课程要点回顾

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. 6

要点1:凸集与凸函数

line segment between x_1 and x_2: all points

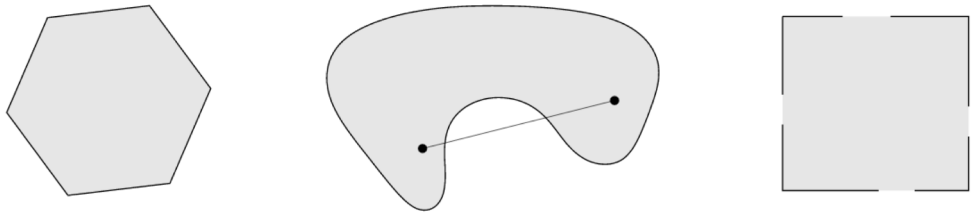
$$x = \theta x_1 + (1-\theta)x_2$$

with $0 \leq \theta \leq 1$

convex set: contains line segment between any two points in the set

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta)x_2 \in C$$

examples (one convex, two nonconvex sets)



注意凸集与仿射集的区别

要点1:凸集与凸函数

- Intersection: the intersection of convex sets is convex
- Scaling and translation: if C is convex, then

$$aC + b = \{ax + b : x \in C\}$$

is convex for any a, b

- Affine images and preimages: if $f(x) = Ax + b$ and C is convex then

$$f(C) = \{f(x) : x \in C\}$$

is convex, and if D is convex then

$$f^{-1}(D) = \{x : f(x) \in D\}$$

is convex

上述关于凸集的保凸运算常被用来验证集合的凸性

要点1:凸集与凸函数

Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for } 0 \leq t \leq 1$$

and all $x, y \in \mathrm{dom}(f)$

In words, function lies below the line segment joining $f(x), f(y)$

Concave function: opposite inequality above, so that

$$f \text{ concave} \iff -f \text{ convex}$$

凹函数也通常被称为下凸函数

要点1:凸集与凸函数

Important modifiers:

- Strictly convex: $f\big(tx + (1-t)y\big) < tf(x) + (1-t)f(y)$ for $x \neq y$ and $0 < t < 1$. In words, f is convex and has greater curvature than a linear function

- Strongly convex with parameter $m > 0$: $f - \frac{m}{2}\|x\|_2^2$ is convex. In words, f is at least as convex as a quadratic function

Note: strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex

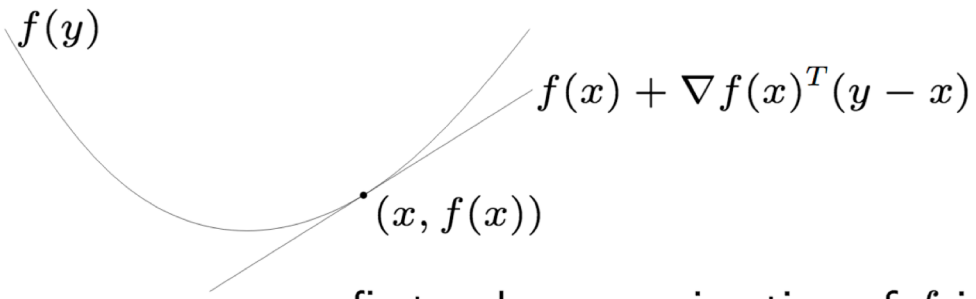
(Analogously for concave functions)

思考：强凸(凹)函数有哪些等价性质？

要点1:凸集与凸函数

1st-order condition: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathbf{dom}\, f$$



first-order approximation of f is global underestimator

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

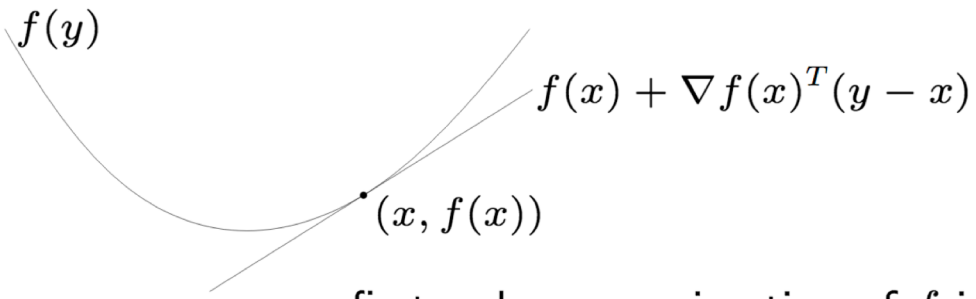
- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

上述两个条件常被用于验证一个函数的凸性

要点1:凸集与凸函数

1st-order condition: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathbf{dom}\, f$$



first-order approximation of f is global underestimator

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then f is strictly convex

上述两个条件常被用于验证一个函数的凸性

要点1:凸集与凸函数

nonnegative multiple: αf is convex if f is convex, $\alpha \geq 0$

sum: $f_1 + f_2$ convex if f_1, f_2 convex (extends to infinite sums, integrals)

composition with affine function: $f(Ax + b)$ is convex if f is convex

examples

- log barrier for linear inequalities

$$f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x), \qquad \mathbf{dom}\, f = \{x \mid a_i^T x < b_i, i = 1, \ldots, m\}$$

- (any) norm of affine function: $f(x) = \|Ax + b\|$

请仔细阅读boyd版3.2节了解更多保凸运算

要点2:凸优化问题

Reminder: a convex optimization problem (or program) is

$$\begin{array}{ll} \min\limits_{x \in D} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & Ax = b \end{array}$$

where f and g_i, $i = 1, \ldots, m$ are all convex, and the optimization domain is $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^{m} \mathrm{dom}(g_i)$ (often we do not write D)

- f is called criterion or objective function
- g_i is called inequality constraint function
- If $x \in D$, $g_i(x) \leq 0$, $i = 1, \ldots, m$, and $Ax = b$ then x is called a feasible point
- The minimum of $f(x)$ over all feasible points x is called the optimal value, written $f^\star$

要点2:凸优化问题

- If x is feasible and $f(x) = f^\star$, then x is called optimal; also called a solution, or a minimizer
- If x is feasible and $f(x) \leq f^\star + \epsilon$, then x is called ϵ-suboptimal
- If x is feasible and $g_i(x) = 0$, then we say g_i is active at x
- Convex minimization can be reposed as concave maximization

$$\begin{array}{lll} \min\limits_x & f(x) & \\ \text{subject to} & g_i(x) \leq 0, & \\ & i = 1, \ldots, m & \\ & Ax = b & \end{array} \iff \begin{array}{lll} \max\limits_x & -f(x) & \\ \text{subject to} & g_i(x) \leq 0, & \\ & i = 1, \ldots, m & \\ & Ax = b & \end{array}$$

Both are called convex optimization problems

要点2:凸优化问题

Let X_{opt} be the set of all solutions of convex problem, written

$$\begin{array}{lll} X_{\text{opt}} = & \text{argmin} & f(x) \\ & \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & & Ax = b \end{array}$$

Key property 1: X_{opt} is a convex set

Proof: use definitions. If x, y are solutions, then for $0 \leq t \leq 1$,

- $g_i(tx + (1-t)y) \leq tg_i(x) + (1-t)g_i(y) \leq 0$
- $A(tx + (1-t)y) = tAx + (1-t)Ay = b$
- $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) = f^\star$

Therefore $tx + (1-t)y$ is also a solution

Key property 2: if f is strictly convex, then solution is unique, i.e., X_{opt} contains one element

思考：如何证明上述性质2？

要点3:梯度法与加速梯度法

Consider the problem

$$\min_x \ f(x)$$

for f convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$. Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

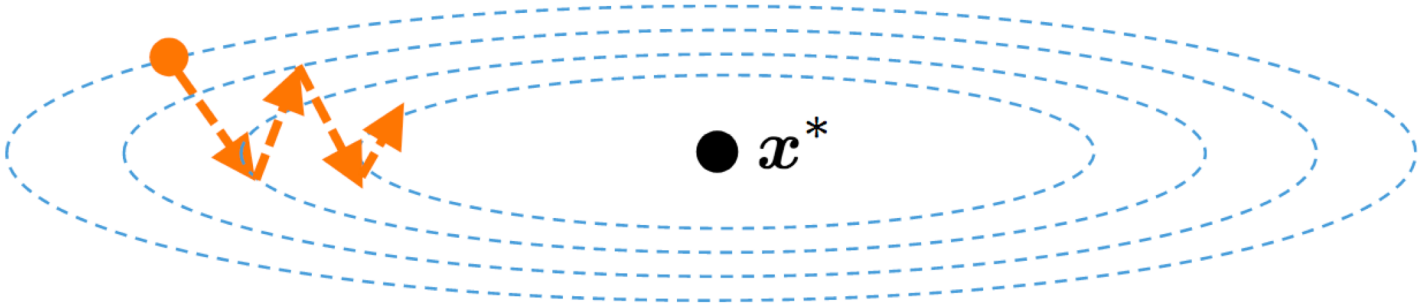
$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Step sizes t_k chosen to be fixed and small, or by backtracking line search

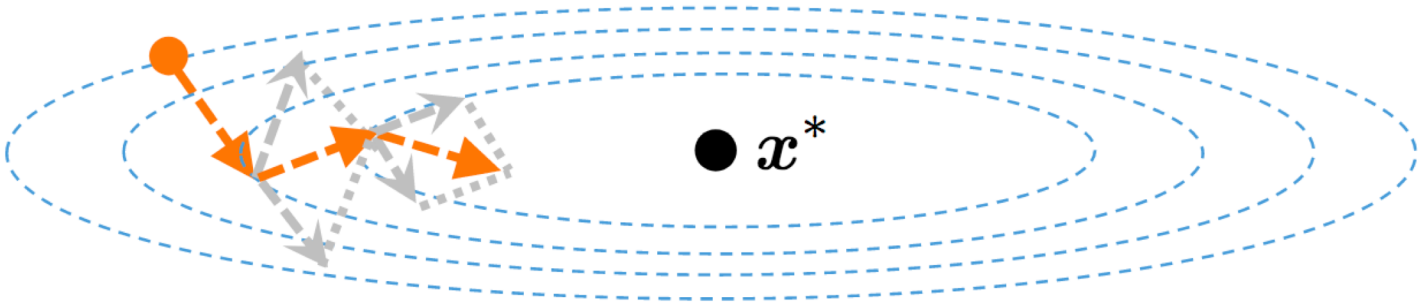
If ∇f is Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$. Downsides:

- Requires f differentiable
- Can be slow to converge

要点3:梯度法与加速梯度法



gradient descent



要点3:梯度法与加速梯度法

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:

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

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

要点3:梯度法与加速梯度法

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

注意：若g强凸，加速策略本质不起作用，即收敛速率均为$O(\log \frac{1}{\varepsilon})$

要点4:次梯度与次微分

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

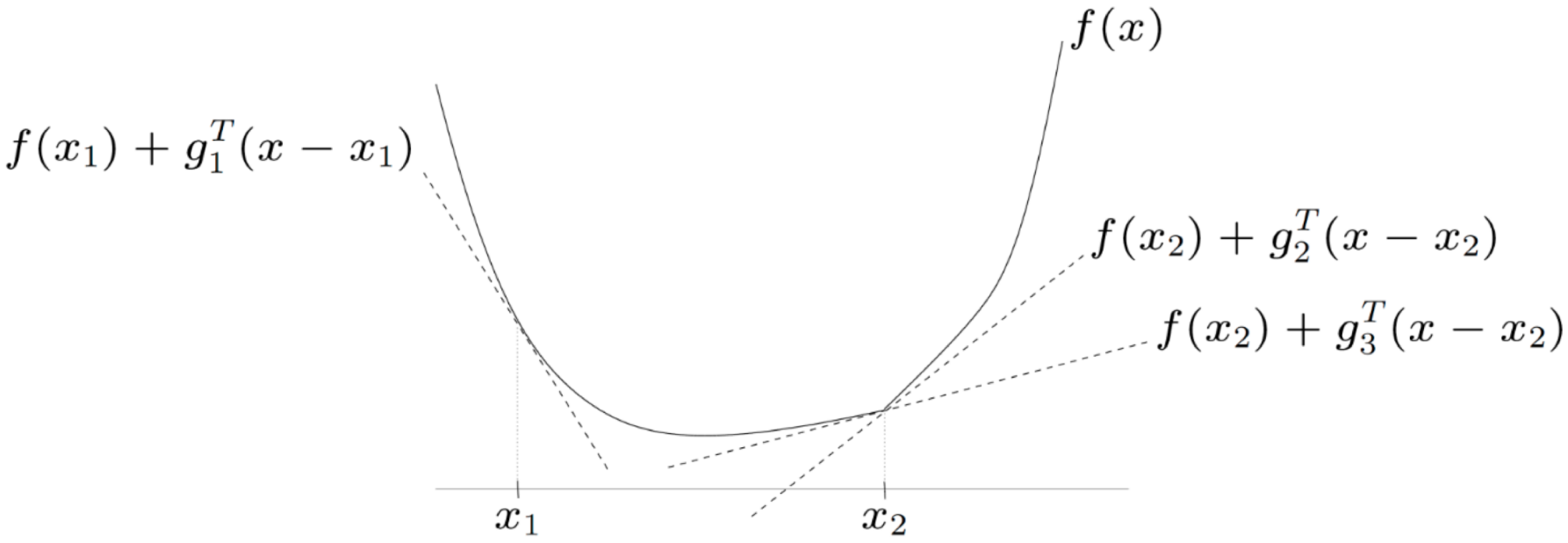
That is, linear approximation always underestimates f

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

- Always exists
- If f differentiable at x, then $g = \nabla f(x)$ uniquely

要点4:次梯度与次微分



g_2, g_3 are subgradients at x_2; g_1 is a subgradient at x_1

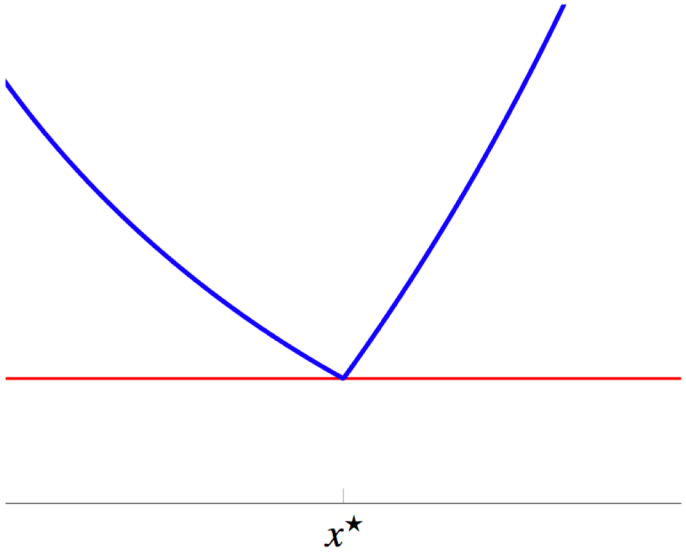
the **subdifferential** $\partial f(x)$ of f at x is the set of all subgradients:

$$\partial f(x) = \{g | g^\top (y - x) \leq f(y) - f(x)\} \quad \forall\, y \in \mathbf{dom}\, f$$

要点4:次梯度与次微分

$x^\star$ minimizes $f(x)$ if and only

$$0 \in \partial f(x^\star)$$



this follows directly from the definition of subgradient:

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) \quad \text{for all } y \qquad \Longleftrightarrow \qquad 0 \in \partial f(x^\star)$$

上述充要条件要求f为凸函数

要点4:次梯度与次微分

Now consider f convex, having $\mathrm{dom}(f) = \mathbb{R}^n$, but not necessarily differentiable

Subgradient method: like gradient descent, but replacing gradients with subgradients. Initialize $x^{(0)}$, repeat:

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

where $g^{(k-1)} \in \partial f(x^{(k-1)})$, any subgradient of f at $x^{(k-1)}$

Subgradient method is not necessarily a descent method, thus we keep track of best iterate $x^{(k)}_{\text{best}}$ among $x^{(0)}, \ldots, x^{(k)}$ so far, i.e.,

$$f(x^{(k)}_{\text{best}}) = \min_{i=0,\ldots,k} f(x^{(i)})$$

要点5：Lagrange对偶与KKT条件

standard form problem (not necessarily convex)

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

variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

Lagrangian: $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(x) \leq 0$
- ν_i is Lagrange multiplier associated with $h_i(x) = 0$

要点5：Lagrange对偶与KKT条件

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\
&= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\end{aligned}
$$

lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$
f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)
$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

要点5：Lagrange对偶与KKT条件

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

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

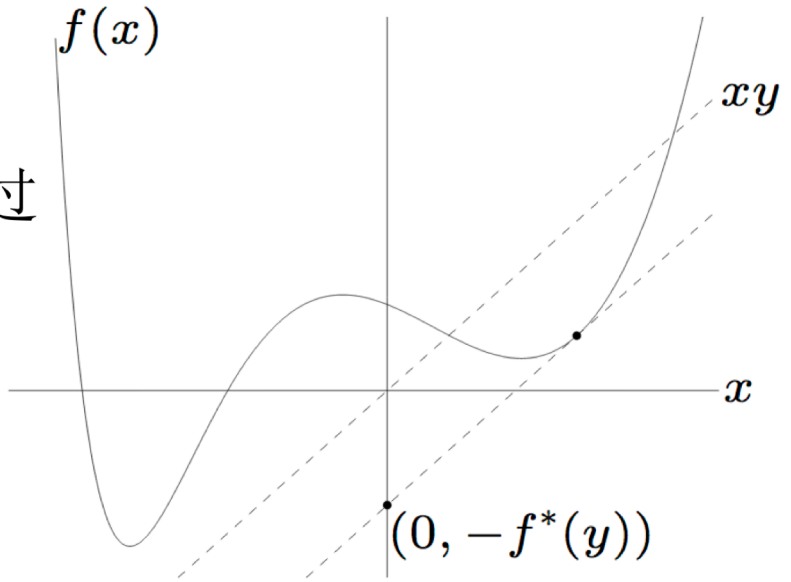
思考：如何证明上述右边问题为左边问题的对偶问题？

要点5：Lagrange对偶与KKT条件

the **conjugate** of a function f is

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

找到f上的一点x，使得通过它的切线斜率为y



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

要点5：Lagrange对偶与KKT条件

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

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

要点5：Lagrange对偶与KKT条件

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: $p^* - d^*$

要点5：Lagrange对偶与KKT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

要点5：Lagrange对偶与KKT条件

Given general problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & \ell_j(x) = 0,\ j = 1, \ldots, r \end{array}$$

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)
- $h_i(x) \leq 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

给定一个凸优化问题，要求可以写出其相应的KKT条件

要点5：Lagrange对偶与KKT条件

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：

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

要点6：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 $x^+ = x - (\nabla^2 f(x))^{-1} \nabla f(x)$

注：若Hessian为正定矩阵，则牛顿方向为下降方向

要点6：牛顿法

- Memory: each iteration of Newton's method requires $O(n^2)$ storage ($n \times n$ Hessian); each gradient iteration requires $O(n)$ storage (n-dimensional gradient)
- Computation: each Newton iteration requires $O(n^3)$ flops (solving a dense $n \times n$ linear system); each gradient iteration requires $O(n)$ flops (scaling/adding n-dimensional vectors)
- Backtracking: backtracking line search has roughly the same cost, both use $O(n)$ flops per inner backtracking step

注：当数据样本不是很大时，牛顿法具有明显的计算优势

要点6：牛顿法

definition

- convex $f : \mathbf{R} \to \mathbf{R}$ is self-concordant if $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \mathbf{dom}\, f$
- $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$

examples on R

- linear and quadratic functions
- negative logarithm $f(x) = -\log x$
- negative entropy plus negative logarithm: $f(x) = x \log x - \log x$

给定一个凸函数，要求能判断它是否自和谐

要点6：牛顿法

summary: there exist constants $\eta \in (0, 1/4]$, $\gamma > 0$ such that

- if $\lambda(x) > \eta$, then

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$$

- if $\lambda(x) \leq \eta$, then

$$2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2$$

(η and γ only depend on backtracking parameters α, β)

complexity bound: number of Newton iterations bounded by

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(1/\epsilon)$$

牛顿法具有明显的局部二次收敛性质

试题类型

- 判断题(10道，共20分)
- 叙述题(4道，共20分)
- 证明与计算题(4道，共60分)

Thank you for your attentions !