Lecture 3. Analysis in vector spaces (II)

First-order necessary condition

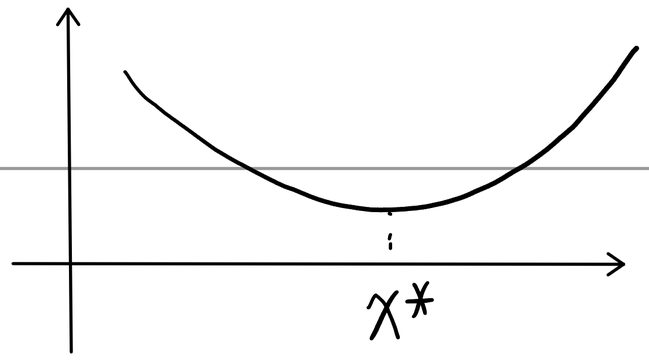
If $f(x^*)$ local minimum. f differentiable at x^*. $\forall$ feasible direction v. $v^T \nabla f(x^*) \geq 0$. (actually $\nabla_v f(x^*)$).

→ 有偏导、方向导数均不够.

However, the condition does not suffice. $f(x) = -x^2$.

鞍点. saddle point $f(x_1, x_2) = x_1^2 - x_2^2$

什么都不是. $f(x) = x^3$. $f(x_1, x_2) = x_1^3 - x_2^3$



$x \uparrow x^*$: $f(x) \downarrow$ $f'(x) < 0$

$x \downarrow x^*$: $f(x) \downarrow$ $f'(x) > 0$.

$f''(x)$?

High-order differentials. $f: \mathbb{R}^n \to \mathbb{R}^m$. $Df: \mathbb{R}^n \to (\mathbb{R}^n \to \mathbb{R}^m)$ $\mathbb{R}^{m \times n}$.

We only consider $f: \mathbb{R}^n \to \mathbb{R}$. $Df: \mathbb{R}^n \to \mathbb{R}^n$. $D^2 f(x): \mathbb{R}^{n \times n}$.

$\nabla f: (x_1, \dots x_n)^T \mapsto \left(\frac{\partial f}{\partial x_1}, \dots \frac{\partial f}{\partial x_n} \right)^T$.

Hessian matrix $H(f) = J(\nabla f) =$

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Schwarz's theorem.

Suppose $f: \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$. $\mathcal{B}(\alpha, \varepsilon) \subseteq \Omega$ for some $\varepsilon > 0$. and f has continuous $\frac{\partial^2 f}{\partial x_i x_j}$ in $\mathcal{B}(\alpha, \varepsilon)$. Then $\frac{\partial^2}{\partial x_i x_j} f(\alpha) = \frac{\partial^2}{\partial x_j x_i} f(\alpha)$. (Hessian is symmetric).

Question: $f''(x) \geq 0 \implies Hf(x) \,/\, \nabla^2 f(x)$??? guess: $v^T \nabla^2 f(x)\, v$

Second-order Taylor expansion

$$f(x_0+\delta) = f(x_0) + f'(x_0)\,\delta + \frac{f''(x_0)}{2}\,\delta^2 + o(\delta^2)$$

$$f(\vec{x}_0+\vec{\delta}) = f(\vec{x}_0) + \nabla f(\vec{x}_0)^T \vec{\delta} + \frac{1}{2}\,\vec{\delta}^T \nabla^2 f(x_0)\,\vec{\delta} + o(\|\vec{\delta}\|^2)$$

Example. $f(x) = w^T x + b$. $\nabla f(x) = w$. $\nabla^2 f(x) = 0$

$f(x) = x^T A x$. $\nabla f(x) = (A + A^T)x$. $\nabla^2 f(x) = A + A^T = 2A$ if A is symmetric.

Verify it. $f(x_0+\delta) = (x_0+\delta)^T A\,(x_0+\delta)$

$= x_0^T A^T \delta$ $= x_0^T A\, x_0 + \delta^T A\, x_0 + x_0^T A \delta + \delta^T A \delta$.

$= x_0^T A\, x_0 + x_0^T (A + A^T)\delta + \delta^T A \delta$.

By Taylor. $f(x_0+\delta) = f(x_0) + \nabla f(x_0)^T \delta + \frac{1}{2}\,\delta^T \nabla^2 f(x)\,\delta$.

Another view: second-order directional derivatives by chain rule.

$h(x) = g(f(x))$. $Dh(x) = Dg(f(x))\,Df(x)$. $D^2 h$?

$f(x) = Ax + b$. $x \in \mathbb{R}^m$ $g: \mathbb{R}^n \to \mathbb{R}$. $\therefore h: \mathbb{R}^m \to \mathbb{R}$. $\nabla h(x) = \nabla g(f(x))\,A$.

$\nabla^2 h(x) = \nabla g(f(x))\,\underbrace{DA}_{=0} + A^T D(\nabla g(f(x))) = A^T \nabla^2 g(f(x))\,A$.

In particular. $f(t) = v \cdot t + x_0$. $h(t) = g(f(t)) = g(x_0 + tv)$.

$h''(t) = \nabla^2 h(t) = v^T \nabla^2 g(f(t))\, v$. $h''(0) \geq 0 \implies v^T \nabla^2 g(x_0)\, v \geq 0$.

Second-order necessary condition.

If $f: \mathbb{R}^n \to \mathbb{R}$. is twice continuously differentiable and x^* is a local minimum of f. then $\forall v \in \mathbb{R}^n$. $v^T \nabla^2 f(x^*) v \geq 0$.

Definite Matrices: Let A be a symmetric matrix. Then A

- positive semidefinite if $\forall v \in \mathbb{R}^n$. $v^T A v \geq 0$.
- positive definite if $\forall v \neq 0 \in \mathbb{R}^n$. $v^T A v > 0$.
- negative semidefinite / definite if $\cdots$ ≤ 0 / < 0.
- indefinite if $\exists u, v \in \mathbb{R}^n$. $u^T A u < 0 < v^T A v$.

Remark. In quadratic forms (二次型) we can always assume symmetry.

since $x^T A x = x^T A^T x = x^T \left(\frac{1}{2}(A + A^T)\right) x$.

Properties of definite matrices. (criterion to verify).

- $A \succeq 0$ iff all eigenvalues ≥ 0. $A \succ 0$ iff eigenvalues > 0.

Example. $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ $\det(\lambda I - A) = (\lambda - 2)^2 - 1 = 0 \Rightarrow \lambda = 1, 3$.

$(a, b) A \begin{pmatrix} a \\ b \end{pmatrix} = (2a - b, -a + 2b) \begin{pmatrix} a \\ b \end{pmatrix} = 2a^2 - 2ab + 2b^2 = a^2 + b^2 + (a - b)^2$.

Why? Let $A \in \mathcal{S}^n$. real symmetric $n \times n$ matrices. eigendecomposition.

$A = U \Lambda U^T$. $\Lambda = \text{diag}\{\lambda_1, \ldots \lambda_n\}$ of n eigenvalues.

$UU^T = I$

$U = (u_1, \dots, u_n)$ u_i is an orthonormal eigenvector of corresponding λ_i.

$A = U\Lambda U^T$. 切换坐标系. 按对应比例伸缩. 再换回原坐标系.

$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T$. $v^T A v = (U^T v)^T \Lambda (U^T v)$

$U^T v = (u_1 \dots u_n)^T v = \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} v = \begin{pmatrix} u_1^T v \\ \vdots \\ u_n^T v \end{pmatrix}$ $= \sum_{i=1}^{n} \lambda_i (u_i^T v)^2$.

Sylvester's criterion. minor of matrices. ≥ 0 if $\forall \lambda_i \geq 0$.
< 0 if $\lambda_i < 0$. let $v = u_i$

Given $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ a $k \times k$ principal submatrix is

a submatrix of A, consisting of k rows and k columns of

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

principal minor 主子式. $|A_I|$ $\det(A_I)$ 主子矩阵特征值.

leading principal minor 顺序主子式 if $I = [k] = \{1, \dots, k\}$.

- $A > 0$ iff $D_k(A) \triangleq \det(A_{[k]}) > 0$ for $k = 1, 2, \dots, n$.

not true for ≥ 0. if only $D_k(A) \geq 0$. counter-example $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$

- $A \geq 0$ iff $\det(A_I) \geq 0$ for all $I \subseteq [n]$.

- (sufficient). $A \geq 0$ if $D_k(A) > 0$ for $k \in [n-1]$. and $D_n(A) \geq 0$.

Proof of second-order necessary condition: $\forall v \in \mathbb{R}^n$. $v^T \nabla^2 f(x^*) v \geq 0$.

otherwise $\exists \lambda < 0$. let v be the eigenvector with respect to λ.

$$f(x^*+tv) = f(x^*) + \underbrace{\nabla f(x^*)^T (tv)}_{=0} + \frac{1}{2}(tv)^T \nabla^2 f(x^*)(tv) + o(\|tv\|^2)$$

$$= f(x^*) + \frac{\lambda}{2} t^2 \|v\|^2 + o(t^2\|v\|^2)$$

$\exists \varepsilon > 0$. if $|t| < \varepsilon$. $|o(t^2\|v\|^2)| < \frac{|\lambda|}{4} t^2 \|v\|^2 \Rightarrow f(x^*+tv) < f(x^*)$. □

sufficient condition? $f(x) = x^3$. $f''(0) = 0$ $= 0$ X. positive definite.

Second-order sufficient condition.

Suppose f is twice continuously differentiable. If $\nabla f(x^*) = 0$, and

$\nabla^2 f(x^*)$ is positive definite. then x^* is a local minimum

Proof. Given $v \neq 0$. $f(x^*+tv) = f(x^*) + \frac{t^2}{2} v^T \nabla^2 f(x^*) v + o(t^2\|v\|^2)$.

$v = a_1 u_1 + \cdots + a_n u_n$. $v^T \nabla^2 f(x^*) v = \sum a_i a_j u_i^T \nabla^2 f(x^*) u_j$

let $\lambda_{min} = \min\{\lambda_1, \ldots, \lambda_n\} > 0$ $= \sum a_i a_j u_i^T \lambda_j u_j$

$= \sum a_i^2 \lambda_i \|u_i\|^2$

$\exists \varepsilon = \varepsilon(v) > 0$, if $|t| < \varepsilon$ $\geq \lambda_{min} \sum a_i^2 \|u_i\|^2$

$o(t^2\|v\|^2) < \frac{\lambda_{min}}{4} t^2 \|v\|^2 = \frac{\lambda_{min}}{4} t^2 \sum a_i^2 \|u_i\|^2 \Rightarrow f(x^*+tv) > f(x^*)$

let $\|v\| = 1$. $\varepsilon(v)$ has extreme value $\varepsilon(v_0) = \min \varepsilon(v) > 0$. □.

In general. optimization problem are difficult to solve. even verifying a local minimum is not easy. We will only focus on some special functions.

Line. Linear. Affine. 最简单的函数：线性函数. $y = kx + b$

Geometry: line: given $x, y \in l$. $z = x + \theta(y-x) = \theta y + (1-\theta)x \in l$ $\forall \theta \in \mathbb{R}$

Remark: line is not a linear space ($ru+v \in l$). it's an affine set.

Affine combination: $\forall \theta_i \in \mathbb{R}$. $\theta_1 + \cdots + \theta_n = 1$. $\theta_1 x_1 + \cdots + \theta_n x_n$

Affine set: given $x_1, \ldots, x_n \in S$. $\forall$ affine combination $\in S$.

Example. solution to a linear equation $S = \{x : Ax = b\}$.

($Ax_1 = Ax_2 = b \Rightarrow \forall \theta_1 + \theta_2 = 1.\ A(\theta_1 x_1 + \theta_2 x_2) = \theta_1 Ax_1 + \theta_2 Ax_2 = b$).

Conversely. any affine set is the solution to a linear equation set.

Why? $S' = S - x_0$ is a linear space. $\forall x_1, x_2 \in S'$ $a_1 x_1 + a_2 x_2 \in S'$.

$x_1 + x_0$, $x_2 + x_0 \in S \Rightarrow a_1 x_1 + a_2 x_2 + x_0 = a_1(x_1 + x_0) + a_2(x_2 + x_0) + (1 - a_1 - a_2)x_0 \in S$

In particular. if $A \in \mathbb{R}^{1 \times n}$. $S = \{x : w^T x = b\}$ is a hyperplane.

If $A \neq 0$. affine set $S \neq \mathbb{R}^n$ is the intersection of finite hyperplanes.