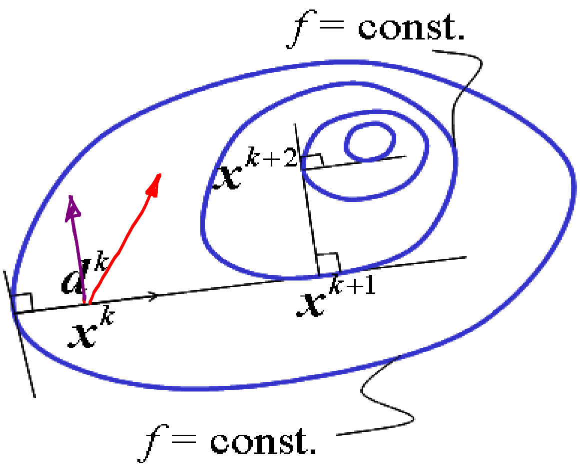
노트 제목

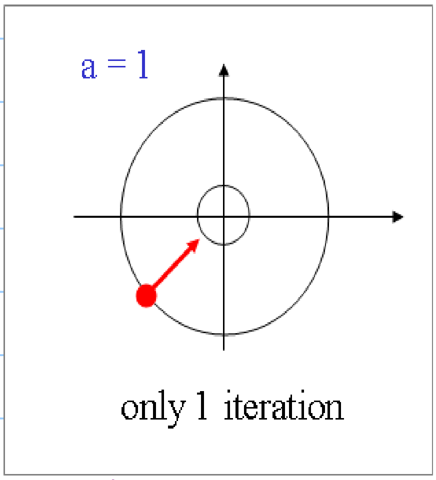
3-3 : Conjugate Gradient Method

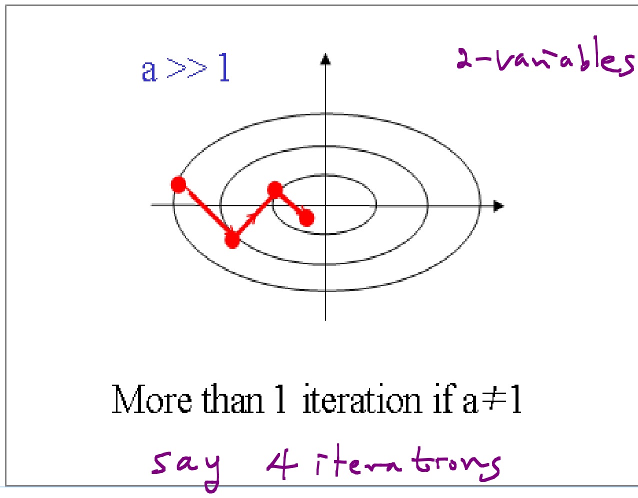
⟨Motivation⟩

i) Recall the convergence pattern of the steepest descent method



ii) Consider $f(x_1, x_2) = x_1^2 + a x_2^2$ ← quadratic ftn of 2 variables





⇒ $\boxed{\text{We want a method yielding a min of } f(x) \text{ within "}n\text{" iterations if } f(x) \text{ is a quadratic function of "}n\text{" variables}}$

⇒ "Conjugate Gradient Method"
(a very powerful method that also works for general non-quadratic functions)

Will study

1) What are conjugate directions?
   i.e., definition of conjugate direction

2) Proof to show n-iteration convergence of $Q(x_1, \cdots, x_n)$ with conjugate search direction.

3) How to find conjugate direction?
   (approximate for non-quadratic ftns)

(1) < Definition of Conjugate Directions >

▫ Given $\underset{\sim}{A}$: $n \times n$ positive-definite symmetric matrix

$\underset{\sim}{d}^i$: conjugate direction
if $(\underset{\sim}{d}^i)^T \underset{\sim}{A} \underset{\sim}{d}^j = 0$ for $i \neq j$

Examples of conjugate directions

Let $\underset{\sim}{A}$: 3×3 sym matrix

$\underline{e}_i$: eigenvector of $\underset{\sim}{A}$ $\quad (\underline{e}_i \underline{A}\, \underline{e}_j \triangleq \delta_{ij})$

Then

$$\begin{cases} (\underline{e}_1, \underline{e}_2, \underline{e}_3) : \text{conjugate direction} \\ (\underline{e}_1+\underline{e}_2,\ \underline{e}_1-\underline{e}_2,\ \underline{e}_3) : \quad \text{''} \\ \quad\vdots \end{cases}$$

※ A special case of conjugate directions = eigenvectors

(2) < proof of n-iteration Convergence >

Min of a quadratic ftn of n variables is found in n iterations or less.

▨ Consider $Q(\underset{\sim}{x}) = \frac{1}{2}\underset{\sim}{x}^T \underset{\sim}{A}\, \underset{\sim}{x} + B^T \underset{\sim}{x} + C$

a quadratic ftn $\qquad$ positive-definite symmetric matrix

$$\underset{\sim}{x} = \{x_1, \cdots, x_n\}^T$$

<u>proof</u>

Let $\underset{\sim}{x}^* =$ min point

then $\underset{\sim}{\nabla} Q(\underset{\sim}{x}^*) = \underline{A}\,\underset{\sim}{x}^* + \underset{\sim}{B} = 0 \qquad (1)$

$$\left(\frac{\partial Q}{\partial x_k} = \frac{\partial}{\partial x_k}\left\{ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i A_{ij} x_j + \sum_{i=1}^{n} B_j x_j + C \right\} \right.$$

$$= \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\delta_{ik} A_{ij} x_j + x_i A_{ij} \delta_{jk} \right) + \sum_{i=1}^{n} B_j \delta_{jk} \right]$$

$$= \frac{1}{2}\left[\sum_{j=1}^{n} A_{kj} x_j + \sum_{i=1}^{n} x_i A_{ik} \right] + B_k$$

$$= \frac{1}{2}\left[\sum_{i=1}^{n} A_{ki} x_i + \sum_{i=1}^{n} \underbrace{A_{ik}}_{A_{ki}} x_i \right] + B_k$$

$$\left. = \sum_{i=1}^{n} A_{ki} x_i + B_k \right)$$

Expand $\underset{\sim}{x}^*$ in terms of orthogonal $\underset{\sim}{d}^i$ $\{i = 0 \,\text{--}, n-1\}$ wrt $\underline{A}$

such that

$$\underset{\sim}{x}^* = \underset{\sim}{x}^0 + \sum_{i=0}^{n-1} \beta^i \underline{d}^i \qquad (2)$$

↑ used to represent an arbitrary starting point

To determine β^i, (2) → (1)

$$\underset{\sim}{B} + \underset{\sim}{A}\,\underset{\sim}{x}^0 + \sum_{i=0}^{n-1} \beta^i A\, \underline{d}^i = 0 \qquad (3)$$

$(\underset{\sim}{d}^{jT})$ (3):

$$\underset{\sim}{d}^{jT}(\underset{\sim}{B} + \underline{A}\,\underline{x}^0) + \underbrace{\sum_{i=0}^{n-1} \beta^i \underbrace{\underline{d}^{jT} \underline{A}\, \underline{d}^i}_{=0 \text{ if } i \neq j}}_{\Downarrow\ \beta^j \underline{d}^{jT} \underline{A}\, \underline{d}^j} = 0 \qquad (4)$$

Thus

$$\boxed{\beta^j = \frac{-(\underset{\sim}{B} + \underline{A}\,\underline{x}^0)^T \underline{d}^j}{(\underline{d}^j)^T \underline{A}\, \underline{d}^j} \qquad (5) \qquad j = 0, 1, 2, \cdots, n-1}$$

⇒ Message: • For "any" starting point $\underline{x}_0$, the min $\underline{x}^*$ of $\underline{Q}(\underline{x})$ is expressed in terms of

$(\underline{d}_i)$; expansion coefficients β_j are computed by (5).

Let us now prove that min of quadratic function $Q(x)$ can be found with n iterations if we update $\underset{\sim}{x}^i$ <u>using $\underline{d}^i$</u> as

$$\boxed{\underset{\sim}{x}^{i+1} = \underset{\sim}{x}^i + \lambda^{i*} \underset{\sim}{d}^i} \qquad (6)$$

$(i = 0, 1, \cdots, n-1)$ $\quad$ conjugate direction

where λ^{i*} is determined by 1-D search by $\min Q(\underset{\sim}{x}^i + \lambda^{i*} \underline{d}^i)$

$$\Leftrightarrow \boxed{0 = \left.\frac{dQ}{d\lambda^i}\right|_{\lambda^i = \lambda^{i*}}} \qquad (7)$$

* If we can prove that

$$\boxed{\lambda^{i*} = \beta^i,} \qquad (i = 0, \cdots, n-1)$$

then
$$\underline{x}^1 = \underline{x}^0 + \lambda^{0*}\underline{d}^0 = \underline{x}^0 + \beta^0 \underline{d}^0$$
$$\underline{x}^2 = \underline{x}^1 + \lambda^{1*}\underline{d}^1 = \underline{x}^0 + \beta^0 \underline{d}^0 + \beta^1 \underline{d}^1$$
$$\vdots$$

$$\underline{x}^n = \underline{x}^0 + \sum_{i=0}^{n-1} \beta^i \underline{d}^i$$

i.e, updating by (6) and (7) yields solutions within n iterations.

☒ Let us prove $\lambda^{i*} = \beta^i$

Use Eq. (7)

$$0 = \left. \frac{dQ(\underline{x}^i + \lambda^i \underline{d}^i)}{d\lambda^i} \right|_{\lambda^i = \lambda^{i*}}$$

$$= \frac{\partial}{\partial \lambda^i} \left[\frac{1}{2} (\underline{x}^i + \lambda^i \underline{d}^i)^T \underline{A} (\underline{x}^i + \lambda^i \underline{d}^i) + \underline{B}^T (\underline{x}^i + \lambda^i \underline{d}^i) + C \right]_{\lambda^i = \lambda^{i*}}$$

$$= \underset{\sim}{d}^{iT} \underline{A} (\underline{x}^i + \lambda^{i*} \underline{d}^i) + \underline{B}^T \underline{d}^i$$

$$= \lambda^{i*} (\underset{\sim}{d}^{iT} \underline{A} \underline{d}^i) + \underbrace{\underline{d}^{iT} \underline{A} \underline{x}^i}_{\text{scalar}} + \underline{B}^T \underline{d}^i$$

$$= \underline{x}^{iT} \overbrace{\underline{A}^T \underline{d}^i}$$

$$\underline{x}^{iT} \underline{A} \underline{d}^i$$

$$\therefore \quad \lambda^{i*} = -\frac{(\underline{B}^T + \underline{x}^{iT}\underline{A})\underline{d}^i}{\underline{d}^{iT}\underline{A}\,\underline{d}^i} \qquad (8)$$

To simply (8), note that

① $$\underline{x}^i = \underline{x}^{i-1} + \lambda^{(i-1)*}\underline{d}^{i-1}$$
$$= \underline{x}^{i-2} + \lambda^{(i-2)*}\underline{d}^{i-2} + \lambda^{(i-1)*}\underline{d}^{i-1}$$
$$\vdots$$
$$= \underline{x}^0 + \sum_{k=0}^{i-1}\lambda^{k*}\underline{d}^k$$

② $$\underline{x}^{iT}\underline{A}\,\underline{d}^i$$
$$= \left(\underline{x}^0 + \sum_{k=0}^{i-1}\lambda^{k*}\underline{d}^k\right)^T \underline{A}\,\underline{d}^i$$
$$= \underline{x}^{0T}\underline{A}\,\underline{d}^i + \sum_{k=0}^{i-1}\lambda^{k*}\underbrace{\underline{d}^{kT}\underline{A}\,\underline{d}^i}_{=0}$$

Because $\underline{d}^k$; conjugate directions $\left\{ i \neq j \right.$

$$= \underline{x}^{0T}\underline{A}\,\underline{d}^i$$
$$= \underline{x}^{0T}\underline{A}^T\underline{d}^i$$
$$= (\underline{A}\,\underline{x}^0)^T\underline{d}^i$$

$$\therefore \quad \lambda^{i*} = -\frac{(\underline{B} + \underline{A}\underline{x}^0)^T\underline{d}^i}{\underline{d}^{iT}\underline{A}\,\underline{d}^i} \equiv \beta^i$$

(3) < How to find $\underset{\sim}{d}$? >

Result :

- for quadratic function $Q(\underline{x}) = \frac{1}{2}\underline{x}^T\underline{A}\underline{x} + \underline{B}^T\underline{x} + c$

  - $\underline{d}^0 = -\underline{g}^0 \quad (\underline{g} = \underset{\sim}{\nabla} Q)$
  - $\underset{\sim}{d}^{k+1} = -\underset{\sim}{g}^{k+1} + \beta^k \underline{d}^k \quad (k \geq 1)$

  where $$\beta^k = \frac{(g^{k+1})^T \underline{A}\, d^k}{\underline{d}^{kT} \underline{A}\, \underline{d}^k}$$

- for non-quadratic function $f(\underline{x})$

  replace β_k as
  $$\beta_k = \frac{\underline{g}^{k+1\,T}\, \underline{g}^{k+1}}{\underline{g}^{k\,T}\, \underline{g}^k}$$

  (Key point : no second-derivative information such as Hessian ($\simeq \underline{A}$) is needed
  $\rightarrow$ due to Fletcher – Reeve)

overall approach

i) derive $\underset{\sim}{d}$ for $Q(\underline{x})$

ii) extend the result for non-quadratic functions

< Derivation >

(A) For $Q(\underline{x}) = \frac{1}{2}\underline{x}^T \underline{A}\, \underline{x} + \underline{B}^T \underline{x} + C$ first

- $\underline{d}_0$: can be any non-zero vector
  to be a conjugate vector at the beginning

- thus, Choose $\underline{d}_0 = -\underline{g}_0$
  where
$$\underset{\sim}{g}^k = = \underset{\sim}{\nabla} Q(\underline{x}^k)$$
$$= \underline{A}\,\underline{x}^k + \underset{\sim}{B} \qquad (a)$$

- Search Algorithm
$$\underline{x}^{k+1} = \underline{x}^k + \underset{\substack{\uparrow \\ \text{found by 1-D search}}}{\alpha^k} \underline{d}^k \qquad (b)$$

- Conjugate direction $\underline{d}^k$

$$\begin{cases} \underline{d}^k = -\underset{\substack{\uparrow \\ \text{known}}}{\underline{g}^k} + \underset{\substack{\uparrow \\ \text{unknown}}}{\beta^{k-1}}\, \underset{\text{known}}{\underline{d}^{k-1}} & (C1) \\ \text{or} \\ \underset{\sim}{d}^{k+1} = -\underline{g}^{k+1} + \beta^k \underline{d}^k & (C2) \end{cases}$$

- To determine β^k, use the conjugate condition.

$$\underline{d}^{kT} \underline{A}\, \underline{d}^{k-1} = 0$$

$$\left[-\underline{g}^k + \beta^{k-1} \underline{d}^{k-1} \right]^T \underline{A}\, \underset{\sim}{d}^{k-1} = 0$$

$$\rightarrow \quad \beta^{k-1} = \frac{(\underline{g}^k)^T \underline{A}\, \underline{d}^{k-1}}{(d^{k-1})^T \underline{A}\, \underset{\sim}{d}^{k-1}}$$

or

$$\boxed{\beta^k = \frac{(\underline{g}^k)^T \underline{A}\, \underline{d}^k}{(\underset{\sim}{d}^k)^T \underline{A}\, \underline{d}^k}} \qquad (d)$$

for non-quadratic ftn, $\underset{\sim}{A}$ can be calculated logically as $\underset{\sim}{H}(x^k)$, but calculation of $\underset{\sim}{H}(x^k)$ should be avoided for computational efficiency

Extension to Non-Quadratic functions

(avoid explicit calculation of $\underset{\sim}{A}$)

Recall

$$\underline{A}\,\underset{\sim}{x}^{k+1} \underset{\text{use (b)}}{=} \underline{A}\,(\underline{x}^k + \alpha^k \underset{\sim}{d}^k)$$

appearing in (d)

$$\Rightarrow \underset{\sim}{A}\,d^k = \frac{1}{\alpha^k}\left(\underset{\sim}{A}\,\underset{\sim}{x}^{k+1} - \underline{A}\,x^k\right)$$

$$= \frac{1}{\alpha^k}\left[\left(A\,\underset{\sim}{x}^{k+1} + B\right) - \left(A\,\underline{x}^k + B\right)\right]$$

$$\underset{\text{use (a)}}{=} \frac{1}{\alpha^k}\underbrace{\left(\underset{\sim}{g}^{k+1} - \underline{g}^k\right)}_{\text{Expressed only in term of gradients}} \qquad \text{(e)}$$

(e) → (d) :

$$\beta^k = \frac{(\underline{g}^{k+1})^T(\underline{g}^{k+1} - \underline{g}^k)}{(\underline{d}^k)^T(\underline{g}^{k+1} - \underline{g}^k)} \qquad \text{(f)}$$

To simplify (f), use

$$\begin{cases} ① & (\underline{g}^{k+1})^T \underline{d}^k = 0 \\ ② & (\underline{d}^k)^T \underline{g}^k = +(\underline{g}^k)^T \underline{g}^k \end{cases}$$

$$③ \quad (\underline{g}^{k+1})^T \underline{g}_k = 0$$

Then
$$\beta^{k}_{(f)} = \frac{(\underset{\sim}{g}^{k+1})^T \underline{g}^{k+1} - \cancel{(\underline{g}^{k+1})^T \underline{g}^{k}}^{\text{ by ③}}}{\underbrace{\cancel{(\underset{\sim}{d}^{k})^T \underset{\sim}{g}^{k+1}}}_{\text{by ①}} - \underbrace{(\underset{\sim}{d}^{k})^T \underline{g}^{k}}_{= -(\underline{g}^{k})^T \underline{g}^{k} \text{ by ②}}}$$

∴
$$\boxed{\beta^{k} = \frac{(g^{k+1})^T \underline{g}^{k}}{(\underset{\sim}{g}^{k})^T \underline{g}^{k}}}$$

proofs

proof of ①: $(\underset{\sim}{g}^{k+1})^T \underline{d}^{k} = 0$

Recall: $f(\underset{\sim}{x}) = f(\underset{\text{known}}{\underline{x}^{k}} + \alpha \underset{\text{known}}{\underline{d}^{k}})$

→ becomes min at $\alpha = \alpha^{k}$ such that

$$0 = \left.\frac{df}{d\alpha}\right|_{\alpha=\alpha^k} = \left[(\underline{\nabla} f)^T \underset{\sim}{d}^{k}\right]_{\alpha=\alpha^k}$$

$$= \left[\underline{\nabla} f\,(\underbrace{\underline{x}^{k} + \alpha^{k} \underline{d}^{k}}_{\underset{\sim}{x}^{k+1}})\right]^T \underline{d}^{k}$$

$$= (\underset{\sim}{g}^{k+1})^T \underline{d}_{k}$$

proof of 2 : $(\underline{d}^k)^T \underline{g}^k = -(\underline{g}^k)^T \underline{g}^k$

$$(\underline{d}^k)^T \underline{g}^k \underset{(c1)}{=} \left[-\underline{g}^k + \beta^{k-1} \underline{d}^{k-1}\right]^T \underline{g}^k$$

$$= -(\underline{g}^k)^T \underline{g}^k + \beta^{k-1} \underbrace{(\underline{d}^{k-1})^T \underline{g}^k}_{\underset{\text{by } ①}{(\underline{g}^k)^T \underline{d}^{k-1} \equiv 0}}$$

$$= -(\underline{g}^k)^T \underline{g}^k$$

proof of 3 : $(\underline{g}^{k+1})^T \underline{g}^k = 0$

$$(\underline{g}^{k+1})^T \underline{g}^k \underset{(c)}{=} (\underline{g}^{k+1})^T \left[-\underline{d}^k + \beta_{k-1} \underline{d}^{k-1}\right]$$

$$= -\underbrace{(\underline{g}^{k+1})^T \underline{d}^k}_{\equiv 0 \text{ by } ①} + \beta_{k-1} (\underline{g}^{k+1})^T \underline{d}^{k-1}$$

$$= \beta_{k-1} (\underline{g}^{k+1})^T \underline{d}^{k-1} \qquad (g)$$

Now rewrite (e) as

$$\underline{g}^{k+1} = \underline{g}^k + \alpha^k \underline{A}\, \underline{d}^k \qquad (h)$$

Then $\underline{g}^{k+1} \underset{(g)}{=} \beta_{k-1} [\underbrace{\underline{g}^k + \alpha_k \underline{A}\, \underline{d}^k}_{(h)}]^T \underline{d}^{k-1}$

$$= \beta_{k-1} \underset{\text{by } ①}{\cancel{(\underline{g}^k)^T \underline{d}^{k-1}}} + \beta_{k-1}\alpha_k \underset{\text{by conjugate cond.}}{\cancel{(\underline{d}^k)^T \underline{A}\, \underline{d}^{k-1}}} = 0$$

< Fletcher-Reeve Conjugate Gradient Method >

- Start with any $\underline{x}_0$
- Search direction

If $k = 0$

$$\underset{\sim}{d}^k = -\underline{g}^k = -\underline{\nabla} f(\underline{x}^k)$$

else

$$\underset{\sim}{d}^k = -\underline{g}^k + \beta^{k-1} \underline{d}^{k-1}$$

$$\beta^{k-1} = \frac{(\underline{g}^k)^T \underline{g}^k}{(\underline{g}^{k-1})^T \underline{g}^{k-1}} = \frac{\| \underline{\nabla} f(\underline{x}^k) \|^2}{\| \underline{\nabla} f(\underline{x}^{k-1}) \|^2}$$

- Repeat until convergence

Remark : ① this method applies to non-quadratic function minimization

② n-iteration convergence is valid only for quadratic ftn

③ may need to reset $\underline{d}^k$ for every n step

Example

Consider $f = x_1^2 + 4x_2^2,\ \boldsymbol{x}_0 = (1,1)^T$

- the steepest descent iteration

$$\boldsymbol{d}_0 = -\nabla f(\boldsymbol{x}_0) = -(2,8)^T$$

$$f(\alpha) = f(\boldsymbol{x}_0 + \alpha \boldsymbol{d}_0) \ \Rightarrow \ \alpha_0 = 0.1308$$

$$\therefore x_1 = x_0 + \alpha_0 \boldsymbol{d}_0 = (0.7385, -0.0462)^T$$

- the conjugate gradient iteration

$$\beta_0 = \frac{\|\nabla f(\boldsymbol{x}_1)\|^2}{\|\nabla f(\boldsymbol{x}_0)\|^2} = 0.0341$$

$$\boldsymbol{d}_1 = -\nabla f^T(\boldsymbol{x}_1) + \beta_0 \boldsymbol{d}_0 = \begin{pmatrix} -1.5451 \\ 0.0966 \end{pmatrix}$$

$$f(\alpha) = f(\boldsymbol{x}_1 + \alpha \boldsymbol{d}_1) \ \Rightarrow \ \alpha_1 = 0.4779$$

$$\therefore x_2 = x_1 + \alpha_1 \boldsymbol{d}_1 \cong (0.0, 0.0)^T$$