches gradient descent rate! (But remember prox cost ...)

注意：类似于梯度法，若g强凸，收敛率可提升到$O(\log \frac{1}{\epsilon})$

回顾：Nestrov加速策略

As before, consider:

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

where g convex, differentiable, and h convex. Accelerated proximal gradient method: choose initial point $x^{(0)} = x^{(-1)} \in \mathbb{R}^n$, repeat:

$$v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$$
$$x^{(k)} = \text{prox}_{t_k}\big(v - t_k \nabla g(v)\big)$$

for $k = 1, 2, 3, \ldots$

- First step $k = 1$ is just usual proximal gradient update
- After that, $v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$ carries some "momentum" from previous iterations
- When $h = 0$ we get accelerated gradient method

上述策略对梯度法也适用

回顾：加速近似点梯度法的收敛性

For criterion $f(x) = g(x) + h(x)$, we assume as before:

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$, and ∇g is Lipschitz continuous with constant $L > 0$
- h is convex, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated

Theorem: Accelerated proximal gradient method with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{2\|x^{(0)} - x^\star\|_2^2}{t(k+1)^2}$$

and same result holds for backtracking, with t replaced by β/L

Achieves optimal rate $O(1/k^2)$ or $O(1/\sqrt{\epsilon})$ for first-order methods

一般的约束优化问题

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x), \\ \text{s.t.} \quad & c_i(x) \le 0,\ i \in \mathcal{I}, \\ & c_i(x) = 0,\ i \in \mathcal{E}, \end{aligned}$$

其中c_i为定义在$\mathbb{R}^n$或其子集上的实值函数，$\mathcal{I}$和$\mathcal{E}$分别表示不等式约束和等式约束对应的下标集合且各下标互不相同.

- 这个问题的可行域定义为

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid c_i(x) \le 0,\ i \in \mathcal{I} \text{ 且 } c_i(x) = 0,\ i \in \mathcal{E}\}.$$

某些时候，约束优化转化为无约束优化问题会使得目标函数性质不好(如不连续)，这导致我们难以分析其理论性质以及设计有效的算法。

拉格朗日函数(Lagrangian)

一般的约束优化问题:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & c_i(x) \le 0,\ i \in \mathcal{I},\ |\mathcal{I}| = m \\ & c_i(x) = 0,\ i \in \mathcal{E},\ |\mathcal{E}| = p \end{aligned}$$

变量$x \in \mathbb{R}^n$, 最优值为p^*, 定义域为

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid c_i(x) \le 0,\ i \in \mathcal{I} \text{ 且 } c_i(x) = 0,\ i \in \mathcal{E}\}$$

拉格朗日函数$L : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$

$$L(x, \lambda, \nu) = f(x) + \sum_{i \in \mathcal{I}} \lambda_i c_i(x) + \sum_{i \in \mathcal{E}} \nu_i c_i(x)$$

- λ_i 为第i 个不等式约束对应的拉格朗日乘子
- ν_i 为第i 个等式约束对应的拉格朗日乘子

Lagrange对偶函数

拉格朗日对偶函数$g:\mathbb{R}_+^m \times \mathbb{R}^p \to [-\infty,+\infty)$

$$g(\lambda,\nu) = \inf_{x\in\mathbb{R}^n} L(x,\lambda,\nu)$$

$$= \inf_{x\in\mathbb{R}^n} \left(f(x) + \sum_{i\in\mathcal{I}} \lambda_i c_i(x) + \sum_{i\in\mathcal{E}} \nu_i c_i(x) \right)$$

定理 (弱对偶原理)

若$\lambda \geq 0$, 则$g(\lambda,\nu) \leq p^*$.

证明:若$\tilde{x} \in \mathcal{X}$, 则

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) \leq L(\tilde{x},\lambda,\nu) \leq f(\tilde{x}),$$

对$\tilde{x}$取下界得

$$g(\lambda,\nu) \leq \inf_{\tilde{x}\in\mathcal{X}} f(\tilde{x}) = p^*.$$

弱队偶原理也称为下界性质(lower bound property)

Lagrange对偶函数

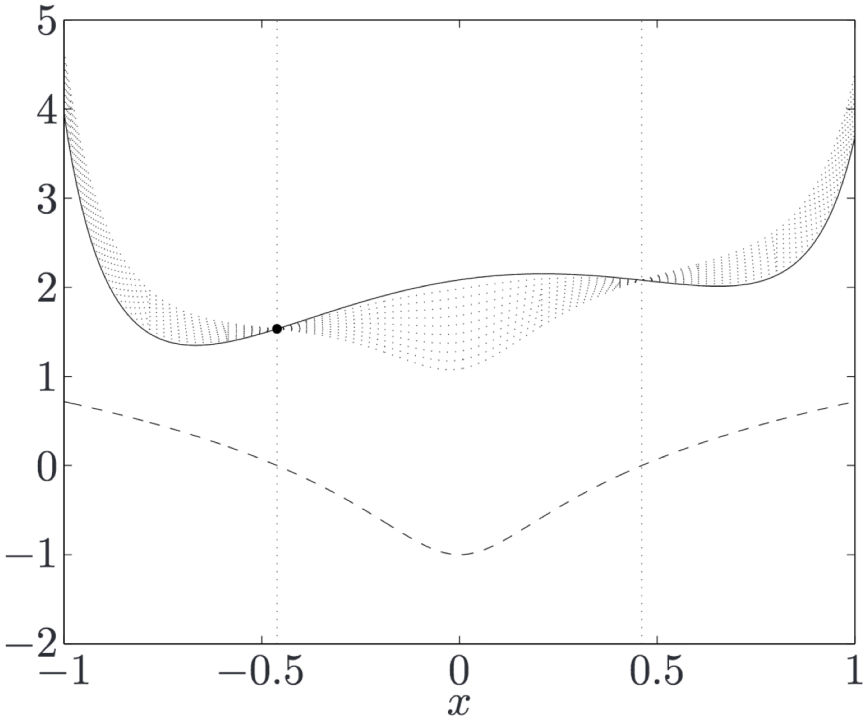


Figure 5.1 *Lower bound from a dual feasible point.* The solid curve shows the objective function f_0, and the dashed curve shows the constraint function f_1. The feasible set is the interval $[-0.46, 0.46]$, which is indicated by the two dotted vertical lines. The optimal point and value are $x^\star = -0.46$, $p^\star = 1.54$ (shown as a circle). The dotted curves show $L(x, \lambda)$ for $\lambda = 0.1,\ 0.2, \ldots, 1.0$.

详见Boyd版216-217页

标准线性规划

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \quad x \succeq 0 \end{array}$$

dual function

- Lagrangian is

$$\begin{aligned} L(x, \lambda, \nu) &= c^T x + \nu^T (Ax - b) - \lambda^T x \\ &= -b^T \nu + (c + A^T \nu - \lambda)^T x \end{aligned}$$

- L is affine in x, hence

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

lower bound property: $p^\star \geq -b^T \nu$ if $A^T \nu + c \succeq 0$

线性方程组的极小范数解

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

dual function

- Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$
- to minimize L over x, set gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \quad \Longrightarrow \quad x = -(1/2) A^T \nu$$

- plug in in L to obtain g:

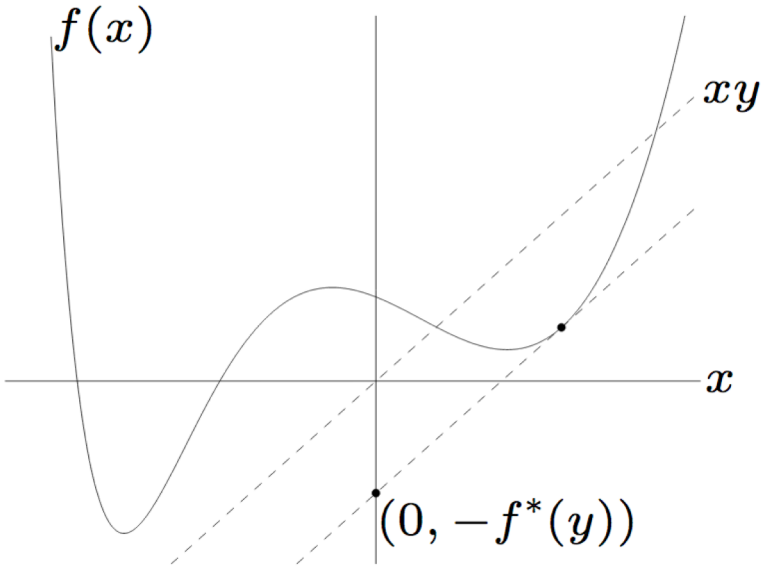
$$g(\nu) = L((-1/2) A^T \nu, \nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$$

lower bound property: $p^\star \geq -(1/4) \nu^T A A^T \nu - b^T \nu$ for all ν

共轭函数

the **conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$



- f^* is convex (even if f is not)

课后作业：证明共轭函数总是为一凸函数

共轭函数

- negative logarithm $f(x) = -\log x$

$$
\begin{aligned}
f^*(y) &= \sup_{x>0}(xy + \log x) \\
&= \begin{cases} -1 - \log(-y) & y < 0 \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
$$

- strictly convex quadratic $f(x) = (1/2)x^T Q x$ with $Q \in \mathbf{S}^n_{++}$

$$
\begin{aligned}
f^*(y) &= \sup_x (y^T x - (1/2)x^T Q x) \\
&= \frac{1}{2} y^T Q^{-1} y
\end{aligned}
$$

Lagrange对偶与共轭函数

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b, \quad Cx = d \end{array}$$

dual function

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathbf{dom}\, f_0} \left(f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu \right) \\ &= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu \end{aligned}$$

- recall definition of conjugate $f^*(y) = \sup_{x \in \mathbf{dom}\, f}(y^T x - f(x))$
- simplifies derivation of dual if conjugate of f_0 is known

等式约束范数极小化

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b \end{array}$$

dual function

$$g(\nu) = \inf_x(\|x\| - \nu^T Ax + b^T\nu) = \begin{cases} b^T\nu & \|A^T\nu\|_* \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

where $\|v\|_* = \sup_{\|u\|\leq 1} u^T v$ is dual norm of $\|\cdot\|$

The conjugate of $f_0 = \|\cdot\|$ is given by

$$f_0^*(y) = \begin{cases} 0 & \|y\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

Lagrange对偶问题

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $d^\star$
- λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathbf{dom}\, g$

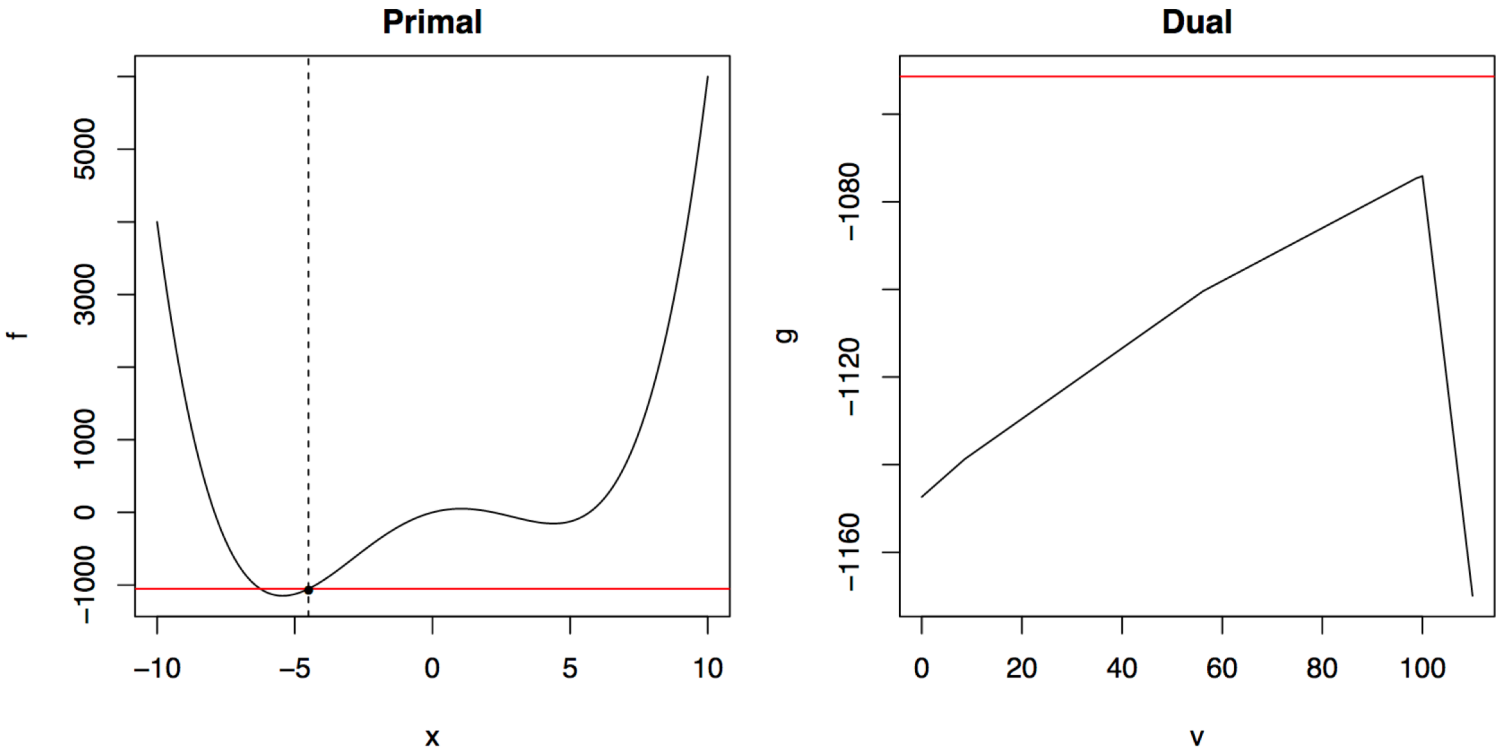
example: standard form LP and its dual

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \qquad \begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

课后作业：证明Lagrange对偶问题总是为一凸优化问题

Example: nonconvex quartic minimization

Define $f(x) = x^4 - 50x^2 + 100x$ (nonconvex), minimize subject to constraint $x \geq -4.5$



弱对偶与强对偶

weak duality: $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications** (约束规范性条件)

Duality gap: $p^* - d^*$

Thank you for your attentions !