Lecture 2. Optimality Condition

2.1 Existence of the optimal solution

Given an optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \Omega\,, \end{array}$$

the *optimal solution* is usually denoted by

$$x^* = \arg\min_{x \in \Omega} f(x)\,.$$

The first question is: for which optimization problems, the optimal solution exist? In general, the question is hard to answer. We only have the following conclusion for some special objective functions and feasible sets.

Theorem (*Weierstrass extreme value theorem*)

Given a compact set S, if function $f : S \to \mathbb{R}$ is continuous on S, then it is bounded and has (both min/max) extreme values.

We now review some definitions in analysis.

Definition (*Open ball*)

For a norm function $\|\cdot\|$ and $n \in \mathbb{N}^+$, an n-dimensional open ball of radius $\epsilon \in \mathbb{R} \geq 0$ is the collection of points of distance less than ϵ. Explicitly, the open ball with center x and radius ϵ is defined by $\mathcal{B}(x, \epsilon) \triangleq \{x' : \|x' - x\| < \epsilon\}$.

Example

The following figure shows the open balls of ℓ_1-norm and ℓ_2-norm:

We can define *open sets* and *closed sets*.

Definition

- (*open set*) A set S is *open* if

$$\forall x \in S, \exists \epsilon > 0, \text{such that } \mathcal{B}(x,\epsilon) \subseteq S$$

- (*closed set*) A set S is *closed* if its complement is open.

For *closed sets*, there is another different but equivalent definition.

Theorem

A set S is *closed* iff for all sequence $\{x_n\}_{n=1}^{\infty}$, where $\forall n$, $x_n \in S$, it holds that

$$\text{if } \lim_{n\to\infty} x_n = x \text{ then } x \in S\,.$$

Example

1. For $(0,1)$, since $\forall x \in (0,1)$, there exists a open ball $\mathcal{B}(x,\epsilon) \subseteq (0,1)$ where $\epsilon = \frac{\min\{x,1-x\}}{2}$, hence, $(0,1)$ is a open set.
2. For $(0,1)$, since $x_n = \frac{1}{2^n} \to 0 \notin (0,1)$, hence, $(0,1)$ is not a closed set.

Then we define *compact sets*.

Definiton (*Compact sets*)

A set S is *compact* if any open cover of it has a finite subcover.

In $\mathbb{R}^n$, there is another definition.

Theorem (Heine–Borel Theorem)

A set $S \subseteq \mathbb{R}^n$ is compact iff it is *bounded* and *closed*.

For optimization problems whose feasible sets are not compact, we usually cannot have simple ways to determine whether optimal solutions exist. However, for continuous function $f : \mathbb{R} \to \mathbb{R}$ and $M \in \mathbb{R}$, if $f(-\infty) = \infty$, $f(\infty) = \infty$, then $\{x : f(x) \leq M\}$ is a compact set, and thus f has minimum values.

2.2 Global minimum and local minimum

Just like the P vs. NP problem, verifying a solution is believed to be easier. So we first study how to justify a solution is indeed an optimal one.

We first identify *global minima* and *local minima*.

Definition

Given a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, where D is $\mathrm{dom}(f)$. A point x is said to be a

- *local minimum point*, if there exists $\varepsilon > 0$ such that

$$\forall\, x' \in \mathcal{B}(x, \varepsilon) \cap D, \quad f(x') \geq f(x)\,;$$

- *global minimum point*, if $\forall\, x' \in D$, $f(x') \geq f(x)$.

The value $f(x)$ is called the *global / local minimum value* of f, respectively.

Similarly, we can also define *strictly global minima* and *strictly local minima*.

Unfortunately, it is too hard to verify global minima in general. It also provides evidence why general optimization problems are difficult to solve. In this course we will study a special type of optimization problem, where local minima are also global minima.

We now give some criteria that can be used to prove local minima.

2.3 First-order optimality condition

Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuous and differentiable. We know that if x^* is a extreme point only if $f'(x) = 0$. Can we have similar results in high dimensions?

The generalization of *derivative* in high dimensions is the *directional derivative*.

Definition (*Directional derivative*)

Given $f : \Omega \to \mathbb{R}$, $\boldsymbol{x}_0 \in \Omega$, $\boldsymbol{v} \in \mathbb{R}^n$, the *directional derivative* of f at $\boldsymbol{x}_0$ with respect to $\boldsymbol{v}$ is defined by

$$\nabla_{\boldsymbol{v}} f(\boldsymbol{x}_0) = \lim_{h \to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{v}) - f(\boldsymbol{x}_0)}{h}$$

if the limit exists.
In particular, if $\boldsymbol{v} = \boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, the *directional derivative* is called the *partial derivative*

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0) = \nabla_{\boldsymbol{e}_i} f(\boldsymbol{x}_0)\,.$$

Given $f : \mathbb{R} \to \mathbb{R}$, we can use $y = f'(x_0)(x - x_0) + f(x_0)$ to do a linear approximation of $f(x)$ at x_0, where $f'(x_0)$ can be seen as a linear mapping. It is natural to define the *differential* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}_0$ by a linear mapping $A : \mathbb{R}^n \to \mathbb{R}$ if $f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + A(\boldsymbol{x} - \boldsymbol{x}_0)$.

Definition (*Differential*)

Given $f : \mathbb{R}^n \to \mathbb{R}^m$, if there exists a matrix $J : \mathbb{R}^n \to \mathbb{R}^m$ (i.e., $J \in \mathbb{R}^{m \times n}$), such that

$$\lim_{x \to x_0} \frac{\|f(x) - f(x_0) - J(x - x_0)\|}{\|x - x_0\|} = 0\,,$$

then we call f is *differentiable* at x_0, and $\mathrm{d}f(x_0) = J$ is the *differential* of f at x_0 (sometimes it also known as the *Jacobian matrix*).
In particular, if $m = 1$, $\nabla f(x_0) = J^\top = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^\top$ is called the *gradient* of f.
If $m \geq 2$, suppose $f : (x_1, \ldots, x_n)^\top \to (f_1, \ldots, f_m)^\top$. Then the Jacobian matrix

is given by

$$\mathrm{d}f = \begin{pmatrix} \nabla f_1^\top \\ \vdots \\ \nabla f_m^\top \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Tip

If f is differentiable at x_0, then the directional derivatives ∇_v at x_0 form a linear mapping with respect to v. Thus it gives that

$$\nabla_v f(x_0) = \nabla f(x_0)^\top v = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot v_i$$

immediately.

Remark

The existence of directional derivatives **cannot** imply the existence of differential.
Consider the following function:

$$f(x,y) = \begin{cases} y^2/x, & x \neq 0 \\ 0, & x = 0 \end{cases}.$$

Then $f(x,y)$ has directional derivative at $(0,0)$ for all direction, but is not differential at $(0,0)$. (Actually, f is even not continuous at $(0,0)$.)

Now we give some examples and calculation rules of differentials.

Example

- $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then $\mathrm{d}f(\boldsymbol{x}) = \boldsymbol{A}$.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ where $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$. Then $\mathrm{d}f(\boldsymbol{x}) = \boldsymbol{w}^\top$ and $\nabla f(\boldsymbol{x}) = \boldsymbol{w}$.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then $\mathrm{d}f(\boldsymbol{x}) = \boldsymbol{x}^\top(\boldsymbol{A} + \boldsymbol{A}^\top)$.

Here is a simple proof of the last example:
$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_{1\le i\le n}\sum_{1\le j\le n} \boldsymbol{A}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j$, so

$$\frac{\partial f}{\partial x_k} = \sum_{1\le i,j\le n} \boldsymbol{A}_{ij}\Big(\frac{\partial x_i}{\partial x_k}(\boldsymbol{x}_i)\cdot \boldsymbol{x}_j + \frac{\partial x_j}{\partial x_k}(\boldsymbol{x}_j)\cdot \boldsymbol{x}_i\Big) = \sum_i \boldsymbol{A}_{ik}\boldsymbol{x}_i + \sum_j \boldsymbol{A}_{kj}\boldsymbol{x}_j\,,$$

which yields that $\nabla f(\boldsymbol{x}) = (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x}$.

Proposition

- *Multiplication*: Given two functions $f, g : \mathbb{R}^n \to \mathbb{R}^m$, let $h : \mathbb{R}^n \to \mathbb{R} = f^\top g$. Then $\mathrm{d}h(x) = f(x)^\top \mathrm{d}g(x) + g(x)^\top \mathrm{d}f(x)$.
- Chain rule: Given $f : \mathbb{R}^n \to \mathbb{R}^m$ differentiable at x_0, $g : \mathbb{R}^m \to \mathbb{R}^\ell$ differentiable at $f(x_0)$, let $h : \mathbb{R}^n \to \mathbb{R}^\ell = g \circ f$ (i.e, $h(x) = g(f(x))$). Then

$$\mathrm{d}h(x_0) = \mathrm{d}g(f(x_0))\,\mathrm{d}f(x_0)\,.$$

We are ready to give the *first-order optimality condition*.

Theorem (*First-order necessary condition*)

Suppose $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is a function differential at some $x^* \in \Omega$ and continuous in $\mathcal{B}(x^*, \varepsilon) \cap \Omega$. If x^* is a local minimum point, then for any feasible direction v (i.e. $\exists\, \varepsilon > 0$ such that $x^* + \delta v \in \Omega$ for any $0 < \delta < \varepsilon$),

$$\nabla_v f(x^*) = \nabla f(x^*)^\top v \ge 0\,.$$

An important idea is to restrict a multivariate function to a line.

Proof

Fix $v \in \mathbb{R}^n$. Define $g : [0, \varepsilon] \to \mathbb{R}$ by $g(t) \triangleq f(x^* + tv)$. Then $g(0) = f(x^*)$. Since x^* is a local minimum point, it holds that $g(t) - g(0) \ge 0$ for any $t > 0$. Therefore, $\frac{g(t)-g(0)}{t} \ge 0$, which gives that
$\nabla_v f(x^*) = g'(0) = \lim_{t\to 0^+} \frac{g(t)-g(0)}{t} \ge 0$.

Corollary

Suppose x^* is further an interior point (i.e., $\exists \varepsilon > 0$ such that $\mathcal{B}(x^*, \varepsilon) \subseteq \Omega$). Then $\nabla f(x^*) = \mathbf{0}$.

Proof

Let $v = -\nabla f(x^*)$. Then $0 \le \nabla_v f(x^*) = -\nabla f(x^*)^\mathsf{T} \nabla f(x^*)$. It implies that $\nabla f(x^*) = \mathbf{0}$.

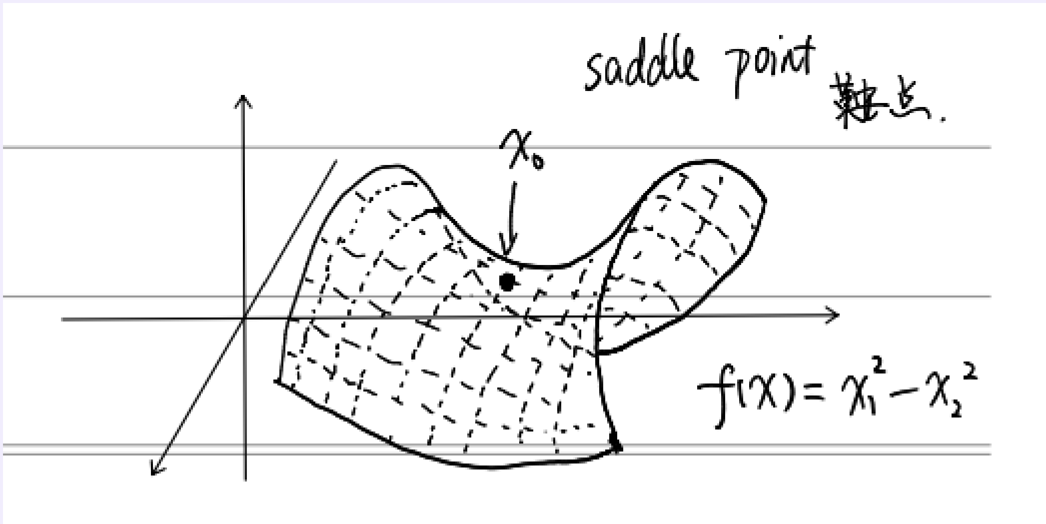
In particular, if Ω is an open set, any point is an interior point. So $\nabla f(x^*) = \mathbf{0}$.

2.4 Second-order optimality condition

Unfortunately, the first-order condition is a necessary condition. If $\nabla f(x^*) = \mathbf{0}$, we still do not know whether x^* is a local minimum. An simple example is function $f(x) = x^3$ and $x^* = 0$. For multivariate functions, there is another case called the *saddle point*.

Example (*Saddle point*)

Consider function $f(x, y) = x^2 - y^2$. Clearly $\nabla f(0, 0) = \mathbf{0}$. But $(0, 0)$ is a *saddle point*, neither a minimum nor a maximum.



We can compute the high-order derivatives to refute saddle points.

For a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, ∇f is a mapping $(x_1, \dots, x_n)^\mathsf{T} \mapsto \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)^\mathsf{T}$. We can further compute the Jacobian matrix of

∇f:

$$J(\nabla f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

The transpose matrix of the Jacobian is called the *Hessian matrix* of f, and denoted by $\boldsymbol{H}(f)$, or $\nabla^2 f$. So $\boldsymbol{H}(f) = J(\nabla f)^\mathsf{T} = \nabla(\nabla f)$.

Theorem (*Schwarz's theorem*, or *Clairaut's theorem*)

Given a function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \Omega$ such that $\mathcal{B}(\boldsymbol{x}, \varepsilon) \subseteq \Omega$ for some $\varepsilon > 0$. If f has continuous $\frac{\partial^2 f}{\partial x_i x_j}$ for all i, j in $\mathcal{B}(\boldsymbol{x}, \varepsilon)$. Then $\frac{\partial^2}{\partial x_i x_j} f(\boldsymbol{x}) = \frac{\partial^2}{\partial x_j x_i} f(\boldsymbol{x})$ for all i, j, which yields that $\boldsymbol{H}(f)(\boldsymbol{x})$ is a symmetric matrix.

We are ready to establish the second-order condition. Consider a function $f : \mathbb{R} \to \mathbb{R}$. Intuitively, if x^* is a local minimum, then we have $f'(x^*) = 0$, $f'(x^* - \varepsilon) < 0$ and $f'(x^* + \varepsilon) > 0$ for sufficiently small $\varepsilon > 0$. Thus $f''(x^*) \geq 0$.

Now let f be a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$. Fix $v \in \mathbb{R}^n$ and consider the restriction of f. Let $g(t) \triangleq f(x^* + tv)$. Using the chain rule, we have

$$\begin{aligned} g'(t) &= \nabla f(x^* + tv) \cdot v = \nabla f(x^* + tv)^\mathsf{T} v\,, \\ g''(t) &= \mathrm{d}g'(t) = \nabla f(x^* + tv)^\mathsf{T}\, \mathrm{d}v + v^\mathsf{T}\, \mathrm{d}(\nabla f(x^* + tv)) = v^\mathsf{T} \nabla^2 f(x^* + tv) v\,. \end{aligned}$$

In particular, we need $g''(0) = v^\mathsf{T} \nabla^2 f(x^*) v \geq 0$.

Another idea is to consider the second-order Taylor series:

$$f(x^* + \delta) = f(x^*) + \nabla f(x^*)^\mathsf{T} \delta + \frac{1}{2} \delta^\mathsf{T} \nabla^2 f(x^*) \delta + o(\|\delta\|^2)\,.$$

Hence we can reasonable guess that $\delta^\mathsf{T} \nabla^2 f(x^*) \delta \geq 0$ since $f(x^* + \delta) \geq f(x^*)$.

Theorem (*Second-order necessary condition*)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function, and x^* is a local minimum. Then $\forall v \in \mathbb{R}^n$,

$$v^\top \nabla^2 f(x^*) v \geq 0\,.$$

Definite matrix

In order to determine whether the Hessian of a function satisfies above condition, we introduce the definition of *definite matrix*.

Definition (*Definite matrix*)

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then A is

- *positive definite* (denoted by $A \succ 0$, or $A > 0$), if $\forall\, v \in \mathbb{R}^n \neq \mathbf{0}$, $v^\top Av > 0$;
- *positive semidefinite* (denoted by $A \succeq 0$, or $A \geq 0$), if $\forall\, v \in \mathbb{R}^n$, $v^\top Av \geq 0$;
- *negative definite* (denoted by $A \prec 0$, or $A < 0$), if $\forall\, v \in \mathbb{R}^n \neq \mathbf{0}$, $v^\top Av < 0$;
- *negative semidefinite* (denoted by $A \preceq 0$, or $A \leq 0$), if $\forall\, v \in \mathbb{R}^n$, $v^\top Av \leq 0$;
- *indefinite*, if $\exists\, v_1, v_2 \in \mathbb{R}^n$, $v_1^\top Av_1 < 0 < v_2^\top Av_2$.

Proposition

Suppose A is a real symmetric matrix, then

- $A \succeq 0$ iff all of its eigenvalues are non-negative,
- $A \succ 0$ iff all of its eigenvalues are positive.

To prove this proposition, we first introduce the *eigendecomposition*, which is a simplified case of SVD (*singular value decomposition*).

Definition (*Eigendecomposition*)

Let $A \in \mathbb{R}^{n\times n}$ be a real symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then A can be decomposed as $A = U\Lambda U^\top$, where $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$ is a diagonal matrix of n eigenvalues, and $U = (u_1, \ldots, u_n)$ consists of orthonormal eigenvectors, namely u_i is an orthonormal eigenvector of corresponding λ_i (i.e., $\forall\, i \neq j$, $\langle u_i, u_j\rangle = 0$ and $\forall\, i$, $\langle u_i, u_i\rangle = 1$, and it implies that $UU^\top = I$).

For any eigenvector u_i, we have $Au_i = \lambda_i u_i$. So $AU = (\lambda_1 u_1, \dots, \lambda_n u_n) = U\Lambda$. Thus $A = U\Lambda U^{-1} = U\Lambda U^\top$;

Proof of the proposition

We use the eigendecomposition of A. Since $A = U\Lambda U^\top$, we have

$$v^\top A v = v^\top U \Lambda U^\top v = (U^\top v)^\top \Lambda (U^\top v)\,.$$

Note that $U^\top v = (u_1, \dots, u_n)^\top v = (u_1^\top v, \dots, u_n^\top v)^\top$. So $v^\top A v = \sum_{i=1}^n \lambda_i (u_i^\top v)^2$. Clearly the result ≥ 0 for all v iff $\lambda_i \geq 0$ for all i (just by letting $v = u_i$).

Example

Consider the following matrix

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Since

$$(a, b) \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 2a^2 - 2ab + 2b^2 = a^2 + b^2 + (a-b)^2 \geq 0\,,$$

and is > 0 if $(a, b) \neq (0, 0)$, A is positive definite.
In addition, each eigenvalue λ of A satisfies $\det(\lambda I - A) = (\lambda - 2)^2 - 1 = 0$. By solving this equation, we obtain that $\lambda = 1, 3$. Since all of the two eigenvalues are positive, A is positive definite.

Sylvester's criterion

Given a matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix},$$

a $k \times k$ *principal submatrix* of A is a submatrix of A, consisting of k rows and k

columns of the same indices $I = \{i_1, \dots, i_k\}$,

$$A_I = \begin{pmatrix} a_{i_1,i_1} & \cdots & a_{i_1,i_k} \\ \vdots & \ddots & \vdots \\ a_{i_k,i_1} & \cdots & a_{i_k,i_k} \end{pmatrix}.$$

The determinant of A_I $\det(A_I)$ is called the *principal minor* (主子式). In particular, if $I = [k] = \{1, \dots, k\}$, $\det(A_I)$ is called the *leading principal minor* (顺序主子式).

Theorem (*Sylvester's criterion*)

Suppose A is a symmetric matrix, then

- $A \succ 0$ iff $D_k(A) \triangleq \det(A_{[k]}) > 0$ for all $k = 1, \dots, n$,
- $A \succeq 0$ iff $D_I(A) \triangleq \det(A_I) \geq 0$ for all $I \subseteq [n]$,
- $A \succeq 0$ if $D_k(A) > 0$ for $k \in [n-1]$, and $D_n(A) \geq 0$.

Remark

We cannot get a criterion for semidefiniteness similar to the first criterion for positive definiteness. Consider the following matrix, all of its principal minor are non-negative. Consider the following example:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}.$$

It is easy to see that $D_k(A) \geq 0$ for all k. However, A is not positive semidefinite.

Second-order sufficient condition

Finally, we give a sufficient condition to assert a local minimum point.

Theorem (*Second-order sufficient condition*)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. Then x^* is a local minimum if $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

Remark

Many minimum points do not satisfy this condition. Consider the function $f(x_1, x_2) = x_1^4 + x_2^4$. Clearly $(0, 0)$ is a local minimum. But the Hessian of f at $(0, 0)$ is $\mathbf{0} \not\succ 0$.