

# Matrix Calculus

Anton Andreytsev

Let  $X \in \mathbb{R}^d, Y \in \mathbb{R}^q$  be some normed linear spaces and let  $f : X \rightarrow Y$   
 $f(X + \Delta X) = f(X) + Df(X)[\Delta X] + \bar{o}(\|\Delta X\|), \quad \|\Delta X\| \rightarrow 0$   
 $Df(X)[\Delta X]$  – the Fréchet derivative

From the definition above it follows, that

$$Df(X)[\Delta X] = \lim_{t \rightarrow +0} \frac{f(X + t \cdot \Delta X) - f(X)}{t}$$

In particular case, when  $X \in \mathbb{R}^n, Y \in \mathbb{R}$  (so  $f : \mathbb{R}^n \rightarrow \mathbb{R} \Rightarrow Df(X)[\Delta X] = \nabla f(X)^T \Delta X$

Generalisations:

$$\begin{aligned} x \in \mathbb{R} &\Rightarrow Df(x)[\Delta x] = \nabla f(x) \cdot \Delta x, \quad \nabla f(x) \in \mathbb{R} \\ x \in \mathbb{R}^n &\Rightarrow Df(x)[\Delta x] = \nabla f(x)^T \Delta x, \quad \nabla f(x) \in \mathbb{R}^n \\ x \in \mathbb{R}^{m \times n} &\Rightarrow Df(x)[\Delta x] = \text{Tr}(\nabla f(x)^T \Delta x), \quad \nabla f(x) \in \mathbb{R}^{m \times n} \end{aligned}$$

**Example 1:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = a^T x$

$$f(x + \Delta x) = a^T(x + \Delta x) = \underbrace{a^T x}_{f(x)} + \underbrace{a^T \Delta x}_{Df(x)[\Delta x]} + \underbrace{0}_{\bar{o}(\|\Delta x\|)}$$

$$Df(x)[\Delta x] = \underbrace{a^T}_{\nabla f(x)^T} \Delta x \Rightarrow \boxed{\nabla(a^T x) = a}$$

**Example 2:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = x^T A x$

$$f(x + \Delta x) = (x + \Delta x)^T A (x + \Delta x) = \underbrace{x^T A x}_{f(x)} + \underbrace{\Delta x^T A x + x^T A \Delta x}_{Df(x)[\Delta x]} + \underbrace{\Delta x^T A \Delta x}_{\bar{o}(\|\Delta x\|)}$$

$$Df(x)[\Delta x] = \underbrace{\Delta x^T A x}_{()=( )^T, \text{const}} + x^T A \Delta x = x^T A^T \Delta x + x^T A \Delta x = x^T (A + A^T) \Delta x =$$

$$\underbrace{((A + A^T)x)^T}_{\nabla f(x)} \Delta x$$

$$\boxed{\nabla(x^T A x) = (A + A^T)x}$$

**Application:** The problem of fitting weights of a linear regression can be formalised as follows:

$$\|X\beta - y\|_2^2 \rightarrow \min_{\beta}$$

The necessary condition of extremum is gradient equals zero:  $\nabla_{\beta} \|X\beta - y\|_2^2 = 0$

$$\|a\|_2^2 = a^T a, \text{ so } \nabla (X\beta - y)^T (X\beta - y) = 0$$

$$\nabla \left( \beta^T X^T X \beta \underbrace{- \beta^T X^T y - y^T X \beta}_{-2y^T X \beta} + y^T y \right) = \left( \underbrace{X^T X + X^T X}_{2X^T X} \right) \beta - 2X^T y = 0$$

$$X^T X \beta = X^T y \Rightarrow \boxed{\beta = (X^T X)^{-1} X^T y}$$

In case of  $l_2$ -regularisation:

$$\|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2 \rightarrow \min_{\beta}$$

$\nabla = 2X^T X \beta - 2y^T X \beta + \lambda (I + I^T) \beta = 0$  where  $I$  - is identity matrix

$$(X^T X + \lambda I) \beta = X^T y \Rightarrow \boxed{\beta = (X^T X + \lambda I)^{-1} X^T y}$$

**Example 3:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = e^{x^T x}$

$$Df(x)[\Delta x] = e^{x^T x} \cdot x^T (I + I^T) \Delta x = \underbrace{2 \cdot x^T \cdot e^{x^T x}}_{\nabla^T} \Delta x$$

$$\nabla (e^{x^T x}) = 2e^{x^T x} x$$

**Example 4:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = x^T e^{xx^T} x, \quad \nabla f(x) = ?$

First consider  $e^A = I + \frac{1}{1!}A + \frac{1}{2!}A^2 + \dots$  - this is by definition.

$$\begin{aligned} \text{So } x^T e^{xx^T} x &= x^T \sum_{i=0}^{+\infty} \frac{(xx^T)^i}{i!} x = \sum_{i=0}^{+\infty} \frac{x^T (xx^T)^i x}{i!} = \sum_{i=0}^{+\infty} \frac{\overbrace{x^T (x \ x^T)(xx^T) \dots (xx^T)x}^{i+1\text{-times}}}{i!} = \\ &= \sum_{i=0}^{+\infty} \frac{(x^T x)^{i+1}}{i!} = x^T x \sum_{i=0}^{+\infty} \frac{(x^T x)^i}{i!} = x^T x \exp(x^T x) \end{aligned}$$

Then take into account, that when  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  the usual differentiation rules

hold:  $\nabla(u \cdot v) = \nabla u \cdot v + u \cdot \nabla v$ , and  $\nabla\left(\frac{u}{v}\right) = \frac{\nabla u \cdot v - u \cdot \nabla v}{v^2}$

$$\begin{aligned} \nabla(x^T x \exp(x^T x)) &= \nabla(x^T x) \cdot \exp(x^T x) + x^T x \cdot \nabla(\exp(x^T x)) = \\ 2 \cdot x \cdot \exp(x^T x) + x^T x \cdot 2 \exp(x^T x) x &= 2 \exp(x^T x) (1 + x^T x) x \end{aligned}$$

$$\boxed{\nabla(x^T \exp(x x^T) x) = 2 \exp(x^T x) (1 + x^T x) x}$$

**Example 5:**  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \text{Det}(2I + x x^T)$

We will need here some properties of matrix differentiation:

$$\boxed{1} \quad h(x) = g(f(x)), \quad f: \mathbb{X} \rightarrow \mathbb{Y}, \quad g: \mathbb{Y} \rightarrow \mathbb{Z} \Rightarrow h: \mathbb{X} \rightarrow \mathbb{Z}$$

$$Dh(x)[\Delta x] = Dg(\underbrace{f(x)}_y) \underbrace{[Df(x)[\Delta x]]}_{\Delta y}$$

$$\boxed{2} \quad D(\text{Tr}(X))[\Delta X] = \text{Tr}(\Delta X)$$

$$\text{So } D(\text{Det}(2I + x x^T))[\Delta x] = \begin{cases} D(\text{Det}(Y))[\Delta Y] & (1) \\ D(2I + x x^T)[\Delta x] & (2) \end{cases}$$

$$(2) : \quad (x + \Delta x)(x^T + \Delta x^T) = x x^T + \underbrace{\Delta x x^T + x \Delta x^T}_{D(x x^T[\Delta x])} + \Delta x \Delta x^T \Rightarrow$$

$$D(2I + x x^T)[\Delta x] = x \Delta x^T + x^T \Delta x$$

$$(1) \quad D(\text{Det}(Y))[\Delta Y] = \text{Det}(Y) \text{Tr}(Y^{-1} \Delta Y) =$$

$$\text{Det}(2I + x x^T) \text{Tr}(\underbrace{(2I + x x^T)^{-1}}_{A=A^T} [x \Delta x^T + x^T \Delta x]) =$$

$$\text{Det}(2I + x x^T) [\text{Tr}(A x \Delta x^T) + \text{Tr}(A \Delta x x^T)] =$$

$$\begin{aligned} &\{ \text{we could rearrange the order inside trace and transpose, because } A = A^T \\ &= \underbrace{(2I + x x^T)}_{\text{constant}} [2 \text{Tr}(x^T A \Delta x)] = \text{Tr}(2 \text{Det}(2I + x x^T) x^T A \Delta x) = \text{Tr}(\nabla f(x)^T \Delta x) \end{aligned}$$

$$\boxed{\nabla f(x) = 2 \text{Det}(2I + x x^T) (2I + x x^T) x}$$

**Example 6:**  $f: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(x) = \log \text{Det}(x)$ ,  $\nabla f(x) = ?$

$$Df(x)[\Delta x] = \lim_{t \rightarrow +0} \frac{f(x+t \cdot \Delta x) - f(x)}{t}$$

$$D(\log \text{Det}(x))[\Delta x] = \begin{cases} D \log Y[\Delta Y] & (1) \\ \Delta Y = D \text{Det}(X)[\Delta X] & (2) \end{cases}$$

$$(1) \quad D \log Y[\Delta Y] = Y^{-1} \cdot \Delta Y = \text{Det}(X)^{-1} \cdot \text{Det}(X) \cdot \text{Tr}(X^{-1} \Delta X) = \text{Tr}(X^{-1} \cdot \Delta X)$$

$$\begin{aligned}
(2) \quad D \text{Det}(X)[\Delta X] &= \lim_{t \rightarrow +0} \frac{\text{Det}(X+t \cdot \Delta X) - \text{Det}(X)}{t} = \\
\lim_{t \rightarrow +0} \frac{\text{Det}(X \cdot [I+t \cdot X^{-1} \cdot \Delta X]) - \text{Det}(X)}{t} &= \lim_{t \rightarrow +0} \frac{\text{Det}(X) \cdot [\text{Det}(I+t \cdot X^{-1} \cdot \Delta X) - 1]}{t} = \\
\lim_{t \rightarrow +0} \frac{\text{Det}(X) \cdot [\exp(\log(\text{Det}(I+t \cdot X^{-1} \cdot \Delta X))) - 1]}{t} &= \lim_{t \rightarrow +0} \frac{\text{Det}(X) \cdot \log(\text{Det}(I+t \cdot X^{-1} \cdot \Delta X))}{t} = \\
\{\text{for small } \epsilon \text{ holds : } \text{Det}(I + \epsilon \cdot A) \approx 1 + \epsilon \cdot \text{Tr}(A) + \bar{o}(\epsilon)\} &= \\
\lim_{t \rightarrow +0} \frac{\text{Det}(X) \cdot \log(1+t \cdot \text{Tr}(X^{-1} \Delta X))}{t} &= \lim_{t \rightarrow +0} \frac{\text{Det}(X) \cdot t \cdot \text{Tr}(X^{-1} \Delta X)}{t} = \boxed{\text{Det}(X) \cdot \text{Tr}(X^{-1} \Delta X)} \\
D(\log \text{Det}(X))[\Delta X] &= \text{Tr}(X^{-1} \Delta X) = \text{Tr}(\nabla f(X)^T \Delta X) \Rightarrow \\
\boxed{\nabla(\log \text{Det}(X)) = X^{-T}}
\end{aligned}$$

**P.S:**  $\boxed{\nabla_X \log \text{Det}(X^{-1}) = -\nabla_X \log \text{Det}(X) = -X^{-T}}$

**Example 7:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}, \quad f(X) = a^T X a$

$$f(X + \Delta X) = \underbrace{a^T X a}_{f(X)} + \underbrace{a^T \Delta X a}_{Df(X)[\Delta X]}$$

$$Df(X)[\Delta X] = \text{Tr}(a^T \Delta X a) = \text{Tr}(a a^T \Delta X) = \text{Tr}(\nabla f(X)^T \Delta X)$$

$$\boxed{\nabla_X(a^T X a) = a a^T}$$

**Application:** Let's consider the problem of finding the maximum likelihood estimations for the multidimensional Normal distribution:  $x_i \sim \mathcal{N}(\mu, \Sigma)$

$$p(x | \mu, \Sigma) = \frac{1}{\sqrt{\text{Det}(2\pi\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu))$$

The Likelihood function will look like that:  $L(\mu, \Sigma) = \prod_{i=1}^n \frac{1}{\sqrt{\text{Det}(2\pi\Sigma)}} \exp(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)) = \text{Det}(2\pi\Sigma)^{-\frac{n}{2}} \exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)) \rightarrow \max_{\mu, \Sigma}$

$$\log L = -\frac{n^2}{2} \log \text{Det}(2\pi) - \frac{n}{2} \log \text{Det}(\Sigma) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \rightarrow \max_{\mu, \Sigma}$$

$$\nabla_{\mu} \log L = -\frac{1}{2} \sum_{i=1}^n 2 \Sigma^{-1}(x_i - \mu) = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i}$$

$$\begin{aligned}
\nabla_{\Sigma} \log L &= \{\text{for convenience let } \Lambda = \Sigma^{-1}\} = \\
\nabla_{\Lambda} \left( -\frac{n}{2} \log \text{Det} \Lambda^{-1} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Lambda (x_i - \mu) \right) &= \frac{n}{2} \underbrace{\Lambda^{-T}}_{\Lambda^{-1}} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T
\end{aligned}$$

$$\mu)^T = 0 \Rightarrow \hat{\Lambda}^{-1} = \boxed{\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T}$$

**Application:** Find the maximum of a function  $f(X) = \text{Det}(X)^{-1} \exp(-\frac{1}{2} \text{Tr}(X^{-1} \cdot A))$

Necessary condition -  $\nabla_X f(X) = 0$ :

First let's consider, that the  $\text{argmax}_X f(X) = \text{argmax}_X \log f(X)$

$$\log f(X) = -\log \text{Det}(X) - \frac{1}{2} \text{Tr}(X^{-1} A)$$

$$D \text{Tr}(-\frac{1}{2} X^{-1} A) = \text{Tr}(\frac{1}{2} X^{-1} \Delta X X^{-1} A) = \text{Tr}(\frac{1}{2} X^{-1} A X^{-1} \Delta X) \Rightarrow \nabla_X = \frac{1}{2} X^{-T} A^{-T} X^{-T}$$

$$\nabla_X f(X) = -X^{-T} + \frac{1}{2} X^{-T} A^{-T} X^{-T} = -X^{-T} (I - \frac{1}{2} A^{-T} X^{-T}) = 0$$

$$A^{-T} X^{-T} = 2I \Rightarrow X^T A^T = \frac{1}{2} I \Rightarrow \boxed{X^* = \frac{1}{2} A^{-1}}$$

**Example 8:**  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \|x\|_2^3$ ,  $\nabla f(x) = ?$

$$f(x) = (x^T x)^{3/2} \Rightarrow \nabla f(x) = \frac{3}{2} (x^T x)^{1/2} 2x = 3(x^T x)^{1/2} x$$

**Example 9:**  $f: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ ,  $f(X) = X^{-1}$ ,  $Df(X)[\Delta X] = ?$

$$\begin{aligned} D(X^{-1})[\Delta X] &= \lim_{t \rightarrow +0} \frac{(X+t \cdot \Delta X)^{-1} - X^{-1}}{t} = \lim_{t \rightarrow +0} \frac{(X \cdot [I + X^{-1} \cdot t \cdot \Delta X])^{-1} - X^{-1}}{t} = \\ \{(AB)^{-1} &= B^{-1} A^{-1}\} = \lim_{t \rightarrow +0} \frac{(I + X^{-1} \cdot t \cdot \Delta X)^{-1} X^{-1} - X^{-1}}{t} = \\ \lim_{t \rightarrow +0} \frac{[(I + X^{-1} \cdot t \cdot \Delta X)^{-1} - I] \cdot X^{-1}}{t} &= \{(I + \epsilon \cdot A)^b - I \approx \epsilon \cdot b \cdot A\} = \lim_{t \rightarrow +0} \frac{-X^{-1} t \Delta X X^{-1}}{t} = \\ &= -X^{-1} \Delta X X^{-1} \end{aligned}$$

$$\boxed{D(X^{-1})[\Delta X] = -X^{-1} \Delta X X^{-1}}$$

**Example 10:**  $f: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \text{Tr}(X)$ ,  $\nabla f(X) = ?$

$$\begin{aligned} D(\text{Tr}(X))[\Delta X] &= \lim_{t \rightarrow +0} \frac{\text{Tr}(X+t \cdot \Delta X) - \text{Tr}(X)}{t} = \lim_{t \rightarrow +0} \frac{\text{Tr}(X+t \cdot \Delta X - X)}{t} = \\ \lim_{t \rightarrow +0} \frac{t \cdot \text{Tr}(\Delta X)}{t} &= \text{Tr}(I \cdot \Delta X) = \text{Tr}(\nabla f(X)^T \Delta X) \Rightarrow \boxed{\nabla_X \text{Tr}(X) = I} \end{aligned}$$

**Example 11:**  $f: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \text{Tr}(AX^{-1}B)$ ,  $\nabla f(X) = ?$

$$DTr(AX^{-1}B)[\Delta X] = \begin{cases} DTr(Y)[\Delta Y] = Tr(\Delta Y) \\ D(AX^{-1}B)[\Delta X], \end{cases} \quad (1)$$

$$(1) \quad D(AX^{-1}B)[\Delta X] = \begin{cases} D(AZB)[\Delta Z], \\ \Delta Z = D(X^{-1})[\Delta X] = -X^{-1}\Delta XX^{-1} \end{cases} \quad (2)$$

$$f(Z + \Delta Z) = \underbrace{AZB}_{f(Z)} + \underbrace{A\Delta ZB}_{Df(Z)[\Delta Z]}$$

$$\text{So } DTr(AX^{-1}B)[\Delta X] = Tr(-AX^{-1}\Delta XX^{-1}B) = Tr(\underbrace{-X^{-1}BAX^{-1}}_{\nabla f(X)^T} \Delta X)$$

$$\nabla_X f(X) = (-X^{-1}BAX^{-1})^T = -X^{-T}A^T B^T X^{-T}$$

**Example 12 :**  $f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ ,  $f(x) = xx^T$ ,  $Df(x)[\Delta x] = ?$

$$f(x + \Delta x) = (x + \Delta x)(x + \Delta x)^T = \underbrace{xx^T}_{f(x)} + \underbrace{x\Delta x^T + \Delta x x^T}_{Df(x)[\Delta x]} + \underbrace{\Delta x \Delta x^T}_{o(\|\Delta x\|)}$$

$$Df(x)[\Delta x] = x\Delta x^T + (\underbrace{x\Delta x^T}_{\text{symmetric}})^T = 2x\Delta x^T$$

**Example 13:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \frac{x^T A x}{x^T x}$ ,  $\nabla f(x) = ?$

$$\begin{aligned} \nabla f(x) &= \nabla \left( \frac{g(x)}{h(x)} \right) = \frac{\nabla g(x) \cdot h(x) - g(x) \cdot \nabla h(x)}{h^2(x)} = \frac{(A + A^T)x \cdot x^T x - 2x^T A x \cdot x}{(x^T x)^2} = \\ &= \frac{(A + A^T)x^T x - 2x^T A x \cdot I}{(x^T x)^2} \cdot x \end{aligned}$$

**Example 14:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = Det(AXB)$ ,  $\nabla f(X) = ?$

$$D(Det(AXB))[\Delta X] = \begin{cases} DDet(Y)[\Delta Y] = Det(Y)Tr(Y^{-1}\Delta Y) \\ \Delta Y = D(AXB)[\Delta X] = A\Delta XB \end{cases}$$

$$Df(X)[\Delta X] = \underbrace{Det(AXB)Tr((AXB)^{-1}A\Delta XB)}_{const} =$$

$$\begin{aligned} Det(AXB)Tr(B^{-1}X^{-1}A^{-1}A\Delta XB) &= Tr(Det(AXB)X^{-1}\Delta X) = \\ Tr(\nabla f(X)^T \Delta X) & \end{aligned}$$

$$\nabla_X Det(AXB) = Det(AXB)X^{-T}$$

**Example 15:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = Tr(AX^{-1})Tr(XB)$ ,  $\nabla f(X) = ?$

Let's apply the property of matrix differentiation:

$h(x) = g(x) \cdot f(x)$ ,  $f : \mathbb{X} \rightarrow \mathbb{Y}$ ,  $g : \mathbb{X} \rightarrow \mathbb{R}$ , then:

$$Dh(x)[\Delta x] = Dg(x)[\Delta x] \cdot f(x) + g(x) \cdot Df(x)[\Delta x]$$

$$D \left( Tr(AX^{-1})Tr(XB) \right) [\Delta X] = Tr(XB) \cdot \underbrace{DTr(AX^{-1})[\Delta X]}_{(1)} + Tr(AX^{-1}) \cdot \underbrace{DTr(XB)[\Delta X]}_{(2)}$$

$$(1) \quad DTr(AX^{-1})[\Delta X] = \begin{cases} DTr(Z)[\Delta Z] = Tr(\Delta Z) \\ \Delta Z = D(AY)[\Delta Y] = A\Delta Y \\ \Delta Y = D(X^{-1})[\Delta X] = -X^{-1}\Delta XX^{-1} \end{cases}$$

$$DTr(AX^{-1})[\Delta X] = Tr(-AX^{-1}\Delta XX^{-1})$$

$$(2) \quad DTr(XB) = Tr(\Delta XB), \text{ so}$$

$$Df(X)[\Delta X] = \underbrace{Tr(XB)}_{const} \cdot Tr(-AX^{-1}\Delta XX^{-1}) + \underbrace{Tr(AX^{-1})}_{const} \cdot Tr(\Delta XB) = Tr(-Tr(XB)X^{-1}AX^{-1}\Delta X + Tr(AX^{-1}B\Delta X)) = Tr([-Tr(XB)X^{-1}AX^{-1} + Tr(AX^{-1})B] \cdot \Delta X) = Tr(\nabla f(X)^T \Delta X)$$

$$\nabla f(X) = Tr(AX^{-1})B - Tr(XB)X^{-1}AX^{-1}$$

**Example 16:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = Det(\exp(X))$ ,  $\nabla f(X) = ?$

$$DDet(\exp(X))[\Delta X] = \begin{cases} DDet(Y)[\Delta Y] = Det(Y)Tr(Y^{-1}\Delta Y) \\ \Delta Y = D\exp(X)[\Delta X] = \exp(X)\Delta X - \text{prove it !} \end{cases}$$

$$DDet(\exp(X))[\Delta X] = Det(\exp(X))Tr(\exp(X)^{-1}\exp(X)\Delta X) = Tr(Det(\exp(X))\Delta X) = Tr(\nabla f(X)^T \Delta X)$$

$$\boxed{\nabla_X (Det(\exp(X))) = Det(\exp(X)) \cdot I}$$

**Example 17:**  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \frac{1}{2}\|X - A\|_F^2$ ,  $\nabla f(X) = ?$

$$\boxed{\|A\|_F^2 = \sum_{i,j} a_{ij}^2 = Tr(A^T A)}, \text{ so } f(X) = \frac{1}{2}Tr((X - A)^T(X - A)) = \frac{1}{2}Tr(X^T X - X^T A - A^T X + A^T A)$$

$$Df(X)[\Delta X] = \frac{1}{2}Tr(X^T \Delta X + \Delta X^T X - \Delta X^T A - A^T \Delta X) = Tr(X^T \Delta X) + Tr(-A^T \Delta X) = Tr((X^T - A^T) \cdot \Delta X) = Tr(\nabla f(X)^T \Delta X)$$

$$\nabla f(X) = X - A$$

**Example 18:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \text{Tr}(Axx^T)$ ,  $\nabla f(x) = ?$

$$\begin{aligned} Df(x)[\Delta x] &= \text{Tr}(D(Axx^T)[\Delta x]) = \text{Tr}(Ax\Delta x^T + A\Delta xx^T) = \\ &= \text{Tr}([x\Delta x^T]^T A^T) + \text{Tr}(x^T A\Delta x) = \text{Tr}(x^T A^T \Delta x + x^T A\Delta x) = \text{Tr}(\nabla f(x)^T \Delta x) \end{aligned}$$

$$\nabla f(x) = x^T (A + A^T)$$

**Example 19:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \frac{1}{2}\|xx^T - A\|_F^2$ ,  $\nabla f(x) = ?$

$$f(x) = \frac{1}{2}\text{Tr}((xx^T - A)^T(xx^T - A)) = \frac{1}{2}\text{Tr}(xx^T xx^T - xx^T A - Axx^T - A^T A)$$

$$Df(x)[\Delta x] = \frac{1}{2}\text{Tr}(4x^T xx^T \Delta x - x^T A\Delta x - x^T A^T \Delta x)$$

$$\begin{aligned} \nabla f(x) &= [2x^T xx^T - \frac{1}{2}x^T(A + A^T)]^T = [x^T(2xx^T - \frac{1}{2}A - \frac{1}{2}A^T)]^T = \\ &= (2xx^T - \frac{1}{2}(A + A^T))x \end{aligned}$$

**Example 20:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \frac{1}{2}(x^T x + (a^T x)^2)$ ,  $\nabla f(x) = ?$

$$\nabla f(x) = x + a^T x a$$

**Example 21:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = x^T \vec{\log}(x)$ ,  $\nabla f(x) = ?$ , where  $\vec{\log}(x) = (\log x_1, \dots, \log x_n)$

$$f(x) = x^T \log(x) = \sum_{i=1}^n x_i \cdot \log(x_i), \quad \nabla f(x) = \left( \frac{\partial f(x_1)}{\partial x_1}, \dots, \frac{\partial f(x_n)}{\partial x_n} \right)$$

$$\frac{\partial f(x_i)}{\partial x_i} = \log(x_i) + 1 \Rightarrow \nabla f(x) = \log(x) + \vec{1}$$

**Example 22:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \text{Det}(A + X^{-1})$ ,  $\nabla f(X) = ?$

$$DDet(A + X^{-1})[\Delta X] = \begin{cases} DDet(Y)[\Delta Y] = Det(Y)Tr(Y^{-1}\Delta Y) \\ \Delta Y = D(A + X^{-1})[\Delta X] = -X^{-1}\Delta X X^{-1} \end{cases}$$

$$\begin{aligned} Df(X)[\Delta X] &= Det(A + X^{-1})Tr(-(A + X^{-1})^{-1}X^{-1}\Delta X X^{-1}) = \\ &= Tr(-Det(A + X^{-1})X^{-1}(A + X^{-1})^{-1}X^{-1}\Delta X) = Tr(\nabla f(X)^T \Delta X) \end{aligned}$$

$$\nabla f(X) = -Det(A + X^{-1})X^{-T}(A + X^{-1})^{-T}X^{-T}$$

**Example 23:** (The problem of constructing surrogate eigenvector)



The problem comes from the fact, that the linear system  $Ax = b$  can equivalently be rewritten as  $f(x) = \frac{1}{2}x^T Ax - b^T x \rightarrow \min_x$ , where equivalence means that the solution of both problems would be the same.

$f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \frac{1}{2}x^T (A - \lambda \cdot I) x \rightarrow \min_x$ , where  $\lambda$  is an eigenvalue of  $A$ ,  $A = A^T > 0$

$$f(x) \rightarrow \min_x \Rightarrow \nabla_x f(x) = 0$$

$\nabla f(x) = (A - \lambda \cdot I)x = 0$ , from this follows, that a minimiser of this function is an eigenvector of a matrix  $A$ , corresponding to an eigenvalue  $\lambda$ , but if we can't find an eigenvector, we can use this gradient to find it iteratively and numerically:

Initialisation:  $x_0$  – some random vector,

$$x_{k+1} = x_k - \alpha_k \cdot \nabla f(x_k) = ((1 + \alpha_k \cdot \lambda) \cdot I - \alpha_k \cdot A) x_k = \{\text{if } \alpha_k \text{ is constant, then } \} = ((1 + \alpha \cdot \lambda) \cdot I - \alpha \cdot A)^{k+1} \cdot x_0$$

**Application:** (Logistic Regression fitting)

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) = -\sum_{i=1}^n \log(\sigma(w^T x_i y_i)) \rightarrow \min_w, \text{ where } \sigma(z) = \frac{1}{1+\exp(-z)}, \quad x_i, w \in \mathbb{R}^n, \quad y_i \in \mathbb{R}$$

$$\sigma(z)'_z = \frac{\exp(-z)}{(1+\exp(-z))^2} = \frac{1}{(1+\exp(z))^2} = \sigma^2(-z)$$

$$\nabla_x f(x) = -\sum_{i=1}^n \frac{\sigma^2(-w^T x_i y_i)}{\sigma(w^T x_i y_i)} y_i x_i$$

**Example 24** Find minimum of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = \frac{x^T A x}{x^T B x}$ , where  $A \geq 0, B \geq 0$

Necessary condition  $\nabla f(x) = 0$

$$\nabla f(x) = \frac{(A+A^T)x \cdot x^T B x - x^T A x \cdot (B+B^T)x}{(x^T B x)^2}$$

**Example 25:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = a^T X^{-1} a$ ,  $\nabla f(X) = ?$

$$D(a^T X^{-1} a) [\Delta X] = \text{Tr}(a a^T D X^{-1} [\Delta X]) = \text{Tr}(-a a^T X^{-1} \Delta X X^{-1}) = \text{Tr}(-X^{-1} a a^T X^{-1} \Delta X) = \text{Tr}(\nabla f(X)^T \Delta X) \Rightarrow \nabla f(X) = -X^{-T} a a^T X^{-T}$$

**Example 26:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \text{Det}(X^2)$ ,  $\nabla f(X) = ?$

There are 2 ways of dealing with it, let's compare these ways:

$$\begin{aligned} 1) \quad & \text{Det}(X^2) = \text{Det}(X)^2 \Rightarrow D\text{Det}(X)^2[\Delta X] = \\ & \left\{ \begin{aligned} DY^2[\Delta Y] &= Y\Delta Y + \Delta Y Y \quad (1) \\ \Delta Y &= D\text{Det}(X)[\Delta X] = \text{Det}(X)\text{Tr}(X^{-1}\Delta X) \end{aligned} \right. \\ & (1) \quad Df(X)[\Delta X] = 2\text{Det}(X) \cdot \text{Det}(X)\text{Tr}(X^{-1}\Delta X) \Rightarrow \nabla f(X) = \\ & 2\text{Det}(X)^2 X^{-T} \end{aligned}$$

$$2) \text{Det}(X^2), \quad D\text{Det}(X^2)[\Delta X] = \left\{ \begin{aligned} D\text{Det}(Y)[\Delta Y] &= \text{Det}(Y)\text{Tr}(Y^{-1}\Delta Y) \\ \Delta Y &= D(X^2)[\Delta X] = X\Delta X + \Delta X X \end{aligned} \right.$$

$Df(X)[\Delta X] = \text{Det}(X^2)\text{Tr}(2X^{-1}\Delta X) = 2\text{Det}(X)^2\text{Tr}(X^{-1}\Delta X)$ , which is indeed the same result as in the first way

**Example 27:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \text{Tr}(AX^{-1})$ ,  $\nabla f(X) = ?$

$$Df(X)[\Delta X] = \left\{ \begin{aligned} D\text{Tr}(AY)[\Delta Y] &= \text{Tr}(A\Delta Y) \\ \Delta Y &= D(X^{-1})[\Delta X] = -X^{-1}\Delta X X^{-1} \end{aligned} \right.$$

$$D\text{Tr}(AX)[\Delta X] = \text{Tr}(-AX^{-1}\Delta X X^{-1}) = \text{Tr}(-X^{-1}AX^{-1}\Delta X)$$

$$\nabla f(X) = -X^{-T}A^{-T}X^{-T}$$

**Application:** Kernel Linear Regression Problem

$$Q(a) = \frac{1}{2}\|\Phi\Phi^T a - y\|_2^2 + \frac{\lambda}{2}a^T\Phi\Phi^T a \rightarrow \min_a, \quad K = \Phi\Phi^T = K^T > 0$$

$$\nabla_a Q = (\Phi\Phi^T)^2 a - \Phi\Phi^T y + \lambda\Phi\Phi^T a = 0 \mid \cdot K^{-1} \Rightarrow Ka - y + \lambda a = 0$$

$$(K + \lambda I)a = y \Rightarrow \boxed{a^* = (K + \lambda I)^{-1}y}$$

**Example 28:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(X) = \log \text{Det}(X) + \text{Tr}(X^{-1}A) \rightarrow \min_X$

$$D \log \text{Det}(X)[\Delta X] = \text{Tr}(X^{-1}\Delta X) \Rightarrow \nabla \log \text{Det}(X) = X^{-T}$$

$$\begin{aligned} D\text{Tr}(X^{-1}A)[\Delta X] &= \text{Tr}(-X^{-1}\Delta X X^{-1}A) = \text{Tr}(-X^{-1}AX^{-1}\Delta X) \Rightarrow \\ \nabla \text{Tr}(X^{-1}A) &= -X^{-T}A^T X^{-T} \end{aligned}$$

$$\nabla f(X) = X^{-T} - X^{-T} A^T X^{-T} = 0 \Rightarrow X^{-T} (I - A^T X^{-T}) = 0$$

$$X^{-1} A = I \Rightarrow X = A$$

**Application:** Robust linear regression

$$\sum_{i=1}^n w_i (< x_i, \beta > - y_i)^2 = (X\beta - y)^T W (X\beta - y)$$

$$Q(\beta) = (X\beta - y)^T W (X\beta - y) \rightarrow \min_{\beta}$$

$$Q(\beta) = \beta^T X^T W X \beta - 2y^T W X \beta + y^T W y$$

$$\nabla_{\beta} Q = 2X^T W X \beta - 2X^T W y = 0$$

$$\boxed{\beta^* = (X^T W X)^{-1} X^T W y} \quad \text{where } W \text{ is a weight matrix}$$

## Higher order derivatives

When  $f : \mathbb{R}^n \rightarrow \mathbb{R} \Rightarrow \boxed{D^2 f(x)[\Delta x_1, \Delta x_2] = \Delta x_1^T H(x) \Delta x_2}$ , where  $H(x)$  – *Hessian*

$$D^k f(x)[\Delta x_1, \dots, \Delta x_k] = \left. \frac{\partial^k}{\partial t_1 \partial t_2 \dots \partial t_k} \right|_{t_1 = \dots = t_k = 0} f(x + t_1 \cdot \Delta x_1 + \dots + t_k \cdot \Delta x_k)$$

**Example 1:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = x^T A x$ ,  $D^2 f(x)[\Delta x_1, \Delta x_2] = ?$ ,  $H(x) = ?$

$$Df(x)[\Delta x_1] = 2x^T A \Delta x_1$$

$$\begin{aligned} D^2 f(x)[\Delta x_1, \Delta x_2] &= \lim_{t \rightarrow +0} \frac{Df(x+t \cdot \Delta x_2)[\Delta x_1] - Df(x)[\Delta x_1]}{t} = \\ \lim_{t \rightarrow +0} \frac{2(x^T + t \cdot \Delta x_2^T) A \Delta x_1 - 2x^T A \Delta x_1}{t} &= \boxed{2\Delta x_2^T A \Delta x_1} \Rightarrow \boxed{H(x) = 2A} \end{aligned}$$

**Example 2:**  $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ ,  $f(X) = \log \text{Det}(X)$ ,  $D^2 f(X)[\Delta X_1, \Delta X_2] = ?$

$$Df(X)[\Delta X_1] = \text{Tr}(X^{-1} \Delta X_1)$$

$$\begin{aligned} D^2 f(X)[\Delta X_1, \Delta X_2] &= \lim_{t \rightarrow +0} \frac{Df(X+t \cdot \Delta X_2)[\Delta X_1] - Df(X)[\Delta X_1]}{t} = \\ \lim_{t \rightarrow +0} \frac{\text{Tr}((X+t \cdot \Delta X_2)^{-1} [\Delta X_1]) - \text{Tr}(X^{-1} [\Delta X_1])}{t} &= \lim_{t \rightarrow +0} \frac{\text{Tr}([(X+t \cdot \Delta X_2)^{-1} - X^{-1}] [\Delta X_1])}{t} = \end{aligned}$$

$$\begin{aligned} \lim_{t \rightarrow +0} \frac{\text{Tr}([X(I+t \cdot X^{-1} \Delta X_2)]^{-1} - X^{-1})[\Delta X_1])}{t} &= \lim_{t \rightarrow +0} \frac{\text{Tr}([(I+t \cdot X^{-1} \Delta X_2)^{-1} - I]X^{-1} \Delta X_1)}{t} = \\ \lim_{t \rightarrow +0} \frac{\text{Tr}(-t \cdot X^{-1} \Delta X_2 X^{-1} \Delta X_1)}{t} &= \text{Tr}(-\Delta X_1 X^{-1} \Delta X_2 X^{-1}) \end{aligned}$$

**Example 3:**  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(\beta) = \|X\beta - y\|_2^2$ ,  $H(\beta) = ?$

$$Df(\beta)[\Delta\beta_1] = (2X^T X\beta - 2X^T y)^T [\Delta\beta_1]$$

$$\begin{aligned} D^2 f(\beta)[\Delta\beta_1, \Delta\beta_2] &= \lim_{t \rightarrow +0} \frac{[2(\beta^T + t \cdot \Delta\beta_2^T)X^T X - 2y^T X - 2\beta^T X^T X + 2y^T X][\Delta\beta_1]}{t} = \\ \Delta\beta_2^T 2X^T X \Delta\beta_1 &= \beta_1^T 2X^T X \Delta\beta_2 \Rightarrow \boxed{H(\beta) = 2X^T X} \end{aligned}$$

Which we could have attained easier:  $\nabla_\beta \nabla_\beta f(\beta) = \nabla_\beta (2X^T X\beta) = 2X^T X$

## Constraint optimisation

**Example 1:** 
$$\begin{cases} \|x\|_2^2 \rightarrow \min_x \\ \text{s.t. } Ax \leq b \end{cases}$$

$$L = x^T x + \lambda^T (Ax - b)$$

K.K.T conditions:

$$\begin{cases} \nabla_x L = 2x + A^T \lambda = 0 \\ Ax - b \leq 0 \\ \lambda^T (Ax - b) = 0 \\ \lambda \geq 0 \end{cases} \Rightarrow \begin{cases} \lambda^T (-\frac{1}{2}AA^T \lambda - b) = 0 \Rightarrow \lambda^* = -2(AA^T)^{-1}b \\ x^* = -\frac{1}{2}A^T \lambda \end{cases}$$

$$\boxed{x^* = A^T (AA^T)^{-1} b}$$

**Application:** OLS under constraints:

$$\begin{cases} \|X\beta - y\|_2^2 \rightarrow \min_\beta \\ \text{s.t. } C\beta = d \end{cases}$$

$$L = \beta^T X^T X \beta - 2y^T X \beta + y^T y + \mu^T (C\beta - d)$$

K.K.T. conditions:

$$\begin{cases} \nabla_\beta L = 2X^T X \beta - 2X^T y + C^T \mu = 0 \\ C\beta - d = 0 \end{cases} \Rightarrow \begin{cases} \beta = (X^T X)^{-1} (X^T y - \frac{1}{2} C^T \mu) \\ C\beta = d \end{cases}$$

$$C\beta = d \Rightarrow C(X^T X)^{-1} X^T y - \frac{1}{2} C(X^T X)^{-1} C^T \mu = d$$

$$\frac{1}{2} C(X^T X)^{-1} C^T \mu = C(X^T X)^{-1} X^T y - d, \quad \mu = 2 \cdot (C(X^T X)^{-1} C^T)^{-1} (C(X^T X)^{-1} X^T y - d)$$

$$\boxed{\beta^* = \underbrace{(X^T X)^{-1} X^T y}_{\beta^{\text{ols}}} - (X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} \left( C \underbrace{(X^T X)^{-1} X^T y}_{\beta^{\text{ols}}} - d \right)}$$

Let's derive  $V(\beta^*)$ :

$$\begin{aligned} V(\beta^*) &= \sigma^2 (X^T X)^{-1} - \sigma^2 \cdot (X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} \\ &\cdot ((X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} C)^T = \sigma^2 (X^T X)^{-1} - \sigma^2 \cdot \\ &(X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} C^T (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} = \\ &\sigma^2 (X^T X)^{-1} - \sigma^2 (X^T X)^{-1} (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} = \\ &\boxed{V(\beta^*) = \sigma^2 (X^T X)^{-1} \left( I - (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} \right)} \end{aligned}$$

When  $C$  is invertable this could be reduced to:

$$\begin{aligned} &\sigma^2 (X^T X)^{-1} \left( I - (C(X^T X)^{-1} C^T)^{-1} C(X^T X)^{-1} C^T C^{-T} \right) = \\ &\boxed{V(\beta^*) = \sigma^2 (X^T X)^{-1} (I - C^{-T})} \end{aligned}$$

**Example 2:**  $\begin{cases} a^T x \rightarrow \min_x \\ \text{s.t. } x^T A x \leq 1 \end{cases}, \text{ where } A = A^T > 0$

$$L = a^T x + \lambda(x^T A x - 1)$$

K.K.T. conditions:

$$\begin{cases} \nabla_x L = a + 2\lambda A x = 0 \\ x^T A x - 1 \leq 0 \\ \lambda(x^T A x - 1) = 0 \\ \lambda \geq 0 \end{cases} \Rightarrow \begin{cases} x^* = -\frac{1}{2\lambda} A^{-1} a \\ \lambda \left( (-\frac{1}{2\lambda} A^{-1} a)^T A (-\frac{1}{2\lambda} A^{-1} a) - 1 \right) = 0 \end{cases} \quad (1)$$

$$(1) \frac{1}{4\lambda} a^T A^{-1} a - \lambda = 0 \mid \cdot \lambda \Rightarrow \lambda^* = \pm \frac{1}{2} \sqrt{a^T A^{-1} a}, \quad \lambda \geq 0 \Rightarrow \lambda^* = \frac{1}{2} \sqrt{a^T A^{-1} a}$$

$$\boxed{x^* = -\frac{A^{-1} a}{\sqrt{a^T A^{-1} a}}}$$

**Example 3:**  $\begin{cases} x^T Q x \rightarrow \min_x \\ \text{s.t. } \|Ax - b\|_2^2 \leq 1 \end{cases}, \text{ where } Q = Q^T > 0, A = A^T > 0$

$$L = x^T Q x + \lambda ((Ax - b)^T (Ax - b) - 1) = x^T Q x + \lambda (x^T A^T A x - 2b^T A x + b^T b - 1)$$

K.K.T. conditions:

$$\begin{cases} \nabla_x L = 2Qx + 2\lambda A^T A x - 2\lambda A^T b = 0 \\ \lambda \geq 0 \\ \lambda (x^T A^T A x - 2b^T A x + b^T b - 1) = 0 \\ \|Ax - b\|_2^2 \leq 1 \end{cases} \Rightarrow$$

$$\begin{cases} \boxed{x^* = \lambda (Q + \lambda A^T A)^{-1} A^T b} \\ \lambda \geq 0 \\ \lambda (x^T A^T A x - 2b^T A x + b^T b - 1) = 0 \\ \|Ax - b\|_2^2 \leq 1 \end{cases}$$

$\lambda^*$  one can find from the dual function, which will look a bit complex here

**Application:** (2-rank update)

$$\begin{cases} \|B - B_k\|_F^2 \rightarrow \min_B \\ \text{s.t. } B s_k = y_k \end{cases}$$

$$L = \text{Tr}(B^T B) + \mu^T (B s_k - y_k)$$

K.K.T. conditions:

$$\begin{cases} \nabla_B L = 2B + \mu s_k^T = 0 \\ B s_k = y_k \end{cases} \Rightarrow B = -\frac{1}{2} \mu s_k^T$$

Dual function  $q(\mu) = \text{Tr}(\frac{1}{4} s_k \mu^T \mu s_k^T) - \frac{1}{2} \mu^T \mu s_k^T s_k - \mu^T y_k \rightarrow \min_{\mu}$

$$\nabla_{\mu} q(\mu) = \frac{1}{2} s_k^T s_k \mu - s_k^T s_k \mu - y_k = 0, \quad \mu^* = \frac{-2y_k}{s_k^T s_k}$$

So  $\boxed{B^* = \frac{y_k s_k^T}{s_k^T s_k}}$

**Example 4:**

$$\begin{cases} \frac{1}{2} x^T A x + b^T x + c \rightarrow \min_x \\ \text{s.t. } x^T x \leq 1 \end{cases} \text{ where } A = A^T > 0$$

$$L = \frac{1}{2} x^T A x + b^T x + c + \lambda (x^T x - 1)$$

K.K.T. conditions:

$$\left\{ \begin{array}{l} \nabla_x L = Ax + b + \lambda x = 0 \\ x^T x - 1 \leq 0 \\ \lambda(x^T x - 1) = 0 \\ \lambda \geq 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \lambda b^T (A + \lambda \cdot I)^{-T} (A + \lambda \cdot I)^{-1} b - \lambda = 0, \end{array} \right. \quad \boxed{x^* = -(A + I \cdot \lambda)^{-1} b} \quad (\star)$$

$$(\star) \lambda(b^T (A + \lambda \cdot I)^{-2} b - 1) = 0 \Rightarrow \lambda^* = \max\{0, \text{solution of } (\star)\}$$

**Example 5:** (Projection on the radius 1 Ball)

$$\left\{ \begin{array}{l} \frac{1}{2} \|x - v\|_2^2 \rightarrow \min_x \\ \text{s.t. } x^T x \leq 1 \end{array} \right.$$

$$L = \frac{1}{2} (x^T x - 2v^T x + v^T v) + \lambda(x^T x - 1)$$

K.K.T. conditions:

$$\left\{ \begin{array}{l} \nabla_x L = x - v + 2\lambda x = 0 \\ \lambda \geq 0 \\ \lambda(x^T x - 1) = 0 \\ x^T x - 1 \leq 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} x^* = \frac{v}{1+2\lambda} \\ \lambda \left( \frac{v^T v}{(1+2\lambda)^2} - 1 \right) = 0 \end{array} \right. \quad (\star)$$

$$(\star) 1 + 2\lambda = \sqrt{v^T v} = \|v\|_2$$

So  $\boxed{x^* = \frac{v}{\|v\|_2}}$ , which is quite intuitive

**Application:** (Hard margin SVM)

$$\left\{ \begin{array}{l} \frac{1}{2} \|w\|_2^2 \rightarrow \min_w \\ y_i (x_i^T w - b) \geq 1 \end{array} \right.$$

$$L = \frac{1}{2} w^T w - \sum_{i=1}^n \lambda_i (y_i (x_i^T w - b) - 1)$$

K.K.T. conditions:

$$\left\{ \begin{array}{l} \nabla_w L = w - \sum_{i=1}^n \lambda_i y_i x_i = 0 \\ \nabla_b L = \sum_{i=1}^n \lambda_i y_i = 0 \\ \lambda_i \geq 0, \forall i = 1, \dots, n \\ \lambda_i (y_i (x_i^T w - b) - 1) = 0, \forall i = 1, \dots, n \end{array} \right.$$

$$\boxed{w^* = \sum_{i=1}^n \lambda_i y_i x_i}$$

**Application:** Consider the multiple regression model  $y = X \cdot \beta + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ , let's propose a linear estimator of the form  $\hat{\beta} = L \cdot y$ . Find the

unbiased estimator  $\hat{\beta}$ , for which  $Tr(Var(\hat{\beta})) \rightarrow \min_{\beta}$

$$\begin{cases} Tr(Var(Ly)) \rightarrow \min_L \\ \mathbb{E}(\hat{\beta}) = \beta \end{cases}$$

$$\mathcal{L} = Tr(Var(L(X\beta + \epsilon))) + \mu^T (\mathbb{E}(\hat{\beta}) - \beta) = Tr(Var(L\epsilon)) + \mu^T (LX\beta - \beta) = Tr(\sigma^2 LL^T) + \mu^T (LX\beta - \beta)$$

$$DTr(LL^T)[\Delta L] = Tr(L\Delta L^T + L^T\Delta L) = Tr(\nabla f(L)^T \Delta L) \Rightarrow \nabla_L = 2L$$

$$D\mu^T LX\beta[\Delta L] = \mu^T \Delta LX\beta = Tr(X\beta\mu^T \Delta L) \Rightarrow \nabla_L = \mu(X\beta)^T = \mu\beta^T X^T$$

$$\text{K.K.T. conditions: } \begin{cases} \nabla_L \mathcal{L} = 2\sigma^2 L + \mu\beta^T X^T = 0 \\ LX\beta = \beta \end{cases} \Rightarrow$$

$$\begin{cases} L^* = -\frac{\mu}{2\sigma^2} \beta^T X^T \\ -\frac{\mu}{2\sigma^2} \beta^T X^T X\beta = \beta \quad (\star) \end{cases}$$

$$(\star) \quad -2\sigma^2 \mu \beta^T X^T X = I \Rightarrow \mu\beta^T = -2\sigma^2 (X^T X)^{-1} \Rightarrow L^* = (X^T X)^{-1} X^T$$

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$$

**P.S:** The same problem, but for  $\epsilon \sim \mathcal{N}(0, \Sigma)$

$$\begin{cases} Tr(Var(Ly)) \rightarrow \min_L \\ \mathbb{E}(Ly) = \beta \end{cases}$$

$$\mathcal{L} = Tr(Var(LX\beta + L\epsilon)) + \mu^T (LX\beta - \beta) = Tr(L\Sigma L^T) + \mu^T (LX\beta - \beta)$$

$$\nabla_L (Tr(L\Sigma L^T)) = 2L\Sigma^T = 2L\Sigma$$

$$DTr(L\Sigma L^T)[\Delta L] = Tr(L\Sigma \Delta L^T + \Delta L \Sigma L^T) = Tr(2\Sigma L^T \Delta L) = Tr(\nabla f(L)^T \Delta L)$$

K.K.T. conditions:

$$\begin{cases} \nabla_L \mathcal{L} = 2L\Sigma + \mu\beta^T X^T = 0 \\ LX\beta = \beta \end{cases} \Rightarrow \begin{cases} L^* = -2\mu\beta^T X^T \Sigma^{-1} \\ -2\mu\beta^T X^T \Sigma^{-1} X\beta = \beta \quad (\star) \end{cases}$$

$$(\star) \quad -2\mu\beta^T X^T \Sigma^{-1} X = I \Rightarrow \mu\beta^T = -\frac{1}{2} (X^T \Sigma^{-1} X)^{-1}$$

$$L^* = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}$$

$$\boxed{\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y}$$



## Application (Principle Component Analysis)

Finding the 1st component:

$$\begin{cases} \|Xa\|_2^2 \rightarrow \max_a \\ \text{s.t. } \|a\|_2^2 = 1 \end{cases}$$

$$L = a^T X^T X a + \mu (a^T a - 1)$$

$$\text{K.K.T. conditions: } \begin{cases} \nabla_a L = 2X^T X a + 2\mu a = 0 & (1) \\ a^T a = 1 \end{cases}$$

From the (1) it follows that,  $X^T X a = -\mu a \Rightarrow a$  is an eigenvector of  $X^T X$ . Let's call the corresponding eigenvalue  $\lambda$ , then the problem  $\|Xa\|_2^2 \rightarrow \max_a$  would take the form:  $a^T X^T X a = a^T \lambda a = \lambda \cdot a^T a = \lambda \underbrace{\|a\|_2^2}_1 \rightarrow \max \Rightarrow \lambda = \text{maximum eigenvalue of } X^T X$ , so the answer would be  $a =$  the eigenvector, corresponding to the maximum eigenvalue of matrix  $X^T X$

Finding the  $k^{th}$  principal component:

$$\begin{cases} \|Xa_k\|_2^2 \rightarrow \max_{a_k} \\ \langle a_k, a_i \rangle = 0, \forall i \neq k \\ \|a_k\|_2^2 = 1 \end{cases}$$

$$L = a^T X^T X a + \mu (a^T a - 1) + \sum_{i=1}^{k-1} \gamma_i a_k^T a_i$$

$$\begin{aligned} \nabla_{a_k} L &= 2X^T X a_k + 2\mu a_k + \sum_{i=1}^{k-1} \gamma_i a_i = 0 \Big| \cdot a_k^T \Rightarrow 2a_k^T X^T X a_k + 2\mu a_k^T + \\ \underbrace{\sum_{i=1}^{k-1} \gamma_i a_k^T a_i}_0 &= 0 \Rightarrow a_k \text{ is an eigenvector of } X^T X, \text{ corresponding to the next} \\ &\text{biggest eigenvalue.} \end{aligned}$$

$$\text{Example : } \begin{cases} x^T A x \rightarrow \min_x \\ \text{s.t. } \|x\|_2^2 \leq 1 \end{cases}, \text{ where } A = A^T > 0$$

$$L = x^T A x + \lambda (x^T x - 1)$$

K.K.T conditions:

$$\left\{ \begin{array}{l} \nabla_x L = 2Ax + 2\lambda x = 0 \\ x^T x \leq 1 \\ \lambda(x^T x - 1) = 0 \\ \lambda \geq 0 \end{array} \right. \Rightarrow Ax = -\lambda x, \text{ so } x \text{ should be the eigenvector of}$$

matrix  $A$  for example with eigenvalue  $\gamma$ , then the dual function:

## Other

**Example 1:** Find  $\mathbb{E}(x^T x)$ , where  $x \sim \mathcal{N}(\mu, \Sigma)$

$$\mathbb{E}x = \mu, \quad \mathbb{E}(x - \mu)(x - \mu)^T = \Sigma$$

$$\begin{aligned} \mathbb{E}(x - \mu)(x - \mu)^T &= \mathbb{E}xx^T - 2\mu^T \mathbb{E}x + \mathbb{E}\mu^T \mu = \mathbb{E}xx^T - 2\mu\mu^T + \mu\mu^T = \\ \mathbb{E}xx^T - \mu\mu^T &= \Sigma \Rightarrow \mathbb{E}xx^T = \Sigma + \mu\mu^T \end{aligned}$$

$$\begin{aligned} \mathbb{E}x^T x &= \text{tr}(\mathbb{E}x^T x) = \mathbb{E}(\text{tr}(x^T x)) = \mathbb{E}(\text{tr}(xx^T)) = \text{tr}(\mathbb{E}xx^T) = \\ \text{tr}(\Sigma + \mu\mu^T) &= \text{tr}\Sigma + \mu^T \mu \end{aligned}$$

$$\boxed{\mathbb{E}x^T x = \text{tr}\Sigma + \mu^T \mu}$$

Here we used the fact, that  $\text{tr}(aa^T) = a^T a$