MATH466/MATH766
Math of machine learning
02/10 Lecture 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

$p \Rightarrow q$

q is necessary to p

p is sufficient to q

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

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

3. property of cts func.

If $f: \mathbb{R}^n \to \mathbb{R}$ 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 $f: \mathbb{R}^n \to \mathbb{R}$ 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

$f(x^*) \leq f(x) \Rightarrow ?$ $\nabla f(x^*) = 0$

$\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ is p.s.d.

Intuition

$f(x+p) = f(x^*) + \left(\nabla f(x^* + tp)\right)^T p$ for some $t \in (0,1)$

if $\nabla f(x^*) \neq \underline{0}$, $\exists\, p$ s.t. $\nabla f(x^*)^T p < 0$

$\exists\, t$ small enough s.t. $\left(\nabla f(x^* + tp)\right)^T p < 0$

$\exists\, x$ very close to x^* s.t. $f(x) < f(x^*)$, contradiction

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

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

if $\nabla^2 f(x^*)$ is not p.s.d., $\exists\, p$ s.t. $p^T \nabla^2 f(x^*) p < 0$

$\exists\, t$ small enough s.t. $p^T \nabla^2 f(x^* + tp) p < 0$

$\exists\, x$ very close to x^* s.t. $f(x) < f(x^*)$, contradiction

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

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

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/. $\nabla f(x^*) = 0$, 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 positive definite.
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/ $\|p\| < r$
we have $x^* + p \in \mathcal{D}$ and $f(x^* + p) = f(x^*) + \frac{1}{2} p^\top \nabla^2 f(x^* + tp)\, p$
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. ?
reason can be seen from Taylor Thm

$\min_{x \in \mathbb{R}^n} f(x)$ f not necessarily differentiable but convex

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

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

- Optimality condition

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

2. Subgradient calculus

⓪ differentiable if f is cvx and differentiable at x

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

① scaling: $\partial(af) = a\partial f$ if $a > 0$

② addition: $\partial(f_1+f_2) = \partial f_1 + \partial f_2 = \{g_1+g_2 \mid g_1 \in \partial f_1,\ g_2 \in \partial f_2\}$

③ affine composition: if $g(x) = f(Ax+b)$, then $\partial g(x) = A^T \partial f(Ax+b)$

$= \{A^T g \mid g \in \partial f(Ax+b)\}$

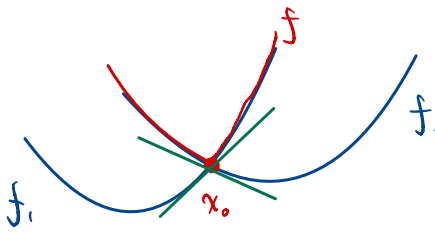
④ finite pointwise maximum: if $f(x) = \max_{i=1,\cdots,m} f_i(x)$

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.

e.g.



$\partial f(x_0)$

$= \{(1-t)\nabla f_1(x_0) + t\nabla f_2(x_0) \mid t \in [0,1]\}$

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

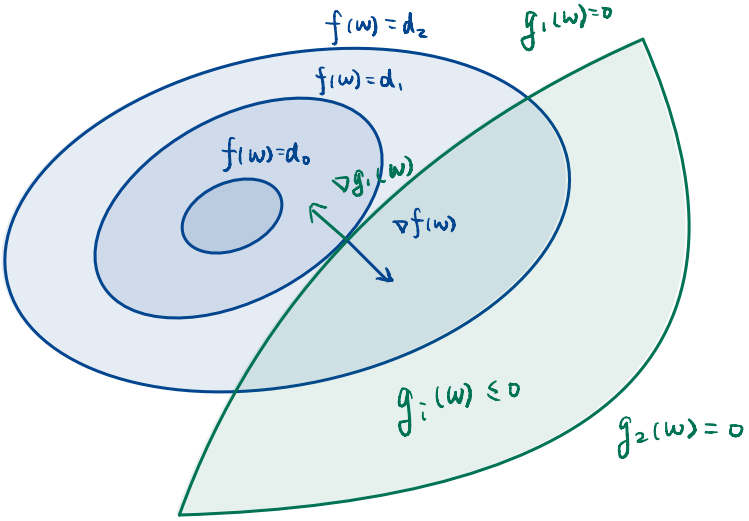
then $\|x\|_p = \max_{\|z\|_q \le 1} z^T x$

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

(Optional)

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$

$$\text{s.t.} \quad g_i(x) \le 0 \qquad i = 1, 2, \cdots, m$$



e.g. SVM $\quad x \leftrightarrow w$

$$f(w) = \frac{1}{2}\|w\|^2$$

$$g_i(w) = 1 - y_i w^T x_i$$

Consider General Optimization Problems

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$

$$\text{Subject to } h_i(x) \le 0, \quad i = 1, \cdots, m \qquad \text{(P)}$$

$$l_j(x) = 0, \quad j = 1, \cdots, p$$

primal prob.

Assume domain $\mathcal{D} := \text{dom} f \cap \bigcap_{i=1}^{m} \text{dom} h_i \cap \bigcap_{j=1}^{p} \text{dom}\, l_j$ is nonempty

feasible set $\mathcal{F} := \mathcal{D} \cap \bigcap_{i=1}^{m} \{x : h_i(x) \le 0\} \cap \bigcap_{j=1}^{p} \{x : l_j(x) = 0\}$

is nonempty

(P) has an optimal value f^*

1. Duality

Define the **Lagrangian** as $L: \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$

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

x is primal variable. u, v are dual variables

$x \in \mathcal{F}$ is primal feasible. $u \geq 0, v$ are dual feasible

then $f(x) = \max_{u \geq 0, v} L(x,u,v)$, $\forall x \in \mathcal{F}$

To see this $L(x,u,v) = f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^p v_j \ell_j(x) \leq f(x)$

($u_i \geq 0$, $h_i(x) \leq 0$, $\ell_j(x) = 0$)

and $L(x, 0, 0) = f(x)$

therefore $f^* = \min_{x \in \mathcal{F}} f(x) = \min_{x \in \mathcal{F}} \max_{u \geq 0, v} L(x,u,v)$

$\Rightarrow$ For any $u \geq 0, v$, $f^* \geq \min_{x \in \mathcal{F}} L(x,u,v) \geq \min_x L(x,u,v) =: g(u,v)$ — Lagrangian dual function

Any dual feasible u, v gives a lower bound on f^*

Thus, consider $\max_{u,v} g(u,v)$ s.t. $u \geq 0$ (D) dual problem

and denote the optimal value as g^*. Then $f^* \geq g^*$

① weak duality $f^* \geq g^*$ always hold (even for nonconvex (P))

② the dual problem (D) is a concave maximization problem

i.e. a convex minimization problem

holds even for nonconvex (P)

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

$$= - \max_x \left\{ -f(x) - \sum_{i=1}^m h_i(x) u_i - \sum_{j=1}^p \ell_j(x) v_j \right\}$$

(linear in u, v thus convex)

($-$ cvx $=$ concave; pointwise max of cvx func. is cvx)

$g(u,v)$ is concave in (u,v), $u \geq 0$ is cvx constraint

thus (D) is equivalent to a convex minimization problem.

e.g.

(P) $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T x$ s.t. $Ax = b$.

$A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$

(D) $\max_{v \in \mathbb{R}^p} -\frac{1}{2} v^T A A^T v - b^T v$

$L(x, v) = \frac{1}{2} x^T x + v^T (Ax - b)$

$g(v) = \min_x L(x, v)$ let $\nabla_x L(x, v) = x + A^T v = 0$

$= L(-A^T v, v) = -\frac{1}{2} v^T A A^T v - b^T v$

③ $f^* - g^*$ is called duality gap, $f^* - g^* \geq 0$

when $f^* - g^* = 0$, we say that strong duality holds

There are some sufficient conditions for strong duality to hold.

Mostly used:

a) (P) is a linear optimization problem

or b) Slater's condition: (P) is a cvx prob. (f, h_i are cvx, l_j are affine)

and there exists at least one x s.t. $h_i(x) < 0$

$l_j(x) = 0$

2. KKT condition

The KKT (Karush-Kuhn-Tucker) conditions are

$$\begin{cases} \nabla f(x) + \sum_{i=1}^{m} u_i \nabla h_i(x) + \sum_{j=1}^{p} v_j \nabla l_i(x) = 0 & \text{(stationarity)} \\ u_i h_i(x) = 0, \quad i = 1, \cdots, m & \text{(Complementary slackness)} \\ h_i(x) \le 0, \ i = 1, \cdots, m, \quad l_j(x) = 0, \ j = 1, \cdots, p & \text{(primal feasibility)} \\ u_i \ge 0, \quad i = 1, \cdots, m & \text{(dual feasibility)} \end{cases}$$

- Necessity

x^*, (u^*, v^*) are primal and dual solutions with zero duality gap

$\Rightarrow$ (x^*, u^*, v^*) satisfy the KKT condition

Pf. $f(x^*) = g(u^*, v^*)$

$$= \min_x f(x) + \sum_{i=1}^{m} u_i^* h_i(x) + \sum_{j=1}^{p} v_j^* l_j(x)$$

$\Rightarrow$ x^* minimize $L(x, u^*, v^*)$ $\Rightarrow$ stationarity

$$\le f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{p} v_j^* l_j(x^*) \quad (1)$$

$$\le f(x^*) \quad (2)$$

$\Rightarrow$ (1) (2) are equalities $\Rightarrow$ complementary slackness $\square$

- Sufficiency

(x^*, u^*, v^*) satisfies KKT conditions

$\Rightarrow$ x^*, (u^*, v^*) are primal and dual solutions

Pf. $g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{p} v_j^* l_j(x^*) = f(x^*)$

(first $=$: By stationarity; second $=$: By complementary slackness)

$\Rightarrow$ zero duality gap

with primal and dual feasibility

$\Rightarrow$ x^*, (u^*, v^*) are primal and dual optimal