MAS473 Homework 1. 20200130 Yujun Kim Sep / 21 / 2023

#1

Singular values are square root of eigenvalues of A^TA.

Let $A=U\Sigma V^T$ be the SVD of A. For $A\in M_{m\times n}$, we have $U\in M_{m\times m}$, $V\in M_{n\times n}$ unitary and $\Sigma\in M_{m\times n}$ rectangular diagonal with diagonal elements $\sigma_1\geq\sigma_2\geq\cdots\geq 0$. Let $\|\cdot\|$ denote $\|\cdot\|_2$.

Claim: $\|A\|_2=\sigma_1$

($\geq$) Let $V=[v_1\cdots v_n]$, $U=[u_1,\cdots,u_m]$. As V,U are unitary, $\|v_1\|_2=\|u_1\|_2=1$.

$Av_1=U\Sigma V^Tv_1=U\Sigma e_1$, where $e_1=(1,0,\cdots,0)\in\mathbb{R}^n$

$=U\sigma_1\bar{e}_1$, where $\bar{e}_1=(1,0,\cdots,0)\in\mathbb{R}^m$

$=\sigma_1(U\bar{e}_1)=\sigma_1u_1$.

Thus, $\|A\|_2=\max_{\|x\|_2=1}\|Ax\|_2\geq\|Av_1\|=\|\sigma_1u_1\|=\sigma_1\|u_1\|=\sigma_1$.

($\leq$) Now, for any $x\in\mathbb{R}^n$ with $\|x\|_2=1$, let $y=V^Tx$. Then, $Vy=VV^Tx=x$.

As V^T is unitary, $\|y\|=\|V^Tx\|=\|x\|=1$

$Ax=U\Sigma V^Tx=U\Sigma y$. As U is unitary, $\|Ax\|=\|U\Sigma y\|=\|\Sigma y\|$. ···①

$\|\Sigma y\|^2=\sum_{i=1}^m(\sigma_iy_i)^2$ (σ_i might be 0 for some i)

$\leq\sum_{i=1}^m(\sigma_iy_i)^2=\sigma_i^2\|y\|^2=\sigma_i^2$. ···②

①,② $\Rightarrow\|Ax\|=\|\Sigma y\|\leq\sigma_1$ and we are done. (By taking max. over $\|x\|_2=1$)

#2

(a) $KL(p\|q):=\int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$

By convention $0\log\left(\frac{0}{0}\right)=0\log\left(\frac{0}{a}\right)=a\log\left(\frac{a}{0}\right)=\infty$ from lecture slide,

i) If $[a,b]\subseteq[c,d]$, then integrant $=0\log\left(\frac{0}{?}\right)=0$ outside $[a,b]$.

Thus, $KL(U_{[a,b]}\|U_{[c,d]})=\int_a^b\frac{1}{b-a}\log\frac{d-c}{b-a}\,dx=\log\frac{d-c}{b-a}$.

ii) Otherwise, $a<c$ or $d<b$. Then, integrant $=\frac{1}{b-a}\log\left(\frac{\frac{1}{b-a}}{0}\right)$ on (a,c) or (d,b).

Integration of fn with value ∞ at positive measure set is ∞.

Thus, $KL(U_{[a,b]}\|U_{[c,d]})=\infty$.

(b) $KL(N(\mu_0,\sigma_0^2)\,\|\,N(\mu_1,\sigma_1^2)) = \int \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma_0^2}(x-\mu_0)^2} \log\left(\frac{\sigma_1}{\sigma_0} e^{\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_0^2}(x-\mu_0)^2}\right) dx$

$= \log\left(\frac{\sigma_1}{\sigma_0}\right) \underbrace{\int \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma_0^2}(x-\mu_0)^2} dx}_{①} + \underbrace{\int \left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right) \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma_0^2}(x-\mu_0)^2} dx}_{②}$.

Let's denote $p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$.

Note that (1) $\int p(x;\mu,\sigma)\, dx = \mathbb{E}[1] = 1$

(2) $\int x\, p(x;\mu,\sigma)\, dx = \mathbb{E}[X] = \mu$ for $X \sim N(\mu, \sigma^2)$

(3) $\int (x-\mu)^2 p(x;\mu,\sigma)\, dx = \mathbb{E}[(X-\mu)^2] = \sigma^2$ for $X \sim N(\mu, \sigma^2)$.

By (1), ① $= \log\left(\frac{\sigma_1}{\sigma_0}\right)$.

② $= \int \left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right) p(x;\mu_0,\sigma_0)\, dx$

$= \underbrace{\int \frac{(x-\mu_1)^2}{2\sigma_1^2} p(x;\mu_0,\sigma_0)\, dx}_{③} - \underbrace{\int \frac{(x-\mu_0)^2}{2\sigma_0^2} p(x;\mu_0,\sigma_0)\, dx}_{④}$.

By (3), ④ $= \frac{\sigma_0^2}{2\sigma_0^2} = \frac{1}{2}$.

Finally, ③ $= \int \frac{1}{2\sigma_1^2} (x-\mu_0+\mu_0-\mu_1)^2 p(x;\mu_0,\sigma_0)\, dx$

$= \int \frac{1}{2\sigma_1^2} \left[(x-\mu_0)^2 + 2(\mu_0-\mu_1)(x-\mu_0) + (\mu_0-\mu_1)^2\right] p(x;\mu_0,\sigma_0)\, dx$

$= \frac{1}{2\sigma_1^2}\left[\underbrace{\sigma_0^2}_{\text{By (3)}} + 2(\mu_0-\mu_1)\underbrace{\times 0}_{\text{By (2)}} + (\mu_0-\mu_1)^2 \underbrace{\times 1}_{\text{By (1)}}\right]$

$= \frac{1}{2\sigma_1^2}\left[\sigma_0^2 + (\mu_0-\mu_1)^2\right]$

Thus, $KL(N(\mu_0,\sigma_0^2)\|N(\mu_1,\sigma_1^2)) = \log\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{1}{2\sigma_1^2}[\sigma_0^2+(\mu_0-\mu_1)^2] - \frac{1}{2}$

$= \log\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{1}{2\sigma_1^2}[\sigma_0^2-\sigma_1^2+(\mu_0-\mu_1)^2]$

#3

Notes

When handling multivariate calculus, $x\in\mathbb{R}^n$, $y=f(x)=\mathbb{R}^m$, we have two choice of representation $\frac{dy}{dx}\in\mathcal{M}_{m\times n}$ or $\frac{dy}{dx}\in\mathcal{M}_{n\times m}$.
For example, $y=Ax$, $A\in\mathcal{M}_{m\times n}$ gives $\frac{dy}{dx}=A$ in first representation
$\frac{dy}{dx}=A^T$ in the 2nd representation

This matters when applying chain rule. $x\in\mathbb{R}^n$, $y=f(x)=\mathbb{R}^m$, $z=g(y)=\mathbb{R}^l$.
$\frac{dz}{dx}=\frac{dz}{dy}\times\frac{dy}{dx}\in\mathcal{M}_{l\times n}$, with $\frac{dz}{dy}\in\mathcal{M}_{l\times m}$, $\frac{dy}{dx}\in\mathcal{M}_{m\times n}$ in 1st rep.
$\frac{dz}{dx}=\frac{dy}{dx}\times\frac{dz}{dy}\in\mathcal{M}_{n\times l}$, with $\frac{dz}{dy}\in\mathcal{M}_{m\times l}$, $\frac{dy}{dx}\in\mathcal{M}_{n\times m}$ in 2nd rep.
Let's be consistent with the 1st representation.
For example, $\frac{d}{dt}t^Tt=2t^T\in\mathcal{M}_{1\times n}$ for $t\in\mathbb{R}^n$.

(a) By chain rule, $\frac{d}{dt}\sin(\log(t^Tt)) = \cos(\log(t^Tt))\cdot\frac{1}{t^Tt}\cdot 2t^T$

(b) For simplicity, for $M\in\mathcal{M}_{m\times n}$, $y=f(M)\in\mathbb{R}$, let $\frac{dy}{dM}\in\mathcal{M}_{m\times n}$ with $\left(\frac{df}{dM}\right)_{ij}=\frac{df}{\partial M_{ij}}$.

$$\begin{aligned}
tr(AXB) &= \sum_{i=1}^{D}(AXB)_{ii}\\
&= \sum_{i=1}^{D}\sum_{j=1}^{E}A_{ij}(XB)_{ji}\\
&= \sum_{i=1}^{D}\sum_{j=1}^{E}\sum_{k=1}^{F}A_{ij}X_{jk}B_{ki}.\\
&= \sum_{j=1}^{E}\sum_{k=1}^{F}X_{jk}\left(\sum_{i=1}^{D}B_{ki}A_{ij}\right) = \sum_{j=1}^{E}\sum_{k=1}^{F}X_{jk}(BA)_{kj}
\end{aligned}$$

Hence, $\frac{d}{dX_{jk}}tr(AXB)=(BA)_{kj}$. i.e. $\frac{d}{dX}tr(AXB)=(BA)^T$

#4

(a) Given samples $x_1, \cdots, x_n$ of $X_1, \cdots, X_n$,

$$l := \log p(X_1 = x_1, \cdots, X_n = x_n) = \log \prod_{i=1}^{n} p(X_i = x_i)$$

(joint distribution)

$$= \sum_{i=1}^{n} \log p(X_i = x_i) = \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{K/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$

$$= nC - \sum_{i=1}^{n} \frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \Sigma^{-1}(x_i-\mu) = \Sigma^{-1}\left(\left(\sum_{i=1}^{n} x_i\right) - n\mu\right)$$

$$= n\Sigma^{-1}(\bar{x}-\mu), \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Thus, $\frac{\partial l}{\partial \mu} = 0$ gives $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. (MLE)

(b) Posterior $p(\mu | x_1, \cdots, x_n) = \frac{p(x_1, \cdots, x_n | \mu)\, p(\mu)}{\int p(x_1, \cdots, x_n | \mu)\, p(\mu)\, d\mu}$

$$= \frac{\prod_{i=1}^{n} \mathcal{N}(x_i | \mu, \Sigma)\, \mathcal{N}(\mu | \mu_0, \Sigma_0)}{\int \prod_{i=1}^{n} \mathcal{N}(x_i | \mu, \Sigma)\, d\mu}$$

$$= \frac{\frac{1}{(2\pi)^{nK/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)} \frac{1}{(2\pi)^{K/2}|\Sigma_0|^{1/2}} e^{-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)}}{(\text{constant wrt } \mu)}$$

As denominator of (*) is indep. of μ,

$$\frac{d}{d\mu} \log p(\mu | x_1, \cdots, x_n) = \frac{d}{d\mu}\Big[(\text{constant wrt } \mu) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\Big]$$

$$= -n\Sigma^{-1}(\bar{x}-\mu) - \Sigma_0^{-1}(\mu-\mu_0)$$

Hence, $\frac{\partial l}{\partial \mu} = 0$ gives $n\Sigma^{-1}(\bar{x}-\mu) = \Sigma_0^{-1}(\mu-\mu_0)$

$$\Rightarrow (n\Sigma^{-1} + \Sigma_0^{-1})\mu = n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0$$

$$\Rightarrow \mu = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}(n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$$

i.e. MAP is $\hat{\mu} = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}(n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$

/ /

#5

Note that if $L: \mathbb{R}^n \to \mathbb{R}$ is linear, $\varphi: \mathbb{R} \to \mathbb{R}$ is convex, then $\varphi \circ L$ is convex.

$\because$) $\varphi \circ L(\lambda x + (1-\lambda)y) = \varphi(\lambda L(x) + (1-\lambda)L(y)) \le \lambda \varphi(L(x)) + (1-\lambda)\varphi(L(y))$.

① First we show $g: \mathbb{R} \to \mathbb{R}$ by $g(t) = -\log \sigma(t)$ is convex.

$\because$) $\sigma(t) = \frac{1}{1+e^{-t}}$, $\sigma'(t) = \frac{e^{-t}}{(1+e^{-t})^2} = e^{-t}\sigma(t)^2$.

$\Rightarrow g'(t) = -\frac{\sigma'(t)}{\sigma(t)} = -e^{-t}\sigma(t)$

$\Rightarrow g''(t) = e^{-t}\sigma(t) - e^{-t}\sigma'(t)$

$= e^{-t}(\sigma(t) - \sigma'(t))$

$= e^{-t}(\sigma(t) - e^{-t}\sigma(t)^2)$

$= e^{-t}\sigma(t)^2(\frac{1}{\sigma(t)} - e^{-t})$

$= e^{-t}\sigma(t)^2(1+e^{-t}-e^{-t}) = e^{-t}\sigma(t)^2 \ge 0$.

Thus, g is convex.

② $h: \mathbb{R} \to \mathbb{R}$ by $h(t) = -\log(1-\sigma(t))$ is convex.

$\because$) $h(t) = -\log(\sigma(-t)) = g(-t)$, where g is convex, and $t \mapsto -t$ is linear.

↑ $\sigma(t) + \sigma(-t) = 1$.

③ f is convex.

$L_i(\theta)$ by $L_i(\theta) = x_i^T\theta$ is linear, g, h are convex.

$f = \sum_{i=1}^{N} y_i g(L_i(\theta)) + (1-y_i) h(L_i(\theta))$.

$\theta \mapsto g(L_i(\theta))$ is convex, $\theta \mapsto h(L_i(\theta))$ is convex.

$y_i \ge 0$, $1-y_i \ge 0$. Thus, f is a positive linear combination of convex fns and so convex.