Matrix Calculus

Anton Andreytsev

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

From the definition above it follows, that

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

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

Generalisations:

$x \in \mathbb{R} \Rightarrow Df(x)[\Delta x] = \nabla f(x) \cdot \Delta x, \quad \nabla f(x) \in \mathbb{R}$
$x \in \mathbb{R}^n \Rightarrow Df(x)[\Delta x] = \nabla f(x)^T \Delta x, \quad \nabla f(x) \in \mathbb{R}^n$
$x \in \mathbb{R}^{mxn} \Rightarrow Df(x)[\Delta x] = Tr\left(\nabla f(x)^T \Delta x\right), \quad \nabla f(x) \in \mathbb{R}^{mxn}$

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

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

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

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

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

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

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

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

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

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

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

In case of l_2- regularisation:

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

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

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

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

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

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

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

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

So $x^T e^{xx^T} x = x^T \sum_{i=0}^{+\infty} \frac{(xx^T)^i}{i!} x = \sum_{i=0}^{+\infty} \frac{x^T (xx^T)^i x}{i!} = \sum_{i=0}^{+\infty} \frac{\overbrace{x^T (x \; x^T)(xx^T)...(xx^T)x}^{i+1\text{-times}}}{i!} =$

$\sum_{i=0}^{+\infty} \frac{(x^T x)^{i+1}}{i!} = x^T x \sum_{i=0}^{+\infty} \frac{(x^T x)^i}{i!} = x^T x \exp(x^T x)$

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

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

$$\nabla\left(x^T x \exp(x^T x)\right) = \nabla(x^T x) \cdot \exp(x^T x) + x^T x \cdot \nabla(\exp(x^T x)) =$$
$$2 \cdot x \cdot \exp(x^T x) + x^T x \cdot 2\exp(x^T x)x = 2\exp(x^T x)\left(1 + x^T x\right)x$$

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

Example 5: $f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) = Det\left(2I + xx^T\right)$

We will need here some properties of matrix differentiation:

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

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

$\boxed{2}$ $D\left(Tr(X)\right)[\Delta X] = Tr(\Delta X)$

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

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

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

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

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

$Det(2I + xx^T)[Tr\left(Ax\Delta x^T\right) + Tr\left(A\Delta x x^T\right)] =$

{we could rearange the order inside trace and transpose, because $A = A^T$

$= \underbrace{(2I + xx^T)}_{constant}[2Tr(x^T A\Delta x)] = Tr\left(2Det(2I + xx^T)x^T A\Delta x\right) = Tr(\nabla f(x)^T\Delta x)$

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

Example 6: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(x) = \log Det(x), \quad \nabla f(x) = ?$

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

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

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

$$(2) \quad DDet(X)[\Delta X] \quad = \quad \lim_{t\to+0} \frac{Det(X+t\cdot\Delta X)-Det(X)}{t} \quad =$$

$$\lim_{t\to+0} \frac{Det(X\cdot[I+t\cdot X^{-1}\cdot\Delta X])-Det(X)}{t} \quad = \quad \lim_{t\to+0} \frac{Det(X)\cdot[Det(I+t\cdot X^{-1}\cdot\Delta X)-1]}{t} \quad =$$

$$\lim_{t\to+0} \frac{Det(X)\cdot[\exp(\log(Det(I+t\cdot X^{-1}\cdot\Delta X)))-1]}{t} \quad = \quad \lim_{t\to+0} \frac{Det(X)\cdot\log(Det(I+t\cdot X^{-1}\cdot\Delta X))}{t} \quad =$$

$$\{\texttt{for small } \epsilon \texttt{ holds} : Det(I + \epsilon \cdot A) \approx 1 + \epsilon \cdot Tr(A) + \bar{o}(\epsilon)\} \quad =$$

$$\lim_{t\to+0} \frac{Det(X)\cdot\log(1+t\cdot Tr(X^{-1}\Delta X))}{t} = \lim_{t\to+0} \frac{Det(X)\cdot t\cdot Tr(X^{-1}\Delta X)}{t} = \boxed{Det(X)\cdot Tr(X^{-1}\Delta X)}$$

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

$$\boxed{\nabla(\log Det(X)) = X^{-T}}$$

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

Example 7: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = a^T X a$

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

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

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

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

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

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

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

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

$$\nabla_\Sigma \log L \quad = \quad \{\texttt{for convenience let } \Lambda = \Sigma^{-1}\} \quad =$$
$$\nabla_\Lambda \left(-\tfrac{n}{2} \log Det\Lambda^{-1} - \tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Lambda(x_i - \mu) \right) = \tfrac{n}{2}\underbrace{\Lambda^{-T}}_{\Lambda^{-1}} - \tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i -$$

4

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

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

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

First let's consider, that the $argmax_X f(X) = argmax_X \log f(X)$

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

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

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

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

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

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

Example 9: $f : \mathbb{R}^{nxn} \to \mathbb{R}^{nxn}, \quad f(X) = X^{-1}, \quad Df(X)[\Delta X] = ?$

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

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

Example 10: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = Tr(X), \quad \nabla f(X) = ?$

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

Example 11: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = Tr(AX^{-1}B), \quad \nabla f(X) = ?$

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

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

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

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

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

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

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

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

Example 13: $f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) = \frac{x^T Ax}{x^T x}, \quad \nabla f(x) = ?$

$$\nabla f(x) \quad = \quad \nabla\left(\frac{g(x)}{h(x)}\right) \quad = \quad \frac{\nabla g(x) \cdot h(x) - g(x) \cdot \nabla h(x)}{h^2(x)} \quad = \quad \frac{(A+A^T)x \cdot x^T x - 2x^T Ax \cdot x}{(x^T x)^2} \quad =$$

$\frac{(A+A^T)x^T x - 2x^T Ax \cdot I}{(x^T x)^2} \cdot x$

Example 14: $f : \mathbb{R}^{n x n} \to \mathbb{R}, \quad f(X) = Det(AXB), \quad \nabla f(X) = ?$

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

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

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

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

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

Let's apply the property of matrix differentiation:

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

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

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

$$\underbrace{DTr(XB)[\Delta X]}_{(2)}$$

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

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

(2) $DTr(XB) = Tr(\Delta X B)$, so

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

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

Example 16: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = Det(\exp(X)), \quad \nabla f(X) =?$

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

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

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

Example 17: $f : \mathbb{R}^{mxn} \to \mathbb{R}, \quad f(X) = \frac{1}{2}\|X - A\|_F^2, \quad \nabla f(X) =?$

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

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

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

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

$$Df(x)[\Delta x] = Tr(D(Axx^T)[\Delta x]) = Tr(Ax\Delta x^T + A\Delta xx^T) = Tr([x\Delta x^T]^T A^T) + Tr(x^T A\Delta x) = Tr(x^T A^T \Delta x + x^T A\Delta x) = Tr(\nabla f(x)^T \Delta x)$$

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

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

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

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

$$\nabla f(x) = [2x^T xx^T - \frac{1}{2}x^T(A + A^T)]^T = [x^T \left(2xx^T - \frac{1}{2}A - \frac{1}{2}A^T\right)]^T = (2xx^T - \frac{1}{2}(A + A^T))x$$

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

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

Example 21: $f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) = x^T \vec{\log}(x), \nabla f(x) =?, \text{where} \quad \vec{\log}(x) = (\log x_1, \ldots, \log x_n)$

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

Example 22: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = Det(A + X^{-1}), \quad \nabla f(X) =?$

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

$$Df(X)[\Delta X] = Det(A + X^{-1})Tr(-(A + X^{-1})^{-1}X^{-1}\Delta XX^{-1}) = Tr\left(-Det(A + X^{-1})X^{-1}(A + X^{-1})^{-1}X^{-1}\Delta X\right) = Tr(\nabla f(X)^T \Delta X)$$

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

Example 23: (The problem of constructing surrogat eigenvector)

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

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

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

$\nabla f(x) = (A - \lambda \cdot I)x = 0$, from this follows, that a minimiser of this function is an eigenvector of a matrix A, corresponding to an eigenvalue λ, 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,

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

Application: (Logistic Regression fitting)

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

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

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

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

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

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

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

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

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

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

1) $Det(X^2) \quad = \quad Det(X)^2 \quad \Rightarrow \quad DDet(X)^2[\Delta X] \quad =$
$\begin{cases} DY^2[\Delta Y] = Y\Delta Y + \Delta Y Y \quad (1) \\ \Delta Y = DDet(X)[\Delta X] = Det(X)Tr(X^{-1}\Delta X) \end{cases}$

(1) $Df(X)[\Delta X] \quad = \quad 2Det(X) \cdot Det(X)Tr(X^{-1}\Delta X) \quad \Rightarrow \quad \nabla f(X) \quad =$
$2Det(X)^2 X^{-T}$

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

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

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

$Df(X)[\Delta X] = \begin{cases} DTr(AY)[\Delta Y] = Tr(A\Delta Y) \\ \Delta Y = DX^{-1}[\Delta X] = -X^{-1}\Delta X X^{-1} \end{cases}$

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

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

Application: Kernel Linear Regression Problem

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

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

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

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

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

$DTr(X^{-1}A)[\Delta X] = Tr(-X^{-1}\Delta X X^{-1}A) = Tr(-X^{-1}AX^{-1}\Delta X) \Rightarrow$
$\nabla Tr(X^{-1}A) = -X^{-T}A^T X^{-T}$

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

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

Higher order derivatives

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

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

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

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

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

Example 2: $f : \mathbb{R}^{nxn} \to \mathbb{R}, \quad f(X) = \log Det(X), \quad D^2 f(X[\Delta X_1, \Delta X_2] = ?$

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

$$D^2 f(X)[\Delta X_1, \Delta X_2] = \lim_{t \to +0} \frac{Df(X + t \cdot \Delta X_2)[\Delta X_1] - Df(X)[\Delta X_1]}{t} =$$
$$\lim_{t \to +0} \frac{Tr((X + t \cdot \Delta X_2)^{-1}[\Delta X_1]) - Tr(X^{-1}[\Delta X_1])}{t} = \lim_{t \to +0} \frac{Tr\left([(X + t \cdot \Delta X_2)^{-1} - X^{-1}][\Delta X_1]\right)}{t} =$$
$$\lim_{t \to +0} \frac{Tr\left(([X(I + t \cdot X^{-1}\Delta X_2)]^{-1} - X^{-1})[\Delta X_1]\right)}{t} = \lim_{t \to +0} \frac{Tr\left([(I + t \cdot X^{-1}\Delta X_2)^{-1} - I]X^{-1}\Delta X_1\right)}{t} =$$
$$\lim_{t \to +0} \frac{Tr(-t \cdot X^{-1}\Delta X_2 X^{-1}\Delta X_1)}{t} = Tr\left(-\Delta X_1 X^{-1}\Delta X_2 X^{-1}\right)$$

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

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

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

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

Constraint optimisation

Example 1: $\begin{cases} \|x\|_2^2 \to \min\limits_x \\ \texttt{s.t.}\, Ax \le b \end{cases}$

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

K.K.T conditions:

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

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

Application: OLS under constraints:

$\begin{cases} \|X\beta - y\|_2^2 \to \min\limits_\beta \\ \texttt{s.t.}\, C\beta = d \end{cases}$

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

K.K.T. conditions:

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

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

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

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

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

$V(\beta^*) = \sigma^2 \left(X^T X \right)^{-1} - \sigma^2 \cdot (X^T X)^{-1} C^T \left(C(X^T X)^{-1} C^T \right)^{-1} C (X^T X)^{-1} \cdot$
$\cdot ((X^T X)^{-1} C^T \left(C(X^T X)^{-1} C^T \right)^{-1} C)^T \quad = \quad \sigma^2 \left(X^T X \right)^{-1} \quad - \quad \sigma^2 \quad \cdot$

$$(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1} =$$
$$\sigma^2(X^TX)^{-1} \quad - \quad \sigma^2\left(X^TX\right)^{-1}\left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1} \quad =$$
$$\boxed{V(\beta^*) = \sigma^2(X^TX)^{-1}\left(I - \left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1}\right)}$$

When C is invertible this could be reduced to:

$$\sigma^2(X^TX)^{-1}\left(I - \left(C(X^TX)^{-1}C^T\right)^{-1}C(X^TX)^{-1}C^TC^{-T}\right) \quad =$$
$$\boxed{V(\beta^*) = \sigma^2(X^TX)^{-1}\left(I - C^{-T}\right)}$$

Example 2: $\begin{cases} a^Tx \to \min\limits_{x} \\ \text{s.t.} x^TAx \le 1 \end{cases}$, where $A = A^T > 0$

$$L = a^Tx + \lambda(x^TAx - 1)$$

K.K.T. conditions:

$$\begin{cases} \nabla_x L = a + 2\lambda Ax = 0 \\ x^TAx - 1 \le 0 \\ \lambda(x^TAx - 1) = 0 \\ \lambda \ge 0 \end{cases} \Rightarrow \begin{cases} x^\star = -\frac{1}{2\lambda}A^{-1}a \\ \lambda\left((-\frac{1}{2\lambda}A^{-1}a)^TA(-\frac{1}{2\lambda}A^{-1}a) - 1\right) = 0 \quad (1) \end{cases}$$

(1) $\frac{1}{4\lambda}a^TA^{-1}a - \lambda = 0 \big| \cdot \lambda \Rightarrow \lambda^\star = \pm\frac{1}{2}\sqrt{a^TA^{-1}a}, \quad \lambda \ge 0 \Rightarrow \lambda^\star = \frac{1}{2}\sqrt{a^TA^{-1}a}$

$$\boxed{x^\star = -\frac{A^{-1}a}{\sqrt{a^TA^{-1}a}}}$$

Example 3: $\begin{cases} x^TQx \to \min\limits_{x} \\ \text{s.t.} \|Ax - b\|_2^2 \le 1 \end{cases}$, where $Q = Q^T > 0, A = A^T > 0$

$$L \quad = \quad x^TQx \quad + \quad \lambda\left((Ax - b)^T(Ax - b) - 1\right) \quad = \quad x^TQx \quad +$$
$$\lambda\left(x^TA^TAx - 2b^TAx + b^Tb - 1\right)$$

K.K.T. conditions:
$$\begin{cases} \nabla_x L = 2Qx + 2\lambda A^TAx - 2\lambda A^Tb = 0 \\ \lambda \ge 0 \\ \lambda\left(x^TA^TAx - 2b^TAx + b^Tb - 1\right) = 0 \\ \|Ax - b\|_2^2 \le 1 \end{cases} \Rightarrow$$

$$\begin{cases} \boxed{x^\star = \lambda\left(Q + \lambda A^TA\right)^{-1}A^Tb} \\ \lambda \ge 0 \\ \lambda\left(x^TA^TAx - 2b^TAx + b^Tb - 1\right) = 0 \\ \|Ax - b\|_2^2 \le 1 \end{cases}$$

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

Application: (2-rank update)
$$\begin{cases} \|B - B_k\|_F^2 \to \min\limits_B \\ \quad \text{s.t.} \, Bs_k = y_k \end{cases}$$
$$L = Tr(B^T B) + \mu^T (Bs_k - y_k)$$

K.K.T. conditions:
$$\begin{cases} \nabla_B L = 2B + \mu s_k^T = 0 \\ \quad\quad Bs_k = y_k \end{cases} \Rightarrow B = -\tfrac{1}{2}\mu s_k^T$$
Dual function $q(\mu) = Tr(\tfrac{1}{4} s_k \mu^T \mu s_k^T) - \tfrac{1}{2}\mu^T \mu s_k^T s_k - \mu^T y_k \to \min\limits_\mu$

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

So $\boxed{B^\star = \dfrac{y_k s_k^T}{s_k^T s_k}}$

Example 4:
$$\begin{cases} \tfrac{1}{2} x^T A x + b^T x + c \to \min\limits_x \\ \quad\quad\quad \text{s.t.} \, x^T x \le 1 \end{cases} \text{ where } A = A^T > 0$$
$$L = \tfrac{1}{2} x^T A x + b^T x + c + \lambda(x^T x - 1)$$

K.K.T. conditions:
$$\begin{cases} \nabla_x L = Ax + b + \lambda x = 0 \\ \quad\quad x^T x - 1 \le 0 \\ \quad\quad \lambda(x^T x - 1) = 0 \\ \quad\quad\quad \lambda \ge 0 \end{cases} \Rightarrow \begin{cases} \boxed{x^\star = -(A + I \cdot \lambda)^{-1} b} \\ \lambda b^T (A + \lambda \cdot I)^{-T}(A + \lambda \cdot I)^{-1} b - \lambda = 0, \quad (\star) \end{cases}$$
$$(\star)\lambda(b^T (A + \lambda \cdot I)^{-2} b - 1) = 0 \Rightarrow \lambda^\star = \max\{0, \text{solution of } (\star)\}$$

Example 5: (Projection on the radius 1 Ball)
$$\begin{cases} \tfrac{1}{2}\|x - v\|_2^2 \to \min\limits_x \\ \quad \text{s.t.} \, x^T x \le 1 \end{cases}$$
$$L = \tfrac{1}{2}\left(x^T x - 2v^T x + v^T v\right) + \lambda(x^T x - 1)$$

K.K.T. conditions:

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

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

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

Application: (Hard margin SVM)
$$\begin{cases} \frac{1}{2}\|w\|_2^2 \to \min_w \\ y_i\left(x_i^T w - b\right) \geq 1 \end{cases}$$

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

K.K.T. conditions:
$$\begin{cases} \nabla_w L = w - \sum\limits_{i=1}^n \lambda_i y_i x_i = 0 \\ \nabla_b L = \sum\limits_{i=1}^n \lambda_i y_i = 0 \\ \lambda_i \geq 0, \forall i = 1, \ldots, n \\ \lambda_i \left(y_i \left(x_i^T w - b\right) - 1\right) = 0, \forall i = 1, \ldots, n \end{cases}$$

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

Application: Consider the multiple regression model $y = X \cdot \beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, lets propose a linear estimator of the form $\hat{\beta} = L \cdot y$. Find the unbiased estimator $\hat{\beta}$, for which $Tr(Var(\hat{\beta})) \to \min\limits_\beta$

$$\begin{cases} Tr(Var(Ly)) \to \min\limits_L \\ \mathbb{E}\left(\hat{\beta}\right) = \beta \end{cases}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$L^\star = \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}$$

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

Application (Principle Component Analysis)

Finding the 1st component:

$$\begin{cases} \|Xa\|_2^2 \to \max_a \\ \texttt{s.t.}\,\|a\|_2^2 = 1 \end{cases}$$

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

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

From the (1) it follows that, $X^T Xa = -\mu a \Rightarrow a$ is an eigenvector of $X^T X$. Let's call the corresponding eigenvalue λ, then the problem $\|Xa\|_2^2 \to \max_a$ would

take the form: $a^T X^T X a = a^T \lambda a = \lambda \cdot a^T a = \lambda \underbrace{\|a\|_2^2}_{1} \to \max \Rightarrow \lambda = \text{maximum}$

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 kth principal component:

$$\begin{cases} \|Xa_k\|_2^2 \to \max\limits_{a_k} \\ <a_k, a_i> = 0, \forall i \neq k \\ \|a_k\|_2^2 = 1 \end{cases}$$

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

$$\nabla_{a_k} L = 2X^T X a_k + 2\mu a_k + \sum_{i=1}^{k-1} \gamma_i a_i = 0 \bigg| \cdot a_k^T \Rightarrow 2a_k^T X^T X a_k + 2\mu a_k^T +$$

$$\underbrace{\sum_{i=1}^{k-1} \gamma_i a_k^T a_i}_{0} = 0 \Rightarrow a_k \text{ is an eigenvector of } X^T X \text{ , corresponding to the next}$$

biggest eigenvalue.

$$\textbf{Example :} \begin{cases} x^T A x \to \min\limits_{x} \\ \texttt{s.t.} \|x\|_2^2 \leq 1 \end{cases} \texttt{,where } A = A^T > 0$$

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

$$\begin{cases} \text{K.K.T conditions:} \\ \nabla_x L = 2Ax + 2\lambda x = 0 \\ \quad x^T x \leq 1 \\ \quad \lambda(x^T x - 1) = 0 \\ \quad \lambda \geq 0 \end{cases} \Rightarrow Ax = -\lambda x, \text{ so x should be the eigenvector of}$$

matrix A for example with eigenvalue γ, then the dual function:

Other

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

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

$$\mathbb{E}\left(x - \mu\right)\left(x - \mu\right)^T = \mathbb{E}xx^T - 2\mu^T \mathbb{E}x + \mathbb{E}\mu^T \mu = \mathbb{E}xx^T - 2\mu\mu^T + \mu\mu^T =$$

$$\mathbb{E}xx^T - \mu\mu^T = \Sigma \Rightarrow \mathbb{E}xx^T = \Sigma + \mu\mu^T$$

$$\mathbb{E}x^T x = tr\left(\mathbb{E}x^T x\right) = \mathbb{E}\left(tr(x^T x)\right) = \mathbb{E}\left(tr(xx^T)\right) = tr\left(\mathbb{E}xx^T\right) = tr\left(\Sigma + \mu\mu^T\right) = tr\Sigma + \mu^T\mu$$

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