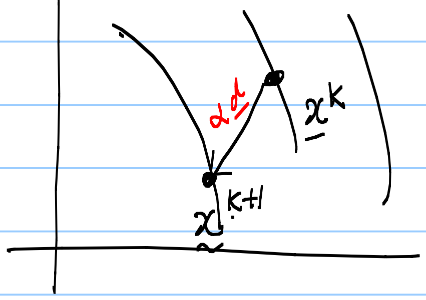
Lecture 3-2 The Steepest Descent Method

노트 제목



▫ Consider the update rule

$$\underset{\sim}{x}^{k+1} = \underline{x}^k + \alpha \, \underset{\sim}{d}^k$$

(with $\alpha > 0$)

to reduce the function value

i.e; $f(\underset{\sim}{x}^{k+1}) < f(\underline{x}^k)$

(repeat until converged)

Question? For what $\underset{\sim}{d}$, $f(\underline{x}^{k+1}) < f(\underline{x}^k)$ valid?

$$f(\underline{x}^{k+1}) = f(\underset{\sim}{x}^k + \alpha \underset{\sim}{d}) \qquad (\alpha > 0)$$

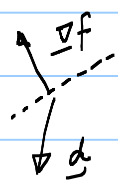
$$\underset{\text{Taylor Expansion}}{=} f(\underline{x}^k) + \alpha \, \underline{\nabla} f^T(\underline{x}^k) \underset{\sim}{d} + O(\alpha^2) \qquad \text{---(1)}$$

To satisfy

$$f(x^{k+1}) < f(x^k) \qquad (\alpha > 0) \qquad \text{--(2)}$$

$\underset{\sim}{d}$ must satisfy

$$\boxed{\nabla f^T(\underline{x}_k) \cdot \underset{\sim}{d} < 0} \qquad (3)$$

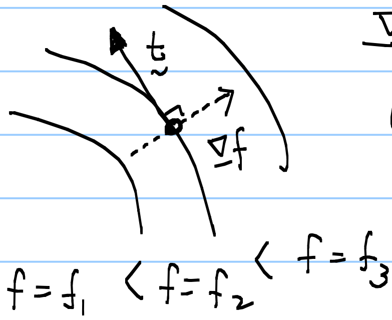


Consequence of (3)

∇f : Function-Increasing direction --(A)

Claim (will be proven)

∇f is orthogonal to the contour of $f = \text{const}$ (B)



$\nabla f^T \underline{t} = 0$

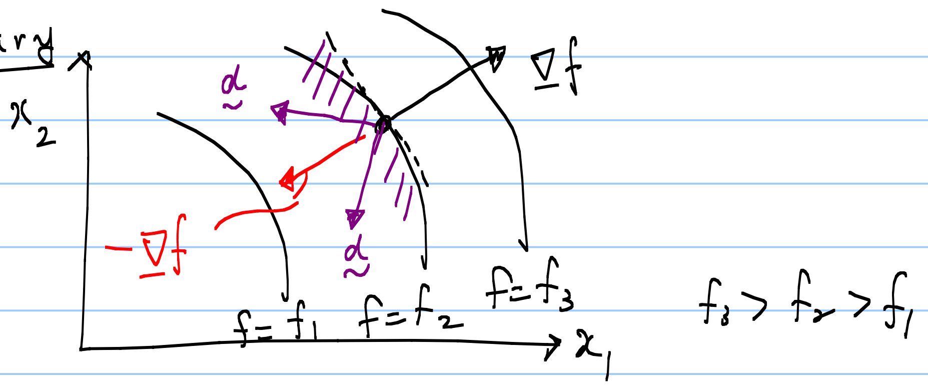
$(\nabla f \cdot \underline{t} = 0)$

($\underset{\sim}{t}$: tangent vector)

$(A, B) \Rightarrow \underset{\sim}{\nabla f}$: the direction of fastest function increase $\rightarrow$ Steepest ascent direction

$\therefore -\underset{\sim}{\nabla f}$: Steepest descencent direction

Summary

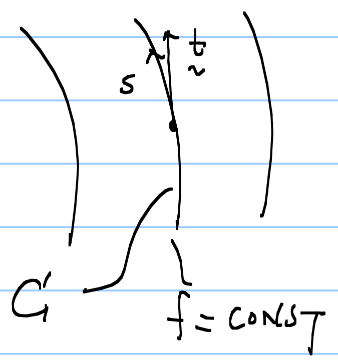


i) f decreases for any $\underset{\sim}{d}$ lying in the shaded region

ii) steepest descent direction $\underset{\sim}{d}$

$$\underset{\sim}{d} = -\underset{\sim}{\nabla} f(\underline{x}^k) \underset{\text{Normalize}}{=} -\frac{\nabla f(\underset{\sim}{x}^k)}{\|\nabla f(\underset{\sim}{x}^k)\|}$$

<proof of Claim B>



Along $f = \text{CONST}$

$$0 = df/ds \quad (s: \text{arclength along } C)$$

$$= \left[\frac{\partial f}{\partial x}\frac{dx}{ds} + \frac{\partial f}{\partial y}\frac{dy}{ds}\right]_{\text{along } C}$$

$$= \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \cdot \left(\frac{dx}{ds}, \frac{dy}{ds}\right)$$

$$= \underline{\nabla} f^T \underset{\sim}{t}$$

For the steepest descent Method

i) search direction $\underset{\sim}{d}_k = -\nabla f(\underline{x}_k) / \|\nabla f(\underline{x}_k)\|$

ii) step $\alpha > 0$?

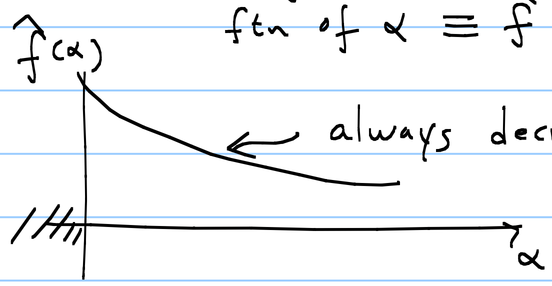
$$\underline{x}^{k+1} = \underline{x}^k + \alpha \underline{d}^k$$

$$\Rightarrow \quad \underset{\sim}{x}^k = \underline{x}^{k+1}(\alpha) \; ; \; \text{1-D problem}$$

$$\therefore \quad \min_{\alpha} \; f(\underbrace{\underline{x}^k + \alpha \underset{\sim}{d}^k}_{}) $$

$$\text{ftn of } \alpha \equiv \hat{f}(\alpha)$$

Remark



Check
$$\left.\frac{d\hat{f}}{d\alpha}\right|_{\alpha=0} = \frac{d}{d\alpha} f(\underbrace{\underline{x}^k + \alpha \underline{d}^k}_{\triangleq \underset{\sim}{x}})\Big|_{\alpha=0}$$

$$\left(x_i = x_i^k + \alpha d_i^k \right) \qquad = \sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i} \frac{\partial}{\partial \alpha} x_i \right]_{\alpha = 0}$$

$$= \sum_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{\underset{\sim}{x}^k} d_i^k = \nabla f^T(\underline{x}^k) \underline{d}^k < 0$$

Recall how $\underline{d}$ was chosen in the steepest descent method

☑ Stopping criteria

i) check the necessary condition for optimality $\| \nabla f(\underset{\sim}{x}^k) \| \leq \epsilon_G$

(usually $O(10^{-6})$)

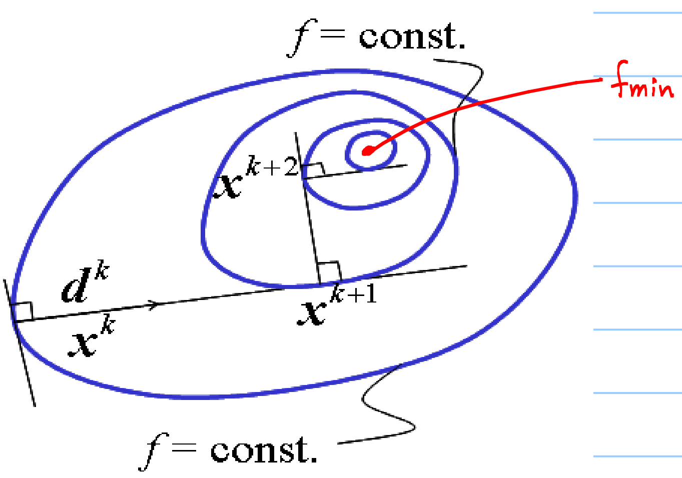
ii) Check the sucessive reduction in f

$$|f(\underset{\sim}{x}^{k+1}) - f(\underset{\sim}{x}^k)| \leq \epsilon_A + \epsilon_R |f(\underset{\sim}{x}^k)|$$

usually $\epsilon_A, \epsilon_R = O(10^{-6})$

(Check at least two sucessive iterations before the stopping the process)

☑ Convergence property = ?



① $\boxed{\underset{\sim}{d}^{k+1} \perp \underset{\sim}{d}^{k} \quad \left(\Leftrightarrow \left(\underset{\sim}{d}^{k+1}\right)^T \underset{\sim}{d}^{k} = 0 \right)}$

→ Every search direction is orthogonal to the previous step

proof: Consider $f(\underset{\sim}{x}^{k+1}) = f(\underset{\sim}{x}^{k} + \alpha^k d^k)$

with
$$\left.\frac{d}{d\alpha} f(\underline{x}^k + \alpha \underline{d}^k)\right|_{\alpha=\alpha_k} = 0 \qquad (a)$$
(by 1-D Search)

Due to (a)

$$0 = \left.\frac{d}{d\alpha} f(\underbrace{\underline{x}^k + \alpha \underline{d}^k}_{\underset{\sim}{x}})\right|_{\alpha=\alpha_k}$$

$$\left(x_i = x_i^k + \alpha d_i^k\right) \qquad = \left[\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial \alpha}\right]_{\alpha=\alpha_k}$$

$$= \sum_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{x^{k+1}} d_i^k = \underline{\nabla} f^T(\underline{x}^{k+1})\, \underline{d}^k$$

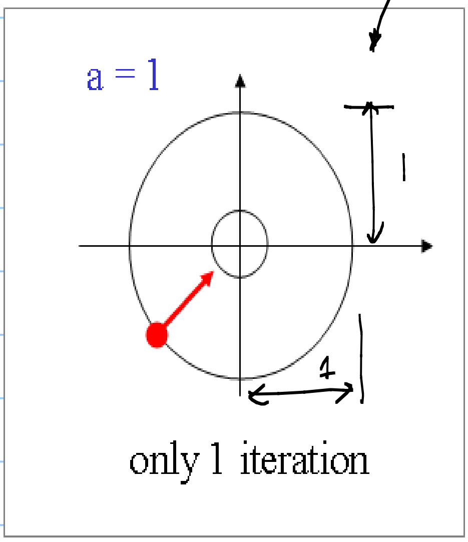
Because $\underset{\sim}{d}^{k+1} = -\underline{\nabla} f(\underset{\sim}{x}^{k+1})$,

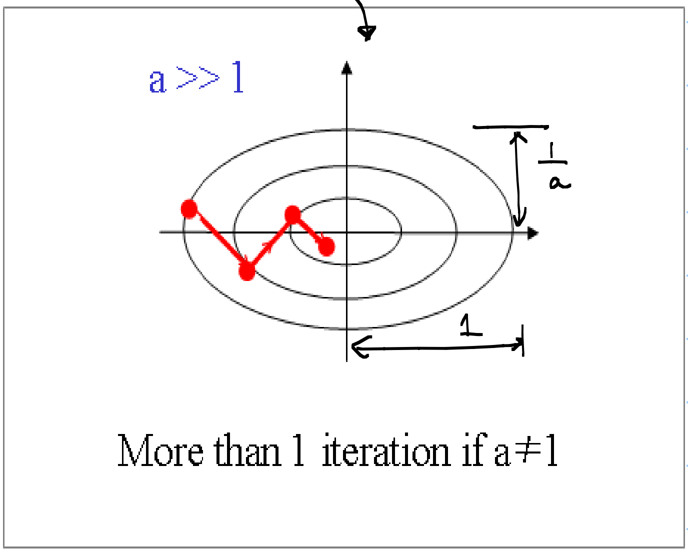
$$\boxed{(\underline{d}^{k+1})^T \underline{d}^k = 0} \Leftarrow$$

However, this relation is valid when the 1-D line search is exact.

② What controls the convergence rate of the steepest descent method?

OBSERVATION : $f(x_1, x_2) = x_1^2 + a x_2^2$





Condition number of Hessian Matrix $\underset{\sim}{H}$ (i.e, $|\lambda_{max}| / |\lambda_{min}|$ of $\underset{\sim}{H}$) affects the convergence property

Consider : $f(x_1, x_2) = x_1^2 + a x_2^2$

■ Check the Hessian of f :

- $\boldsymbol{H} = \left[\dfrac{\partial^2 f}{\partial x_i \partial x_j} \right] = \begin{bmatrix} 2 & 0 \\ 0 & 2a \end{bmatrix}$

- Eigenvalue of $\boldsymbol{H}$

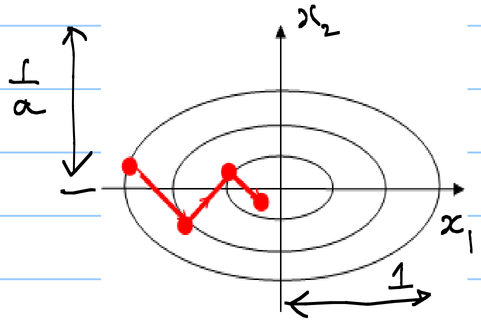
$$\boldsymbol{HZ} - \lambda \boldsymbol{IZ} = 0 \qquad \begin{bmatrix} 2-\lambda & 0 \\ 0 & 2a-\lambda \end{bmatrix} \begin{Bmatrix} z_1 \\ z_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$$

$$det. = 0 \quad \rightarrow \quad \lambda = 2, \; \lambda = 2a$$

$$if \quad a > 1, \qquad Condition \;\; Number = \frac{\lambda_{max}}{\lambda_{min}} = \frac{2a}{2} = a$$

< Trick to Improve Convergence ? >

$f = x_1^2 + a x_2^2$ $\qquad\qquad$ $f = y_1^2 + y_2^2$

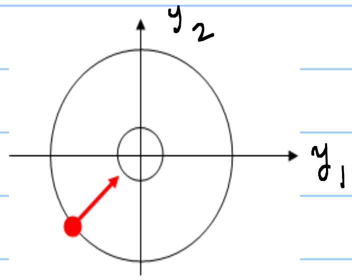


Scaling!!

$\Longrightarrow$

$y_1 = x_1$

$y_2 = \sqrt{a}\, x_2$



$$\Rightarrow \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{a}} \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} \triangleq \underset{\sim}{T} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix}$$

$$\boxed{\underset{\sim}{x}^k = \underset{\sim}{D}^k \underset{\sim}{y}^k}$$ ← Transform $\underset{\sim}{x}$ to $\underset{\sim}{y}$

$$\underset{\sim}{H}_x = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right], \quad \underset{\sim}{H}_y = \left[\frac{\partial^2 f}{\partial y_i \partial y_j} \right]$$

$$\frac{\partial^2 f}{\partial y_i \partial y_j} = D_i \frac{\partial^2 f}{\partial x_i \partial x_j} D_j$$

$$\underset{\sim}{H}_y = \underset{\sim}{D}^T \underset{\sim}{H}_x \underset{\sim}{D}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{a}} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2a \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{a}} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\mathrm{Cond}(H_y) = 1.$$

More General Approch:

$$\underset{\sim}{x} = \underset{\sim}{T} \underset{\sim}{y}$$

several approches to Choose T available

can be non-diagonal matrix though Diagonal matrix is common

Application of $\underset{\sim}{x} = \underset{\sim}{I}\,\underset{\sim}{y}$

$$\boxed{\text{Solve } \underset{\sim}{A}\,\underset{\sim}{x} = \underset{\sim}{b} \quad \text{for "very" large } n}$$

① Convert it as a minimization problem:

$$\min_{\underset{\sim}{x}} Q = \frac{1}{2}\underset{\sim}{x}^T \underset{\sim}{A}\,\underset{\sim}{x} - \underset{\sim}{b}^T \underset{\sim}{x} \qquad (A_{ij} = A_{ji}^T)$$

(• $Q = \frac{1}{2}\sum_i \sum_j A_{ij} x_i x_j - \sum b_i x_j$

• NC for Q to be min

$$\frac{\partial Q}{\partial x_k} = 0 \; ; \; \frac{1}{2}\Big[\sum_i \sum_j A_{ij} \underbrace{\left(\frac{\partial x_i}{\partial x_k}\right)}_{\delta_{ik}} x_j + \sum_i \sum_j A_{ij} x_i \underbrace{\left(\frac{\partial x_j}{\partial x_k}\right)}_{\delta_{jk}}\Big] - \sum_i b_i \underbrace{\left(\frac{\partial x_i}{\partial x_k}\right)}_{\delta_{ik}}$$

$$\left[\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{else} \end{cases}\right]$$

$$= \frac{1}{2}\Big(\sum_j A_{kj} x_j + \sum_i A_{ik} x_i\Big) - b_k$$

$$= \frac{1}{2}\Big(\sum A_{ki} x_i + \sum A_{ik} x_i\Big) - b_k$$

$$= \sum A_{ki} x_i - b_k = 0$$)

$$\Leftrightarrow \underset{\sim}{A}\,\underset{\sim}{x} = \underset{\sim}{b} \quad)$$

② Solve min Q using by an "iterative" optimization algorithm
(such as steepest descent method)

③ To speed up the convergence, may tranform $\underset{\sim}{x}$ as

$$\underset{\sim}{x} = \underset{\sim}{I}\,\underset{\sim}{y}$$

with $\underset{\sim}{I} = [D_{ii}]$ ← Diagonal matrix

$$D_{ii} = 1/\sqrt{|A_{ii}|}$$

Numerical Example of the Steepest descent Method

Consider $f = (x_1-2)^4 + (x_1-2x_2)^2$, $\boldsymbol{x}_0 = (0,3)^T$

$$f(\boldsymbol{x}_0) = 52 \ , \ \nabla f(\boldsymbol{x}) = \left[4(x_1-2)^3 + 2(x_1-2x_2), \ -4(x_1-2x_2)\right]^T$$

- Thus, the search direction is

$$\boldsymbol{d}_0 = -\nabla f(x_0) = [44, -24]^T$$

$$\underset{normalize}{\Rightarrow} \quad \boldsymbol{d}_0 = [0.8779, -0.4789]^T$$

- Line search

Minimize $f(\alpha) = f(\boldsymbol{x}_0 + \alpha\, \boldsymbol{d}_0)$ with $\alpha >0$

$$\Rightarrow \alpha = 3.0841$$

$\therefore$ The new point is

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_0 \boldsymbol{d}_0 = [2.707, 1.523]^T$$