

MATH466/MATH766

Math of machine learning

02/05-02/10 Lecture 8-9 Optimality Conditions

References:

- <https://sites.math.washington.edu/~burke/crs/408f/notes/nlp/unoc.pdf>
- Ch5 of Convex Optimization by Stephen Boyd and Lieven Vandenberghe

Todays contents: optimality conditions for

- unconstrained smooth optimization problems
- unconstrained nonsmooth optimization problems
- constrained optimization problems

Important concepts:

-

Recommend reading:

-

Review

1. necessary condition and sufficient condition

$$p \Rightarrow q \quad \begin{array}{l} q \text{ is necessary to } p \\ p \text{ is sufficient to } q \end{array}$$

2. positive semi-definite (p.s.d.)

positive definite (p.d.)

A square matrix $A \in \mathbb{R}^{n \times n}$ is p.s.d. (p.d.) iff

$$\text{for any } p \in \mathbb{R}^n, p \neq 0, \quad p^T A p \geq 0 \quad (p^T A p > 0)$$

3. property of cts func.

If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous, $f(x_0) > 0$

then there exist $r > 0$ s.t. for any $\|x - x_0\| < r$, $f(x) > 0$

4. Taylor's Thm.

Suppose that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$.

Then we have that

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad \text{for some } t \in (0,1)$$

Moreover, if f is twice continuously differentiable, we have that

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p \quad \text{for some } t \in (0,1)$$

Consider $\min_{x \in \mathbb{R}^n} f(x)$, f is differentiable

1. Necessary condition

$$f(x^*) \leq f(x) \Rightarrow ?$$

Intuition

$$f(x+p) = f(x^*) + (\nabla f(x^* + tp))^T p \quad \text{for some } t \in (0,1)$$

$$\text{if } \nabla f(x^*) \neq 0,$$

$$f(x+p) = f(x^*) + \nabla f(x^*)^T p + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p \quad \text{for some } t \in (0,1)$$

$$= f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p$$

$$\text{if } \nabla^2 f(x^*) \text{ is not p.s.d.,}$$

Thm 1.1. If x^* is a local minimizer and f is continuously differentiable

in an open neighbourhood of x^* , then $\nabla f(x^*) = 0$.

Thm 1.2. If x^* is a local minimizer of f and $\nabla^2 f$ exist and is continuous
in an open neighbourhood of x^* , then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is p.s.d.

2. Sufficient condition

Recall first-order condition for convex function

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

If f is convex and $\nabla f(x^*) = 0$

then $f(y) \geq f(x^*)$ for any $y \Rightarrow x^*$ is a global minimizer

If f is nonconvex, we don't have the inequality

but if for x^* w.l. $\nabla f(x^*) = 0$, f is convex in an open neighbourhood of x^*

Then we can argue locally x^* is a minimizer

Thm 1.3. Suppose that $\nabla^2 f$ is continuous in an open neighbourhood of x^*

and that $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ is _____

Then x^* is a local minimizer of f .

Pf. Because the Hessian is continuous and positive definite at x^* ,

we can choose $r > 0$ so that $\nabla^2 f(x)$ remains p.d. for all x

in $\mathcal{D} = \{z \mid \|z - x^*\| < r\}$. Taking any nonzero vector p w.l. $\|p\| < r$

we have $x^* + p \in \mathcal{D}$ and $f(x^* + p) = f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p$

for some $t \in (0, 1)$. Since $x^* + tp \in \mathcal{D}$, $\nabla^2 f(x^* + tp)$ is p.d.

and therefore $f(x^* + p) > f(x^*)$

□

Rmk. Actually, if the conditions are satisfied,

then x^* is a strictly local minimizer.

Why the 2nd order condition for convex function only requires p.s.d.

but here to have convexity on x^* we require p.d. ?

$$\min_{x \in \mathbb{R}^n} f(x) \quad \begin{array}{l} f \text{ not necessarily differentiable} \\ \text{but convex} \end{array}$$

1. Subgradient and optimality conditions

- Def. (subgradient) Let f be a convex function on a convex set C in \mathbb{R}^n .
 $g \in \mathbb{R}^n$ is called a subgradient of f at $x \in C$ if
 $f(y) \geq f(x) + g^T(y - x)$ holds for any $y \in C$.

Remark. For f convex, $x \in C$, a subgradient always exists
if f is in addition differentiable, then $g = \nabla f(x)$ is unique.

The same definition works for nonconvex function,

but in that case, a subgradient may not exist.

- Def (subdifferential) The set of all subgradients at x
is called the subdifferential at x , denoted as $\partial f(x)$

e.g. ① $f: \mathbb{R} \rightarrow \mathbb{R}, x \mapsto |x|$. $\partial f(x) = \begin{cases} \{\text{sign}(x)\} & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$

② $f: \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto \|x\|_2$ $\partial f(x) = \begin{cases} \left\{ \frac{x}{\|x\|_2} \right\} & \text{if } x \neq 0 \\ \{z \mid \|z\|_2 \leq 1\} & \text{if } x = 0 \end{cases}$

- Optimality condition

x^* is a minimizer of $\min_x f(x)$ iff $0 \in \partial f(x^*)$

2. Subgradient calculus

① differentiable if f is cvx and differentiable at x

$$\text{then } \partial f(x) = \{\nabla f(x)\}$$

② scaling: $\partial(af) = a\partial f$ if $a > 0$

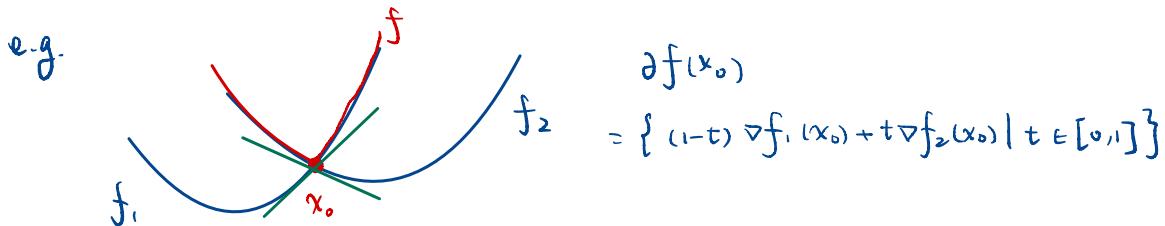
③ addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2 = \{g_1 + g_2 \mid g_1 \in \partial f_1, g_2 \in \partial f_2\}$

④ affine composition: If $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
 $= \{A^T g \mid g \in \partial f(Ax + b)\}$

⑤ finite pointwise maximum: If $f(x) = \max_{i=1, \dots, m} f_i(x)$

$$\text{then } \partial f(x) = \text{conv} \left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

the convex hull of union of subdifferentials
of all active functions at x .



⑤ norms. $f(x) = \|x\|_p$. Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$ ($\frac{p=1, q=\infty}{p=\infty, q=1}$)

$$\text{then } \|x\|_p = \max_{\|z\|_q \leq 1} z^T x$$

$$\text{Hence } \partial f(x) = \operatorname{argmax}_{\|z\|_q \leq 1} z^T x$$