17. Optimization

- Unconstrained optimization problem（制約なし最適化問題）
  - Find $\boldsymbol{x}$ that maximizes (or minimizes) function $f(\boldsymbol{x})$.

Taylor expansion of $f(\boldsymbol{x})$ around $\boldsymbol{x}_0$, $\boldsymbol{x} = \boldsymbol{x}_0 + \Delta\boldsymbol{x},\ \Delta\boldsymbol{x} = \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix}$

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) = f(\boldsymbol{x}_0) + \sum_{i=1}^{n} \left. \frac{\partial f}{\partial x_i} \right|_{\boldsymbol{x}=\boldsymbol{x}_0} \cdot \Delta x_i$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left. \frac{\partial^2 f}{\partial x_i \partial x_j} \right|_{\boldsymbol{x}=\boldsymbol{x}_0} \cdot \Delta x_i \cdot \Delta x_j + \dots$$

$$= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta\boldsymbol{x} + \frac{1}{2} \Delta\boldsymbol{x}^T H(\boldsymbol{x}_0) \Delta\boldsymbol{x} + \dots$$

17. Optimization

The condition under which $f(\boldsymbol{x})$ attains its maximal value（極大値） at $\boldsymbol{x}_0$:

$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_0 + \Delta\boldsymbol{x})$ for $\forall\Delta\boldsymbol{x}$ where $\|\Delta\boldsymbol{x}\|$ is sufficiently small.

$$\Updownarrow$$

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^T H(\boldsymbol{x}_0)\Delta\boldsymbol{x} + \dots$$

$$0 \geq \nabla f(\boldsymbol{x}_0)^T \Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^T H(\boldsymbol{x}_0)\Delta\boldsymbol{x} + \dots$$

The following are the sufficient conditions for $f(\boldsymbol{x})$ to attain its maximal at $\boldsymbol{x}_0$:

$$\begin{cases} \nabla f(\boldsymbol{x}_0) = \boldsymbol{0}, \\ \Delta\boldsymbol{x}^T H(\boldsymbol{x}_0)\Delta\boldsymbol{x} < 0 \text{ for } \forall\Delta\boldsymbol{x}\text{, i.e., } H(\boldsymbol{x}_0) \text{ is negative definite.} \end{cases}$$

(For scalar variable x, the corresponding conditions are,

$$\frac{d}{dx}f(x_0) = 0,\ \frac{d^2}{dx^2}f(x_0) < 0.$$
)

17. Optimization

The conditions described on the previous slide are necessary conditions as well.

When $\Delta \boldsymbol{x}$ is so small that we can ignore the second and higher order terms of $\Delta \boldsymbol{x}$, it is necessary that the following holds,

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x}$$

for any small $\Delta \boldsymbol{x}$. This requires that $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

Then for any $\Delta \boldsymbol{x}$ such that we can ignore its third and higher order terms,

$$\begin{aligned} f(\boldsymbol{x}_0) &\geq f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x} + \frac{1}{2} \Delta \boldsymbol{x}^T H(\boldsymbol{x}_0) \Delta \boldsymbol{x} \\ &= f(\boldsymbol{x}_0) + \frac{1}{2} \Delta \boldsymbol{x}^T H(\boldsymbol{x}_0) \Delta \boldsymbol{x} \end{aligned}$$

$$0 \geq \Delta \boldsymbol{x}^T H(\boldsymbol{x}_0) \Delta \boldsymbol{x}$$

17. Optimization

- Optimization with equality constraints（等式制約つき最適化）
  - Find $\boldsymbol{x}$ that maximizes (or minimizes) function $f(\boldsymbol{x})$ under the condition that,

$$\boldsymbol{h}(\boldsymbol{x}) = \begin{bmatrix} h_1(\boldsymbol{x}) \\ \vdots \\ h_m(\boldsymbol{x}) \end{bmatrix} = \mathbf{0},$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $m < n$.

17. Optimization

– Let us suppose that $\boldsymbol{x}_0$ is the solution, and let us find the conditions that $\boldsymbol{x}_0$ must satisfy

$\boldsymbol{x}_0$ must, of course, satisfy

$$\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{0},$$

and, for sufficiently small $\Delta\boldsymbol{x}$ that satisfies $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) = \boldsymbol{0}$,

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}) = f(\boldsymbol{x}_0 + \Delta\boldsymbol{x}).$$

$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}_0 + \Delta\boldsymbol{x})$ can be approximated as

$$\boldsymbol{h}(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) \simeq \boldsymbol{h}(\boldsymbol{x}_0) + \boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{x}_0)^T \Delta\boldsymbol{x},$$

where $\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{x}_0) = \begin{bmatrix} \nabla h_1(\boldsymbol{x}_0) & \dots & \nabla h_m(\boldsymbol{x}_0) \end{bmatrix}$ is an $n \times m$ matrix.

Since $\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{0},\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, we have $\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{x}_0)^T \Delta\boldsymbol{x} = \boldsymbol{0}$ or

$$\Delta\boldsymbol{x} \in \mathcal{N}(\boldsymbol{\nabla}\boldsymbol{h}(\boldsymbol{x}_0)^T).$$

17. Optimization

Also we can approximate $f(\boldsymbol{x}_0 + \Delta \boldsymbol{x})$ as

$$f(\boldsymbol{x}_0 + \Delta \boldsymbol{x}) \simeq f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x},$$

which leads to

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_0 + \Delta \boldsymbol{x}) \simeq f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x},$$

and therefore

$$0 \geq \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x} \text{ for } \forall \Delta \boldsymbol{x} \in \mathscr{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T).$$

In the above condition, the case that $0 > \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x}$ must be excluded because if $0 > \nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x}$ holds for some $\Delta \boldsymbol{x} \in \mathscr{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T)$, then $-\Delta \boldsymbol{x}$ is also in $\mathscr{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T)$, but $0 < \nabla f(\boldsymbol{x}_0)^T(-\Delta \boldsymbol{x})$.

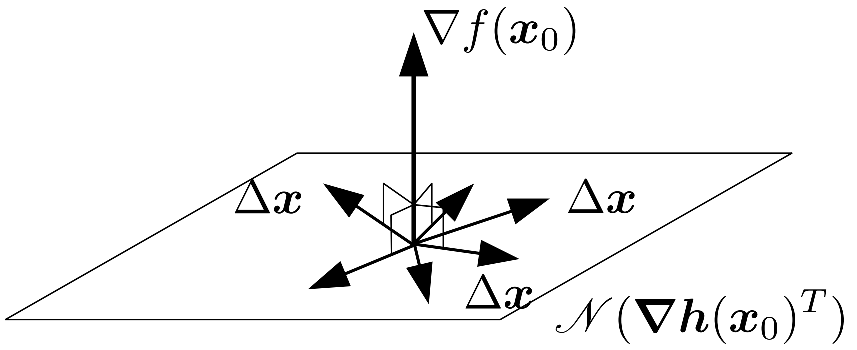
So the condition that must be satisfied is

$$\nabla f(\boldsymbol{x}_0)^T \Delta \boldsymbol{x} = 0 \text{ for } \forall \Delta \boldsymbol{x} \in \mathscr{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T).$$

Note that the above condition is also obtained by starting with $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x}_0 + \Delta \boldsymbol{x})$. So the condition is a necessary condition.

17. Optimization

The condition $\nabla f(\boldsymbol{x}_0)^T \Delta\boldsymbol{x} = 0$ for $\forall \Delta\boldsymbol{x} \in \mathcal{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T)$ implies that the vector $\nabla f(\boldsymbol{x}_0)$ is orthogonal to any vector in $\mathcal{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T)$, namely $\nabla f(\boldsymbol{x}_0)$ is orthogonal to $\mathcal{N}(\boldsymbol{\nabla h}(\boldsymbol{x}_0)^T)$.



17. Optimization

In general,

a vector $\boldsymbol{y}$ is orthogonal to $\mathscr{N}(A^T) \quad \Longleftrightarrow \quad \boldsymbol{y} \in \mathscr{R}(A)$

$A \in \mathbb{R}^{m\times n}$

First, we have $\mathscr{N}(A^T) \perp \mathscr{R}(A)$.

For, $\forall \boldsymbol{x} \in \mathscr{N}(A^T)$,

$$A^T\boldsymbol{x} = \boldsymbol{0},$$

$$\begin{bmatrix} {\boldsymbol{a}_1}^T \\ \vdots \\ {\boldsymbol{a}_n}^T \end{bmatrix} \boldsymbol{x} = \boldsymbol{0},$$

$${\boldsymbol{a}_i}^T\boldsymbol{x} = 0, i = 1, \ldots, n.$$

For, $\forall \boldsymbol{y} \in \mathscr{R}(A)$,

$$\exists \boldsymbol{z}, \boldsymbol{y} = A\boldsymbol{z}$$

$$= \begin{bmatrix} \boldsymbol{a}_1 & \ldots & \boldsymbol{a}_n \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

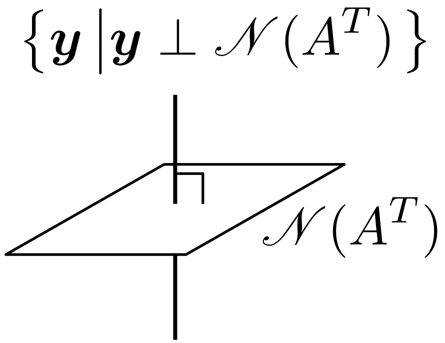
$$= \sum_{i=1}^{n} z_i \boldsymbol{a}_i$$

$$\boldsymbol{y}^T\boldsymbol{x} = \left(\sum_{i=1}^{n} z_i \boldsymbol{a}_i\right)^T \boldsymbol{x} = \sum_{i=1}^{n} z_i {\boldsymbol{a}_i}^T \boldsymbol{x} = 0$$

$$\therefore \boldsymbol{x} \perp \boldsymbol{y}$$

17. Optimization

Next, we examine $\dim\left\{\boldsymbol{y}\,\middle|\,\boldsymbol{y}\perp\mathcal{N}(A^T)\right\}$.



$$\begin{aligned}\dim\left\{\boldsymbol{y}\,\middle|\,\boldsymbol{y}\perp\mathcal{N}(A^T)\right\} &= m-\dim\mathcal{N}(A^T)\\ &= m-\left(m-\dim\mathscr{R}(A^T)\right)\\ &= \dim\mathscr{R}(A^T)\\ &= \operatorname{rank}A^T=\operatorname{rank}A\\ &= \dim\mathscr{R}(A)\end{aligned}$$

So, $\mathcal{N}(A^T)\perp\mathscr{R}(A)$ and the dimension of linear space that consists of all the vectors orthogonal to $\mathcal{N}(A^T)$ is equal to the dimension of $\mathscr{R}(A)$. Therefore $\left\{\boldsymbol{y}\,\middle|\,\boldsymbol{y}\perp\mathcal{N}(A^T)\right\}=\mathscr{R}(A)$.

Therefore $\nabla f(\boldsymbol{x}_0)\in\mathscr{R}(\boldsymbol{\nabla h}(\boldsymbol{x}_0))$.

17. Optimization

The condition $\nabla f(\boldsymbol{x}_0) \in \mathscr{R}(\boldsymbol{\nabla h}(\boldsymbol{x}_0))$ tells that there is a vector $-\boldsymbol{v}$ such that $\nabla f(\boldsymbol{x}_0) = \boldsymbol{\nabla h}(\boldsymbol{x}_0)(-\boldsymbol{v})$ holds.

Therefore we have

$$\nabla f(\boldsymbol{x}_0) + \boldsymbol{\nabla h}(\boldsymbol{x}_0)\boldsymbol{v} = \boldsymbol{0},$$

$$\nabla f(\boldsymbol{x}_0) + \begin{bmatrix} \nabla h_1(\boldsymbol{x}_0) & \dots & \nabla h_m(\boldsymbol{x}_0) \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = \boldsymbol{0},$$

$$\nabla f(\boldsymbol{x}_0) + \sum_{i=1}^{m} v_i \nabla h_i(\boldsymbol{x}_0) = \boldsymbol{0}.$$

This is the necessary condition, together with the condition $\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{0}$, for $\boldsymbol{x}_0$ to be the solution to the optimization problem with equality constraints.

v_i is called a Lagrange multiplier（ラグランジュ乗数）.

17. Optimization

– Example

Maximize

$$f(\boldsymbol{x}) = x_1 + x_2$$

subject to

$$h(\boldsymbol{x}) = {x_1}^2 + {x_2}^2 - 1 = 0.$$

The necessary conditions are

$$\nabla f(\boldsymbol{x}) + v \nabla h(\boldsymbol{x}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + v \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = \begin{bmatrix} 2vx_1 + 1 \\ 2vx_2 + 1 \end{bmatrix} = \mathbf{0},$$

$$h(\boldsymbol{x}) = {x_1}^2 + {x_2}^2 - 1 = 0.$$

We derive

$$x_1 = x_2 = -\frac{1}{2v} \text{ and } {x_1}^2 + {x_2}^2 - 1 = 2\left(\frac{1}{2v}\right)^2 - 1 = 0,$$

and therefore $v = \pm\dfrac{\sqrt{2}}{2}$ and $x_1 = x_2 = \pm\dfrac{\sqrt{2}}{2}.$

17. Optimization

We have two solutions:

$$\begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}, \quad \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}.$$