Anar Rzayev Homework Set 1 ID: 20180788

2) a) Taking into account that Beta distribution is normalized, it implies $\int_0^1 Beta(\mu|a,b)d\mu = 1$ where

$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$. We'll prove the normalization of Dirichlet distribution using induction principle. Since Beta distribution is a special case of Dirichlet for $M=2$, and considering Beta distribution is normalized, we can assume that Dirichlet distribution is normalized for $(M-1)$ variables and will guarantee to show that it is also normalized for M variables.

Consider Dirichlet distribution over M variables. As we have $\sum_{k=1}^{M} \mu_k = 1 \Rightarrow$ We write $p_M(\mu_1,\dots,\mu_{M-1}) = D_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k - 1} \left(1 - \sum_{j=1}^{M-1} \mu_j\right)^{\alpha_M - 1}$ for Dirichlet

When integrating over μ_{M-1}, lower limit of integration $= 0$, upper limit of integration $= 1 - \sum_{j=1}^{M-2} \mu_j$. Therefore, we get

$p_{M-1}(\mu_1,\dots,\mu_{M-2}) = \int_0^{1-\sum_{j=1}^{M-2}\mu_j} p_M(\mu_1,\dots,\mu_{M-1})\, d\mu_{M-1} =$

$= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \int_0^{1-\sum_{j=1}^{M-2}\mu_j} \mu_{M-1}^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-1} \mu_j\right)^{\alpha_M - 1} d\mu_{M-1}$

changing integration variables from μ_{M-1} to t where

$$t = \frac{\mu_{M-1}}{1-\sum_{j=1}^{M-2}\mu_j} \Rightarrow \mu_{M-1} = t\left(1-\sum_{j=1}^{M-2}\mu_j\right) \text{ and}$$

$$\left(\mu_{M-1} \in \left[0, 1-\sum_{j=1}^{M-2}\mu_j\right] \Rightarrow t \in [0,1]\right)$$

$$p_{M-1}(\mu_1,\dots,\mu_{M-2}) = D_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1} \cdot$$

$$\cdot \int_0^1 t^{\alpha_{M-1}-1}(1-t)^{\alpha_M-1}\,dt. \text{ Since Beta distribution is normalized,}$$

$$\int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1}\,d\mu = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \Rightarrow \int_0^1 t^{\alpha_{M-1}-1}(1-t)^{\alpha_M-1} =$$

$$= \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \Rightarrow p_{M-1}(\mu_1,\dots,\mu_{M-2}) =$$

$$= D_M \prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1} \cdot \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \quad (\star)$$

RHS is normalized Dirichlet distribution of $M-1$ variables, where coefficients $\rightsquigarrow \alpha_1,\dots,\alpha_{M-2},\alpha_{M-1}+\alpha_M$ (integrating 2 last groups) and from the formula, $\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$ where $\alpha_0 = \sum_{k=1}^{K}\alpha_k$

By determining normalization coefficient from previous formula and $(\star)$, we obtain

$$D_M = \frac{\Gamma(\alpha_1 + \dots + \alpha_M)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_{M-2}) \Gamma(\alpha_{M-1} + \alpha_M)} \cdot \frac{\Gamma(\alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_{M-1}) \Gamma(\alpha_M)} =$$

$$= \frac{\Gamma(\alpha_1 + \dots + \alpha_M)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_{M-2}) \Gamma(\alpha_{M-1}) \Gamma(\alpha_M)} \quad \text{as desired } \boxed{V} \blacksquare$$

$\Rightarrow$ $\boxed{\text{Dir}(\alpha) \text{ is normalized}}$

b) Based on the definition of Expectation and formula

$$\text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \quad \text{where } \alpha_0 = \sum_{k=1}^{K} \alpha_k \Rightarrow$$

$$E[\mu_j] = \int \mu_j \, \text{Dir}(\mu|\alpha) \, d\mu = \int \mu_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\mu =$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \int \mu_j \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\mu = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \cdot \Gamma(\alpha_1) \dots \Gamma(\alpha_{j-1}) \cdot \Gamma(\alpha_{j+1}) \dots \Gamma(\alpha_K) \cdot \Gamma(\alpha_j + 1)$$

$$= \frac{\Gamma(\alpha_0) \Gamma(\alpha_j + 1)}{\Gamma(\alpha_j) \Gamma(\alpha_0 + 1)} = \frac{\Gamma(\alpha_0) \alpha_j \Gamma(\alpha_j)}{\Gamma(\alpha_j) \alpha_0 \Gamma(\alpha_0)} = \boxed{\frac{\alpha_j}{\alpha_0}} \Rightarrow \boxed{E[\mu_j] = \frac{\alpha_j}{\alpha_0}} \quad \frac{\Gamma(\alpha_0 + 1)}{\Gamma(\alpha_0 + 1)}$$

For variance, we calculate $E[\mu_j^2] \Rightarrow E[\mu_j^2] = \int \mu_j^2 \, \text{Dir}(\mu|\alpha) d\mu$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \Gamma(\alpha_2) \dots \Gamma(\alpha_K)} \int \mu_j^2 \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\mu =$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \cdot \frac{\Gamma(\alpha_1) \dots \Gamma(\alpha_{j-1}) \Gamma(\alpha_j + 2) \Gamma(\alpha_{j+1}) \dots \Gamma(\alpha_K)}{\Gamma(\alpha_0 + 2)} =$$

$$= \frac{\Gamma(\alpha_0)\,\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)\,\Gamma(\alpha_0+2)} = \frac{\Gamma(\alpha_0)(\alpha_j+1)\alpha_j\Gamma(\alpha_j)}{\Gamma(\alpha_j)(\alpha_0+1)\alpha_0\Gamma(\alpha_0)} = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} \Rightarrow$$

$E[\mu_j^2] = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)}$ and from previous result, $E[\mu_j] = \frac{\alpha_j}{\alpha_0}$

$$\Rightarrow \mathrm{var}[\mu_j] = E[\mu_j^2] - E[\mu_j]^2 = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j^2}{\alpha_0^2} =$$

$$= \frac{\alpha_j(\alpha_j+1)\alpha_0 - \alpha_j^2(\alpha_0+1)}{\alpha_0^2(\alpha_0+1)} = \frac{\alpha_j\alpha_0 - \alpha_j^2}{\alpha_0^2(\alpha_0+1)} \Rightarrow \boxed{\mathrm{var}[\mu_j] = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}}$$

Since $\mathrm{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])] = E[X_iX_j] -$

$- E[X_i]E[X_j] = E[X_iX_j] - \frac{\alpha_i}{\alpha_0}\frac{\alpha_j}{\alpha_0}$ from previous results

$$E[X_iX_j] = \int_{\Delta^{k-1}} X_iX_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)} X_1^{\alpha_1-1}\dots X_k^{\alpha_k-1}\,dX =$$

$$= \int_{\Delta^{k-1}} \frac{\Gamma(\alpha_0+2)}{\alpha_0(\alpha_0+1)} \frac{\alpha_0(\alpha_0+1)}{\Gamma(\alpha_1)\dots\frac{\Gamma(\alpha_j+1)}{\alpha_j}\dots\frac{\Gamma(\alpha_i+1)}{\alpha_i}\dots\Gamma(\alpha_k)} X_1^{\alpha_1-1}\dots X_j^{\alpha_j}\dots X_i^{\alpha_i}\dots X_k^{\alpha_k-1}\,dX$$

$$= \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)} \int_{\Delta^{k-1}} \frac{\Gamma(\alpha_0+2)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_j+1)\dots\Gamma(\alpha_i+1)\dots\Gamma(\alpha_k)} X_1^{\alpha_1-1}\dots X_j^{\alpha_j}\dots X_i^{\alpha_i}\dots X_k^{\alpha_k-1}\,dX$$

$= \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}$ since Dirichlet distribution is normalized, where $\Gamma(\alpha_0+2) = \alpha_0(\alpha_0+1)\Gamma(\alpha_0)$ and $\Gamma(\alpha_i+1) = \alpha_i\Gamma(\alpha_i)$

$$\frac{\partial\left(-\frac{1}{2}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right)}{\partial\mu} = -\frac{1}{2}(x_j-\mu)^T\left(\Sigma^{-1}+(\Sigma^{-1})^T\right)\cdot\frac{\partial(x_j-\mu)}{\partial\mu} =$$

$$= \frac{1}{2}(x_j-\mu)^T\left(\Sigma^{-1}+(\Sigma^{-1})^T\right) = \frac{1}{2}(x_j-\mu)^T(2\Sigma^{-1}) =$$

$= (x_j-\mu)^T\Sigma^{-1}$, considering the fact $\Sigma^{-1}\rightsquigarrow$ positive definite such that $\Sigma^{-1} = (\Sigma^{-1})^T$ and independent of μ ($x_j\rightsquigarrow$ independent of μ)

(Σ - positive definite $\Rightarrow \Sigma^{-1}\rightsquigarrow$ positive definite)

Since $\frac{\partial \log(L)}{\partial\mu} = 0$ for finding MLE of $\mu \Rightarrow$ should satisfy

$\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1} = 0$, and $x_j-\mu\rightsquigarrow$ vector columns $\Rightarrow$

$\sum_{j=1}^{n}(x_j-\mu) = 0$, $\Rightarrow$ $\boxed{\mu_{MLE} = \frac{x_1+x_2+\dots+x_n}{n}}$ (Y).

b) It's well-known that posterior distribution $P(\mu|X) \propto P(X|\mu)\,P(\mu)$ or equivalently,

$N(\mu|X,\Sigma) \propto N(X|\mu,\Sigma)\,N(\mu|\mu_0,\Sigma_0)$

where considering $N(X|\mu,\Sigma) = \prod_{j=1}^{n} N(x_j|\mu,\Sigma)$

$\Rightarrow N(X|\mu,\Sigma)\,N(\mu|\mu_0,\Sigma_0)$ becomes identical to

Thus, $\boxed{E[x_i x_j] = \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)}} \Rightarrow Cov(x_i, x_j) = E[x_i x_j] - \frac{\alpha_i \alpha_j}{\alpha_0^2} =$

$= \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)} - \frac{\alpha_i \alpha_j}{\alpha_0^2} = \frac{\alpha_i \alpha_j (\alpha_0 - \alpha_0 - 1)}{\alpha_0^2(\alpha_0+1)} = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)} \Rightarrow \boxed{Cov(x_i, x_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}}$

(V) • (+)

3) a) We consider the likelihood function where product of n random variables $X_1, X_2, \dots, X_n$ is taken

$$\mathcal{L}(X_1, X_2, \dots X_n | \mu) = \mathcal{N}(X_1 | \mu, \Sigma)\, \mathcal{N}(X_2 | \mu, \Sigma) \dots$$

$$\mathcal{N}(X_n | \mu, \Sigma) = \left(\frac{1}{(2\pi)^{K/2}}\right)^n \cdot \left(\frac{1}{|\Sigma|^{1/2}}\right)^n \cdot \exp\left(-\frac{1}{2}\sum_{j=1}^{n} (x_j - \mu)^T \Sigma^{-1} (x_j - \mu)\right)$$

To compute MLE, we use $\log(\mathcal{L}(X_1, \dots, X_n | \mu))$ $\Rightarrow$

$$\log(\mathcal{L}) = -\frac{nK}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n} (x_j - \mu)^T \Sigma^{-1} (x_j - \mu)$$

Since Σ and π do not depend on μ, we'll take derivative of $\log(\mathcal{L})$ wrt μ to obtain μ_{MLE}. In order to do that, observe from lecture notes that $\frac{\partial (x^T A x)}{\partial x} = x^T (A + A^T)$ where $A \sim$ independent of x

Thus, $\frac{\partial\left(-\frac{nK}{2}\log(2\pi)\right)}{\partial \mu} = \frac{\partial\left(-\frac{n}{2}\log|\Sigma|\right)}{\partial \mu} = 0$ and

as $\frac{\partial (x^T A x)}{\partial \mathcal{I}} = x^T (A + A^T) \frac{\partial x}{\partial \mathcal{I}}$ (where $A \sim$ indep. of $\mathcal{I}$, $x \sim$ dependent of $\mathcal{I}$)

$$\left(\frac{1}{(2\pi)^{k/2}}\right)^{n+1} \cdot \left(\frac{1}{|\Sigma|^{1/2}}\right)^{n+1} \cdot \left(\frac{1}{|\Sigma_0|^{1/2}}\right)^{n+1} \cdot \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right)$$

From lecture notes, we know $\mu_{MAP} = \text{argmax}_\mu \left(\log\left(p(X|\mu)p(\mu)\right)\right)$ and using the same derivative/log strategy from last problem

$$\log\left(p(\mu|X)\right) = \log\left(p(X|\mu)p(\mu)\right) = -\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0) - \frac{n+1}{2}\log(|\Sigma_0|) - \frac{n+1}{2}\log(|\Sigma|) - \frac{(n+1)k}{2}\log 2\pi$$

$\Rightarrow \frac{\partial\left(\log p(\mu|X)\right)}{\partial\mu} = 0$ should satisfy for obtaining MAP estimator of μ

$\frac{\partial}{\partial\mu}\left(-\frac{n+1}{2}\log|\Sigma_0| - \frac{n+1}{2}\log(|\Sigma|) - \frac{(n+1)k\log 2\pi}{2}\right) = 0$ where $\frac{\partial\left(\frac{-1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right)}{\partial\mu} = \sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}$ had been proved in previous section

along with $\frac{\partial\left(\frac{-1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\right)}{\partial\mu} = (\mu-\mu_0)^T(-1)\Sigma_0^{-1}$

$\Rightarrow \boxed{\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1} = (\mu-\mu_0)^T\Sigma_0^{-1}}$ becomes true.

$(A+B)^T = A^T + B^T$ Using this formula for LHS operations,

$\sum_{j=1}^{n}(x_j^T-\mu^T)\Sigma^{-1} = \left(\left(\sum_{j=1}^{n}x_j\right)^T - n\mu^T\right)\Sigma^{-1} = \left(\sum_{j=1}^{n}x_j\right)^T\Sigma^{-1} - n\mu^T\Sigma^{-1}$

$(\mu-\mu_0)^T\Sigma_0^{-1} = \mu^T\Sigma_0^{-1} - \mu_0^T\Sigma_0^{-1} \Rightarrow \left(\sum_{j=1}^{n} x_j\right)^T\Sigma^{-1} - n\mu^T\Sigma^{-1} =$

$= \mu^T\Sigma_0^{-1} - \mu_0^T\Sigma_0^{-1} \Rightarrow$ $\boxed{\mu^T\left(\Sigma_0^{-1} + n\Sigma^{-1}\right) = \left(\sum_{j=1}^{n} x_j\right)^T\Sigma^{-1} + \mu_0^T\Sigma_0^{-1}}$

Considering $n\Sigma^{-1} + \Sigma_0^{-1}$ is invertible, we find that

$\mu^T = \left(\sum_{j=1}^{n} x_j^T\Sigma^{-1} + \mu_0^T\Sigma_0^{-1}\right)\left(n\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}$ where $(AB)^T = B^TA^T \Rightarrow$

$\boxed{\mu_{MAP} = \left(\left(n\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1}\right)^T\left(\sum_{j=1}^{n} x_j^T\Sigma^{-1} + \mu_0^T\Sigma_0^{-1}\right)^T}$

1)

		True Label	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

Truth Table (confusion matrix)

score $> X \Rightarrow$ pred is positive

score $\leq X \Rightarrow$ pred is negative

a) $X=15 \Rightarrow$ TP=8 FP=5 FN=2 TN=35 $\rightarrow$ Sensitivity $\rightsquigarrow \frac{TP}{TP+FN}$ and

$\boxed{\text{Sensitivity} \rightsquigarrow \frac{8}{10} = \frac{4}{5}}$ $\boxed{\text{Specificity} \rightsquigarrow \frac{TN}{TN+FP} = \frac{35}{40} = \frac{7}{8}}$

b) Precision $\rightsquigarrow \frac{TP}{TP+FP}$ and Recall $\rightsquigarrow \frac{TP}{TP+FN}$

$F_1 \rightsquigarrow \frac{2TP}{2TP+FP+FN} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$

$X=5 \Rightarrow$ TP=10 FP=84 FN=0 TN=6 $\rightarrow$ $\boxed{F_1 = \frac{20}{54} = \frac{10}{27}}$ ✓

$X=10 \Rightarrow TP=9 \quad FP=14$
$FN=1 \quad TN=26$ $\Rightarrow \boxed{F_1 = \frac{18}{33} = \frac{6}{11}}$ ✓

$X=15 \Rightarrow TP=8,\ FN=2,\ FP=5,\ TN=35 \Rightarrow \boxed{F_1 = \frac{16}{23}}$ ✓

$X=20 \Rightarrow TP=5,\ FN=5,\ FP=1 \Rightarrow \boxed{F_1 = \frac{10}{16} = \frac{5}{8}}$ ✓

Finding the best value for x means to choose highest precision and recall $\rightsquigarrow$ highest F_1-score

$\frac{16}{23} > \frac{10}{27}, \quad \frac{16}{23} > \frac{6}{11}, \quad \frac{16}{23} > \frac{5}{8} \Rightarrow \boxed{\text{highest } F_1\text{-score} \rightsquigarrow \frac{16}{23}}$

and thus, it is achieved at $\boxed{X=15}$ ✱ ✓ ● (+)

5) $A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & & & \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix}$

a) $(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$ and $tr(AB) = \sum_{i=1}^{n} (AB)_{ii} =$

$= \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{ki} \Rightarrow$ According to definition in the problem statement

$y \in \mathbb{R}, \ X \in \mathbb{R}^{p \times q} \Rightarrow \frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \dots & \frac{\partial y}{\partial x_{1q}} \\ \vdots & & \\ \frac{\partial y}{\partial x_{p1}} & \dots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}^T$ $tr(AB) \in \mathbb{R}$

$A \rightsquigarrow n \times n$ matrix

We use numerator-layout notation

$\frac{\partial\, tr(AB)}{\partial A} = \begin{bmatrix} \frac{\partial\, tr(AB)}{\partial a_{11}} & \dots & \frac{\partial\, tr(AB)}{\partial a_{n1}} \\ \vdots & & \\ \frac{\partial\, tr(AB)}{\partial a_{1n}} & \dots & \frac{\partial\, tr(AB)}{\partial a_{nn}} \end{bmatrix}$ and using the $tr(AB)$ value from previous

$$\left[\frac{\partial\, tr(AB)}{\partial A}\right]_{ij} = \frac{\partial\, tr(AB)}{\partial a_{ji}} = \frac{\partial\left(\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki}\right)}{\partial a_{ji}}$$

In the summation $\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki}$, term containing a_{ji} coefficient (dot product of j^{th} row of A and i^{th} column of B) will only be $a_{ji} \times b_{ij} \Rightarrow \frac{\partial\left(\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki}\right)}{\partial a_{ji}} = \frac{\partial\,(a_{ji} \times b_{ij})}{\partial a_{ji}}$ according to indices rule

$= b_{ij}$, thus $\left[\frac{\partial\, tr(AB)}{\partial A}\right]_{(i^{th}\text{ row},\, j^{th}\text{ column})}{}_{ij} = b_{ij} \Rightarrow \boxed{\frac{\partial\, tr(AB) = B}{\partial A}}$ ✓

where $B = I \Rightarrow \frac{\partial\, tr(AI)}{\partial A} = \boxed{\frac{\partial}{\partial A} tr(A) = I}$ ✓

c) $\Rightarrow$ We have to prove $\frac{\partial}{\partial A} \log|A| = A^{-1}$ where $|A| > 0$

According to chain rule, $\frac{\partial}{\partial A} \log|A| = \frac{1}{|A|} \cdot \frac{\partial\,|A|}{\partial A}$

It's a well-known fact that $A^{-1} = \frac{1}{|A|} \cdot [adj(A)]^T$. Thus, we have to prove $\Rightarrow \frac{\partial\,|A|}{\partial A} = adj(A)$ ✓

According to the definition, if we denote C_{ij} by $(-1)^{i+j} \det\left|\begin{matrix}\text{matrix after removal of } i^{th}\text{ row} \\ \text{and } j^{th}\text{ column}\end{matrix}\right|$, then adjugate is the transpose of its cofactor matrix

$$C=\begin{bmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & & & \\ C_{n1} & C_{n2} & \dots & C_{nn} \end{bmatrix} \Rightarrow adj(A)=\begin{bmatrix} C_{11} & C_{21} & \dots & C_{n1} \\ C_{12} & C_{22} & \dots & C_{n2} \\ \vdots & & & \\ C_{1n} & C_{2n} & \dots & C_{nn} \end{bmatrix}$$

It's also well-known rule that $\det(A)=|A|=\sum_{t=1}^{n} a_{kt}C_{kt}$ along the k^{th} row. According to the definition of cofactor matrix, neither any of a_{jk} nor C_{jk} are dependent on a_{ji}, where the only exception is a_{ji} for $k=i$

$$|A|=\sum_{t=1}^{n} a_{jt}C_{jt} \text{ and } \frac{\partial |A|}{\partial a_{ji}}=\frac{\partial}{\partial a_{ji}}\left(a_{ji}C_{ji}\right)=C_{ji}$$

$[adj(A)]_{ij}=C_{ji}$ from the construction (i^{th} row, j^{th} column of $adj(A)$ is C_{ji})

Taking into account numerator-layout notation $\Rightarrow$

$$\left[\frac{\partial}{\partial A}|A|\right]_{ij}=\frac{\partial |A|}{\partial a_{ji}}, \text{ since } \frac{\partial}{\partial A}|A|=\begin{bmatrix} \frac{\partial |A|}{\partial a_{11}} & \dots & \frac{\partial |A|}{\partial a_{n1}} \\ \vdots & & \\ \frac{\partial |A|}{\partial a_{1n}} & \dots & \frac{\partial |A|}{\partial a_{nn}} \end{bmatrix}$$

(i^{th} row, j^{th} column)

As we showed $\frac{\partial}{\partial a_{ji}}|A|=C_{ji} \Rightarrow \left[\frac{\partial}{\partial A}|A|\right]_{ij}=C_{ji}$ (i^{th} row, j^{th} col of $\frac{\partial}{\partial A}|A|$ is C_{ji})

From the definition, $[adj(A)]_{ij}=C_{ji}$, concluding that

$$\boxed{\frac{\partial}{\partial A}|A|=adj(A)} \Rightarrow \text{Therefore, claim is satisfied and so,}$$

$$\boxed{\frac{\partial}{\partial A}\log|A|=A^{-1}}, \text{ where } \det(A)=|A|>0$$

(+) (V) ■

4) a) From famous Jensen's Inequality, it is evident that for real convex function φ, $\varphi(a_1x_1+\dots+a_mx_m) \le a_1\varphi(x_1)+ + a_2\varphi(x_2)+\dots+a_m\varphi(x_m)$ where $a_1+\dots+a_m=1$ and $a_i>0$

From the definition, $H(x) = -E[\log P(x)] = E\left[\log\left(\frac{1}{P(x)}\right)\right]$

$$= \sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{m} p_i \log p_i = \sum_{i=1}^{m} p_i(-\log p_i)$$

Take $\varphi(x) = \log\left(\frac{1}{x}\right)$, where $\varphi'(x) = -\frac{1}{x}$, $\varphi''(x) = \frac{1}{x^2} > 0 \Rightarrow$

$$\sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right) = \sum_{i=1}^{m} p_i \varphi(p_i) \ge \Rightarrow \varphi(x_i) = \log\left(\frac{1}{x_i}\right)$$

$$\sum_{i=1}^{m} p_i \log(p_i) = \sum_{i=1}^{m} p_i \log\left(\frac{1}{1/p_i}\right) = \sum_{i=1}^{m} p_i \varphi\left(\frac{1}{p_i}\right) \ge$$

$$\ge \varphi(1+1+\dots+1) = \varphi(m) = \log\left(\frac{1}{m}\right) = -\log m \Rightarrow$$ we find that

$$\sum_{i=1}^{m} p_i \log(p_i) \ge -\log m \Rightarrow -\sum_{i=1}^{m} p_i \log(p_i) \le \log m \text{ or just}$$

$$\log m \ge \sum_{i=1}^{m} p_i(-\log p_i) = \sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right) = -E[\log P(x)] = H(x)$$

In conclusion, $\boxed{H(x) \le \log m}$ ☒ ✓ ■

Note: Discrete p_i's are greater than 0 with $P(X=x_i)=p_i$ for $1 \le i \le m$

b) From previously solved problem 8, it's easy to observe that univariate Gaussian distributions are

$$\mathcal{N}(m, s^2) = \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{|s|} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{s^2}\right)$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{|\sigma|} \cdot \exp\left(\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Therefore, $D_{KL}\left(\mathcal{N}(\mu, \sigma^2) \,||\, \mathcal{N}(m, s^2)\right) =$

$$= E\left[\log\left(\frac{\mathcal{N}(x|\mu, \sigma^2)}{\mathcal{N}(x|m, s^2)}\right)\right] = E\left[\log\left(\mathcal{N}(x|\mu, \sigma^2)\right)\right]$$

$$- E\left[\log\left(\mathcal{N}(x|m, s^2)\right)\right] = E\left[\log\left(\frac{|s|}{|\sigma|}\exp\left(\frac{-1}{2}\left(\frac{(x-\mu)^2}{\sigma^2} - \frac{(x-m)^2}{s^2}\right)\right)\right)\right]$$

$D_{KL}\left(\mathcal{N}(\mu, \sigma^2) \,||\, \mathcal{N}(m, s^2)\right) =$

$$= E\left[\log\left(\frac{|s|}{|\sigma|}\exp\left(\frac{1}{2}\left(\frac{(x-m)^2}{s^2} - \frac{(x-\mu)^2}{\sigma^2}\right)\right)\right)\right] =$$

$$= E\left[\log\frac{|s|}{|\sigma|}\right] + \frac{1}{2}E\left[\frac{(x-m)^2}{s^2} - \frac{(x-\mu)^2}{\sigma^2}\right] =$$

$$= \log\frac{|s|}{|\sigma|} + \frac{1}{2s^2}E\left[(x-m)^2\right] - \frac{1}{2\sigma^2}E\left[(x-\mu)^2\right]$$

$\mathrm{Var}(X) = E[X^2] - \mu^2 \Rightarrow E[X^2] = \mu^2 + \sigma^2$, $E\left[x^2 - 2mx + m^2\right] =$

$= \mu^2 + \sigma^2 - 2m\mu + m^2$, similarly $E\left[(x-\mu)^2\right] = \mu^2 + \sigma^2 -$

$-2\mu^2 + \mu^2 = \sigma^2$ after plugging $m = \mu$ in general form

So, answer is $\log\frac{|s|}{|G|} + \frac{1}{2s^2}(\mu^2+\sigma^2-2m\mu+m^2) - \frac{1}{2\sigma^2}\cdot\sigma^2 = \log\frac{|s|}{|G|} - \frac{1}{2} + \frac{1}{2s^2}(\mu^2+\sigma^2-2m\mu+m^2)$

Ⓥ • Ⓣ