CS 6190: Probabilistic Machine Learning Spring 2022

Homework 1

Meysam Alishahi (U1323606)

Handed out: 1 Feb, 2022
Due: 11:59pm, 18 Feb, 2022

Analytical problems [80 points + 30 bonus]

1. [8 points] A random vector, $\mathbf{x} = \left[\begin{array}{c}\mathbf{x}_1\\ \mathbf{x}_2\end{array}\right]$ follows a multivariate Gaussian distribution,

$$p(\mathbf{x}) = \mathcal{N}\left(\left[\begin{array}{c}\mathbf{x}_1\\ \mathbf{x}_2\end{array}\right] \middle| \left[\begin{array}{c}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{array}\right]\right).$$

Show that the marginal distribution of $\mathbf{x}_1$ is $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Answer: Set $\boldsymbol{\Sigma} = \left[\begin{array}{cc}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{array}\right]$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda} = \left[\begin{array}{cc}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12}\\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{array}\right]$. In the class, it was proved that

$$p(\mathbf{x}_2|\mathbf{x}_1) = \mathcal{N}\big(\mathbf{x}_2|\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}\big),$$

where $\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Lambda}_{22}^{-1}$ and $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 - \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)$. Using Bayes' Rule,

$$p(\mathbf{x}_1) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_2|\mathbf{x}_1)} = C\exp\left\{-\frac{1}{2}(A - B)\right\}, \tag{1}$$

where C is a constant with respect to $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$A = \left[(\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top, (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top\right]\left[\begin{array}{cc}\boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12}\\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{array}\right]\left[\begin{array}{c}\mathbf{x}_1 - \boldsymbol{\mu}_1\\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{array}\right]$$

and $B = (\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top\boldsymbol{\Sigma}_{2|1}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})$. So, to know $p(\mathbf{x}_1)$, we need to compute $A - B$. Expanding A, we have

$$\begin{aligned} A =& (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Lambda}_{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\\ &+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Lambda}_{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Lambda}_{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2).\end{aligned}$$

We saw in the class that $p(\mathbf{x}_2|\mathbf{x}_1) = \mathcal{N}\big(x_2|\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}\big)$, where $\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Lambda}_{22}^{-1}$ and $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 - \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)$. Also, expanding B and also using the facts that $\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Lambda}_{22}^{-1}$ and $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 - \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)$, we conclude

$$\begin{aligned}
B &= (\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top \boldsymbol{\Sigma}_{2|1}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_{2|1}) \\
&= \Big(\mathbf{x}_2 - \boldsymbol{\mu}_2 + \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)\Big)^\top \boldsymbol{\Lambda}_{22}\Big(\mathbf{x}_2 - \boldsymbol{\mu}_2 + \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)\Big) \\
&= (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Lambda}_{22} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Lambda}_{22}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1) \\
&+ (x_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Lambda}_{21}^\top\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&+ (x_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Lambda}_{21}^\top\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{22}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1) \\
&= (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Lambda}_{22} (\mathbf{x}_2 - \boldsymbol{\mu}_2) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1) \\
&+ (x_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Lambda}_{21}^\top(\mathbf{x}_2 - \boldsymbol{\mu}_2) + (x_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Lambda}_{21}^\top\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(x_1 - \boldsymbol{\mu}_1)
\end{aligned}$$

Therefore,

$$\begin{aligned}
A - B &= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \underbrace{\big(\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{21}^\top\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\big)}_{=\boldsymbol{\Sigma}_{11}^{-1}}(x_1 - \boldsymbol{\mu}_1) \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)
\end{aligned}$$

Substituting in 1,

$$p(\mathbf{x}_1) = C\exp\left\{-\frac{1}{2}\Big((\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\Big)\right\}.$$

Since $p(\mathbf{x}_1)$ is a pdf, C should be $|2\pi\Sigma_{11}|^{-1}$ and consequently, $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ as desired.

2. [**Bonus**][10 points] Given a Gaussian random vector, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We have a linear transformation, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{z}$, where $\mathbf{A}$ and $\mathbf{b}$ are constants, $\mathbf{z}$ is another Gaussian random vector independent to $\mathbf{x}$, $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \boldsymbol{\Lambda})$. Show $\mathbf{y}$ follows Gaussian distribution as well, and derive its form. Hint: using characteristic function. You need to check the materials by yourself.
Answer: First note that since since $\mathbf{x}$ and $\mathbf{z}$ are independent, $\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $\mathbf{z}$ are independent as well. We remind that the characteristic function of a random variable $\mathbf{x}$ with probability distribution function $p(\mathbf{x})$ is defined as

$$\phi_{\mathbf{x}}(\mathbf{t}) = \mathbb{E}_p(e^{i\mathbf{t}^\top \mathbf{x}}).$$

Also, it is a known fact that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if $\phi_{\mathbf{x}}(\mathbf{t}) = e^{\mathbf{t}^\top(i\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\Sigma}\mathbf{t})}$. Let us compute $\phi_{\mathbf{y}}(\mathbf{t})$.

$$\begin{aligned}
\phi_{\mathbf{y}}(\mathbf{t}) &= \mathbb{E}_y(e^{i\mathbf{t}^\top\mathbf{y}}) = \mathbb{E}_{\mathbf{y}}\left(e^{i\mathbf{t}^\top(\mathbf{A}\mathbf{x}+\mathbf{b}+\mathbf{z})}\right) \\
&= e^{i\mathbf{t}^\top\mathbf{b}}\mathbb{E}_{\mathbf{x}}\left(e^{i\mathbf{t}^\top\mathbf{A}\mathbf{x}}\right)\mathbb{E}_{\mathbf{z}}\left(e^{i\mathbf{t}^\top\mathbf{z}}\right) && \text{since } \mathbf{A}\mathbf{x} \text{ and } \mathbf{z} \text{ are independent and } \mathbf{b} \text{ is constant} \\
&= e^{i\mathbf{t}^\top\mathbf{b}}e^{\mathbf{t}^\top(i\mathbf{A}\boldsymbol{\mu} - \frac{1}{2}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\mathbf{t})}e^{\mathbf{t}^\top(-\frac{1}{2}\boldsymbol{\Lambda}\mathbf{t})} && \text{since } \mathbf{A}\mathbf{x} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top) \text{ and } \mathbf{z} \sim \mathcal{N}(0, \boldsymbol{\Lambda}) \\
&= e^{\mathbf{t}^\top\left(i(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}) - \frac{1}{2}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \boldsymbol{\Lambda})\mathbf{t}\right)}.
\end{aligned}$$

Therefore, $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \boldsymbol{\Lambda})$.

3. [8 points] Show the differential entropy of the a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\mathrm{H}[\mathbf{x}] = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}(1 + \log 2\pi)$$

where d is the dimension of $\mathbf{x}$. Answer: If $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $f(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$.

Accordingly,

$$
\begin{aligned}
H[\mathbf{x}] &= -\int_{x\in\mathbb{R}^d} f(\mathbf{x})\log f(\mathbf{x})\mathrm{d}\mathbf{x} = -\int_{x\in\mathbb{R}^d} f(\mathbf{x})\left[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\mathrm{d}\mathbf{x}\\
&= \int_{x\in\mathbb{R}^d} f(\mathbf{x})\left[\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\mathrm{d}\mathbf{x}\\
&= \frac{d}{2}\log 2\pi \underbrace{\int_{x\in\mathbb{R}^d} f(\mathbf{x})\mathrm{d}\mathbf{x}}_{=1 \text{ because } f \text{ is a pdf}} + \frac{1}{2}\log|\boldsymbol{\Sigma}| \underbrace{\int_{x\in\mathbb{R}^d} f(\mathbf{x})\mathrm{d}\mathbf{x}}_{=1 \text{ because } f \text{ is a pdf}} + \frac{1}{2}\int_{x\in\mathbb{R}^d} f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathrm{d}\mathbf{x}\\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right]\\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\right)\right]\\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{tr}\left(\underbrace{\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]}_{=\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right)\\
&= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathrm{tr}\left(\mathbf{I}_{d\times d}\right) = \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}.
\end{aligned}
$$

4. [8 points] Derive the Kullback-Leibler divergence between two Gaussian distributions, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m},\Lambda)$, i.e., $\mathrm{KL}(q||p)$. Answer: Lets remind that

$$
\begin{aligned}
\mathrm{KL}(q||p) &= \int_{x\in\mathbb{R}^d} q(\mathbf{x})\log\frac{q(\mathbf{x})}{p(\mathbf{x})}\mathrm{d}\mathbf{x} = \underbrace{\int_{x\in\mathbb{R}^d} q(\mathbf{x})\log q(\mathbf{x})\mathrm{d}\mathbf{x}}_{=-H_p[\mathbf{x}],\text{ See question 3.}} - \underbrace{\int_{x\in\mathbb{R}^d} q(\mathbf{x})\log p(\mathbf{x})\mathrm{d}\mathbf{x}}_{\text{is computed in the following}}\\
&= -\frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{d}{2}(1+\log 2\pi) + (\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) + \frac{1}{2}\mathrm{tr}\left(\left[\Lambda + \mathbf{m}\mathbf{m}^\top - \mathbf{m}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{m}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right]\boldsymbol{\Sigma}^{-1}\right)\\
&= \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{d}{2} - \frac{1}{2}\log|\Lambda| + \frac{1}{2}\mathrm{tr}\left(\left[\Lambda + \mathbf{m}\mathbf{m}^\top - \mathbf{m}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{m}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right]\boldsymbol{\Sigma}^{-1}\right).\\
&= \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}|}{|\Lambda|} - \frac{d}{2} + \frac{1}{2}\left(\mathrm{tr}(\Lambda\boldsymbol{\Sigma}^{-1}) + \mathbf{m}^\top\boldsymbol{\Sigma}^{-1}\mathbf{m} - 2\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mathbf{m} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)
\end{aligned}
$$

So, we need to compute $\int_{x\in\mathbb{R}^d} q(\mathbf{x})\log p(\mathbf{x})\mathrm{d}\mathbf{x}$.

$$
\begin{aligned}
\int_{x\in\mathbb{R}^d} q(\mathbf{x})\log p(\mathbf{x})\mathrm{d}\mathbf{x} &= -\int_{x\in\mathbb{R}^d} q(\mathbf{x})\left[\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\mathrm{d}\mathbf{x}\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|)\underbrace{\int_{x\in\mathbb{R}^d} q(\mathbf{x})\mathrm{d}\mathbf{x}}_{=1} - \frac{1}{2}\underbrace{\int_{x\in\mathbb{R}^d} q(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^\top\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathrm{d}\mathbf{x}}_{=\mathbb{E}_q\left[(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathbb{E}_q\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right]\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathbb{E}_q\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\right)\right]\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathrm{tr}\left(\mathbb{E}_q\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\boldsymbol{\Sigma}^{-1}\right)\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathrm{tr}\left(\mathbb{E}_q\left[\mathbf{x}\mathbf{x}^\top - \mathbf{x}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{x}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right]\boldsymbol{\Sigma}^{-1}\right)\\
&= -(\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathrm{tr}\left(\left[\Lambda + \mathbf{m}\mathbf{m}^\top - \mathbf{m}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{m}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right]\boldsymbol{\Sigma}^{-1}\right)
\end{aligned}
$$

5. [8 points] Given a distribution in the exponential family,

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big).$$

Show that

$$\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = \mathrm{cov}(\mathbf{u}(\mathbf{x})),$$

where cov is the covariance matrix. Answer: Notice $Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathrm{d}\mathbf{x}$. Consequently,

$$\begin{aligned}
\frac{\partial \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{1}{Z(\boldsymbol{\eta})} \frac{\partial Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{1}{Z(\boldsymbol{\eta})} \frac{\partial}{\partial \boldsymbol{\eta}} \int h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathrm{d}\mathbf{x} \\
&= \frac{1}{Z(\boldsymbol{\eta})} \int \frac{\partial}{\partial \boldsymbol{\eta}} \Big(h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big)\Big) \mathrm{d}\mathbf{x} = -\frac{1}{Z(\boldsymbol{\eta})} \int h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= - \int \underbrace{\frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big)}_{=p(\mathbf{x}|\boldsymbol{\eta})} \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} = -\mathbb{E}(\mathbf{u}(\mathbf{x})).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} &= -\frac{\partial}{\partial \boldsymbol{\eta}} \left(\frac{1}{Z(\boldsymbol{\eta})} \int h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \right) \\
&= -\frac{1}{Z(\boldsymbol{\eta})^2} \underbrace{\frac{\partial Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}}_{=\int h(\mathbf{x}) \exp\big(-\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x}} \int h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \\
&+ \frac{1}{Z(\boldsymbol{\eta})} \int \frac{\partial}{\partial \boldsymbol{\eta}} \Big(h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \Big) \mathrm{d}\mathbf{x} \\
&= - \left[\int \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} \right] \left[\int \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \right] \\
&+ \int \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(- \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \\
&= - \left[\int p(\mathbf{x}|\boldsymbol{\eta}) \mathbf{u}(\mathbf{x}) \mathrm{d}\mathbf{x} \right] \left[\int p(\mathbf{x}|\boldsymbol{\eta}) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \right] + \int p(\mathbf{x}|\boldsymbol{\eta}) \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top \mathrm{d}\mathbf{x} \\
&= -\mathbb{E}(\mathbf{u}(\mathbf{x})) \mathbb{E}(\mathbf{u}(\mathbf{x}))^\top + \mathbb{E}(\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top) = \mathrm{cov}(\mathbf{u}(\mathbf{x}))
\end{aligned}$$

6. [4 points] Is $\log Z(\boldsymbol{\eta})$ convex or nonconvex? Why? Answer: It is convex since the covariance matrix of any random variable is positive semi-definite.
Observation. *If* $\mathbf{z}$ *is a random variable, then* $\mathrm{cov}(\mathbf{z})$ *is positive semi-definite.*
Proof. Let $\mathbf{y}$ be an arbitrary vector, we need to prove that $\mathbf{y}^t \mathrm{cov}(\mathbf{z}) \mathbf{y} \geq 0$. To this end;

$$\begin{aligned}
\mathbf{y}^t \mathrm{cov}(\mathbf{z}) \mathbf{y} &= \mathbf{y}^t \mathbb{E}\big(\mathbf{z}\mathbf{z}^\top\big) \mathbf{y} = \mathbb{E}\big(\mathbf{y}^\top \mathbf{z}\mathbf{z}^\top \mathbf{y}\big) \\
&= \mathbb{E}\left[(\mathbf{z}^\top \mathbf{y})(\mathbf{z}^\top \mathbf{y})\right] = \mathbb{E}\left[(\mathbf{z}^\top \mathbf{y})^2\right] \geq 0,
\end{aligned}$$

since $(\mathbf{z}^\top \mathbf{y})^2 \geq 0$ and expectation of any nonnegative random variable is nonnegative.

7. [8 points] Given two random variables $\mathbf{x}$ and $\mathbf{y}$, show that

$$I(\mathbf{x}, \mathbf{y}) = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}]$$

where $I(\cdot,\cdot)$ is the mutual information and $H[\cdot]$ the entropy.
Answer:

$$
\begin{aligned}
\mathbf{I}[\mathbf{x},\mathbf{y}] &= \mathrm{KL}\Big(p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})\Big) = -\int\int p(\mathbf{x},\mathbf{y})\ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= -\int\int p(\mathbf{x},\mathbf{y})\ln\left(\frac{p(\mathbf{x})\cancel{p(\mathbf{y})}}{p(\mathbf{x}|\mathbf{y})\cancel{p(\mathbf{y})}}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = -\int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \underbrace{\int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}}_{=-H[\mathbf{x}|\mathbf{y}]} \\
&= -\int \underbrace{\left[\int p(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y}\right]}_{=p(\mathbf{x})}\ln p(\mathbf{y})\mathrm{d}\mathbf{x} - H[\mathbf{x}|\mathbf{y}] = \underbrace{-\int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x}}_{=H[\mathbf{x}]} - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}].
\end{aligned}
$$

8. [24 points] Convert the following distributions into the form of the exponential-family distribution. Please give the mapping from the expectation parameters to the natural parameters, and also represent the log normalizer as a function of the natural parameters.

- Dirichlet distribution
- Gamma distribution
- Wishart distribution

Answer: Any distribution can be written as the following form is a member of exponential family;

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})}h(\mathbf{x})\exp\big(-\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\big).$$

Dirichlet distribution. Let us remind that for $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_d)$, and $\alpha_0 = \sum_{i=1}^d \alpha_i$, Dirichlet distribution $\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha})$ is a probability density function over $(d-1)$-simplex (i.e., $\mathbf{x} = (x_1,\ldots,\mathbf{x}_d) \in \Delta_{d-1}$) with the following density function;

$$
\begin{aligned}
\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) &= \underbrace{\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_d)}}_{=C(\boldsymbol{\alpha})}\prod_{i=1}^d x_i^{\alpha_k-1} = C(\boldsymbol{\alpha})\exp\left(\sum_{i=1}^d(\alpha_i-1)\ln x_i\right) \\
&= C(\boldsymbol{\alpha})\frac{1}{\prod_{i=1}^d x_i}\exp\left(\sum_{i=1}^d \alpha_i\ln x_i\right) = \frac{1}{Z(\boldsymbol{\eta})}h(\mathbf{x})\exp\left(-\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right),
\end{aligned}
$$

where

$$h(\mathbf{x}) = \frac{1}{\prod_{i=1}^d x_i}, \qquad \boldsymbol{\eta} = (\alpha_1,\ldots,\alpha_d)^\top,$$

$$\mathbf{u}(\mathbf{x}) = (-\ln x_1,\ldots,-\ln x_d)^\top, \qquad Z(\boldsymbol{\eta}) = \frac{\Gamma\left(\sum_{i=1}^d \boldsymbol{\eta}_i\right)}{\Gamma(\boldsymbol{\eta}_1)\cdots\Gamma(\boldsymbol{\eta}_d)}.$$

Gamma distribution.

$$
\begin{aligned}
\mathrm{Gam}(x|a,b) &= \frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx) \qquad \text{where } a>0, b>0 \\
&= \frac{1}{\Gamma(a)}b^a x^{-1}\exp(-bx + a\ln x) = \frac{1}{\Gamma(a)b^{-a}}\frac{1}{x}\exp(-[a,b][-\ln x, x]^\top) \\
&= \frac{1}{Z(\boldsymbol{\eta})}h(x)\exp\left(-\mathbf{u}(x)^\top\boldsymbol{\eta}\right),
\end{aligned}
$$

where $\quad h(x) = \frac{1}{x}, \qquad \boldsymbol{\eta} = [a,b], \qquad \mathbf{u}(x) = [-\ln x, x], \qquad Z(\boldsymbol{\eta}) = \Gamma(a)b^{-a} = \Gamma(\boldsymbol{\eta}_1)\boldsymbol{\eta}_2^{-\boldsymbol{\eta}_1}.$

Wishart distribution. Let $\mathbf{X}$ be a $d \times d$ symmetric matrix of random variables that is positive definite. If $\nu \geq d$, we say $\mathbf{X}$ has a Wishart distribution with ν degrees of freedom if it has the probability density function

$$\omega(\mathbf{X}|\mathbf{W},\nu) = \frac{|\mathbf{X}|^{\frac{\nu-d-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}\mathbf{X}\right)\right)}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\frac{\nu}{2}}\Gamma_d(\frac{\nu}{2})},$$

where $\Gamma_d(\cdot)$ is the multivariate gamma function.

$$\begin{aligned}\omega(\mathbf{X}|\mathbf{W},\nu) &= \frac{|\mathbf{X}|^{\frac{\nu-d-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}\mathbf{X}\right)\right)}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\frac{\nu}{2}}\Gamma_d(\frac{\nu}{2})} \\ &= \frac{|\mathbf{X}|^{\frac{-d-1}{2}}\exp\left(\frac{\nu}{2}\ln|\mathbf{X}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}\mathbf{X}\right)\right)}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\frac{\nu}{2}\Gamma_d(\frac{\nu}{2})}} \\ &= \frac{1}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\frac{\nu}{2}\Gamma_d(\frac{\nu}{2})}}|\mathbf{X}|^{\frac{-d-1}{2}}\exp\left\{-\left(\frac{1}{2}\mathrm{vec}(\mathbf{X}), -\frac{1}{2}\ln|\mathbf{X}|\right)\left(\mathrm{vec}(\mathbf{W}^{-1}),\nu\right)^\top\right\} \\ &= \frac{1}{Z(\boldsymbol{\eta})}h(\mathbf{X})\exp\left(-\mathbf{u}(\mathbf{X})^\top\boldsymbol{\eta}\right),\end{aligned}$$

where

$$h(\mathbf{X}) = |\mathbf{X}|^{\frac{-d-1}{2}} \qquad Z(\boldsymbol{\eta}) = 2^{\frac{d\nu}{2}}|\mathbf{W}|^{\frac{\nu}{2}}\Gamma_d(\frac{\nu}{2})$$

$$\boldsymbol{\eta} = \left(\mathrm{vec}(\mathbf{W}^{-1}),\nu\right)^\top, \qquad \mathbf{u}(\mathbf{X}) = \left(\frac{1}{2}\mathrm{vec}(\mathbf{X}), -\frac{1}{2}\ln|\mathbf{X}|\right)^\top.$$

9. [6 points] Does student t distribution (including both the scalar and vector cases) belong to the exponential family? Why?

Answer. The student t distribution is given as $f(x) = c(v)(\nu + x^2)^{-\frac{\nu+1}{2}}$. If we were able to write it as an exponential family, then

$$c(v)\exp\left(-\frac{\nu+1}{2}\ln(\nu+x^2)\right) = c(v)\exp\left(-\eta(\nu)\mathbf{u}(x)\right)$$

which implies $\eta(\nu)\mathbf{u}(x) = \frac{\nu+1}{2}\ln(\nu+x^2)$. Setting $\nu = 1$, we have $u(x) = c\ln(1+x^2)$ where $c = 1/\eta(1)$. Plugging $u(x)$ in the above formula, we get

$$\eta(\nu) = \frac{\frac{\nu+1}{2}\ln(\nu+x^2)}{c\ln(1+x^2)}$$

which is a function of x, a contradiction.

10. [6 points] Does the mixture of Gaussian distribution belong to the exponential family? Why?

$$p(\mathbf{x}) = \frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(\mathbf{x}|\mathbf{m},\boldsymbol{\Lambda})$$

Answer. We know that it is impossible to write $\alpha e^x + \beta e^y$ as $c^{f(x,y)}$ if $\alpha \neq \beta$. Therefore, the mixture of Gaussian distribution with different parameters cannot belong to the exponential family.

11. [**Bonus**][20 points] Given a distribution in the exponential family $p(\mathbf{x}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the natural parameters. As we discussed in the class, the distributions in the exponential family are often parameterized by their expectations, namely $\boldsymbol{\theta} = \mathbb{E}\big(\mathbf{u}(\mathbf{x})\big)$ where $\mathbf{u}(\mathbf{x})$ are the sufficient statistics (recall Gaussian and Bernoulli distributions). Given an arbitrary distribution $p(\mathbf{x}|\boldsymbol{\alpha})$, the Fisher information matrix in terms of the distribution parameters $\boldsymbol{\alpha}$ is defined as $\mathbf{F}(\boldsymbol{\alpha}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\alpha})}[-\frac{\partial^2\log(p(\mathbf{x}|\boldsymbol{\alpha}))}{\partial\boldsymbol{\alpha}^2}]$.

(a) [5 points] Show that if we calculate the Fisher Information matrix in terms of the natural parameters, we have $\mathbf{F}(\boldsymbol{\eta}) = \text{cov}\big(\mathbf{u}(\mathbf{x})\big)$.

Answer: To compute $\mathbf{F}(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}[-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\alpha}^2}]$, we first focus on computing $\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\alpha}^2}$.

$$\begin{aligned}\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} &= \frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\\ &= \frac{\partial}{\partial \boldsymbol{\eta}} \log\left[\frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\big(-\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big)\right]\\ &= \frac{\partial}{\partial \boldsymbol{\eta}}\left[-\log Z(\boldsymbol{\eta}) + \log h(\mathbf{x}) - \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right]\\ &= -\frac{\partial \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} - \mathbf{u}(\mathbf{x})\end{aligned}$$

Consequently, using the result proved in Question

$$\begin{aligned}\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}^2} &= -\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} - \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \boldsymbol{\eta}}\\ &= -\text{cov}(\mathbf{u}(x)) + 0 \qquad \text{using Question 5}\\ &= -\text{cov}_{p(\mathbf{x}|\boldsymbol{\eta})}(\mathbf{u}(x))\end{aligned}$$

Finally, since $\text{cov}_{p(\mathbf{x}|\boldsymbol{\alpha})}(\mathbf{u}(x))$ is a constant with respect to $\mathbf{x}$, we conclude that

$$\begin{aligned}\mathbf{F}(\boldsymbol{\eta}) &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\alpha})}\left[-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}^2}\right]\\ &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\alpha})}\left[\text{cov}(\mathbf{u}(x))\right] = \text{cov}(\mathbf{u}(x)).\end{aligned}$$

(b) [5 points] Show that $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \mathbf{F}(\boldsymbol{\eta})$. Answer: In Question 5, we proved that

$$\boldsymbol{\theta} = \mathbb{E}(\mathbf{u}(\mathbf{x})) = \frac{\partial \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

Therefore,

$$\begin{aligned}\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} &= \frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}\\ &= \text{cov}(\mathbf{u}(\mathbf{x})) \qquad \text{proved in Question 5}\\ &= \mathbf{F}(\boldsymbol{\eta}) \qquad \text{proved in previous section.}\end{aligned}$$

(c) [10 points] Show that the Fisher information matrix in terms of the expectation parameters is the inverse of that in terms of the natural parameters, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$.

Answer: According to the definition,

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}[-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}].$$

So, we need to compute $\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}$.

$$\begin{aligned}\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} &= \frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\\ &= \frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\eta}) \qquad \text{from previous part, } \tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \mathbf{F}(\boldsymbol{\eta})\end{aligned}$$

Consequently,

$$\begin{aligned}
\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}^2} &= \frac{\partial}{\partial \boldsymbol{\eta}} \left[\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right]^\top = \frac{\partial}{\partial \boldsymbol{\eta}} \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}^\top \left[\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right]^\top\right] \\
&= \frac{\partial^2 \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^2} \left[\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right]^\top + \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}^\top \left[\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right] \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \\
&= \frac{\partial^2 \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^2} \left[\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right]^\top + \mathbf{F}(\boldsymbol{\eta})^\top \left[\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right] \mathbf{F}(\boldsymbol{\eta}) \quad \text{from previous part, } \tfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} = \mathbf{F}(\boldsymbol{\eta})
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathbf{F}(\boldsymbol{\eta}) &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}^2}\right) \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial^2 \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^2}\left[\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right]^\top\right) - \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\mathbf{F}(\boldsymbol{\eta})^\top \left[\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right] \mathbf{F}(\boldsymbol{\eta})\right) \\
&= -\frac{\partial^2 \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^2} \underbrace{\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right)^\top}_{=\,\mathbf{0}\text{, see the following for the proof.}} + \mathbf{F}(\boldsymbol{\eta})^\top \underbrace{\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(-\frac{\partial^2 \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right)}_{=\mathbf{F}(\boldsymbol{\theta})} \mathbf{F}(\boldsymbol{\eta}) \\
&= \mathbf{F}(\boldsymbol{\eta})\mathbf{F}(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\eta}) \qquad \mathbf{F}(\boldsymbol{\eta}) = \operatorname{cov}(\mathbf{u}(x)) \text{ is symmetric.}
\end{aligned}$$

Therefore, $\mathbf{F}(\boldsymbol{\eta}) = \mathbf{F}(\boldsymbol{\eta})\mathbf{F}(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\eta})$ which yields $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$ as desired.

Proof of $\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}$.

The proof follows using the fact that $\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right) = 0$ (proved in the class),

$$\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left(\frac{\partial \log(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right) = \mathbf{0}.$$

(d) [5 points] Suppose we observed dataset $\mathcal{D}$. Show that

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mathbf{F}(\boldsymbol{\eta})^{-1} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\theta})^{-1} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

Answer: Note that

$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\eta}) \qquad \text{See Part (b)}
\end{aligned}$$

Consequently,

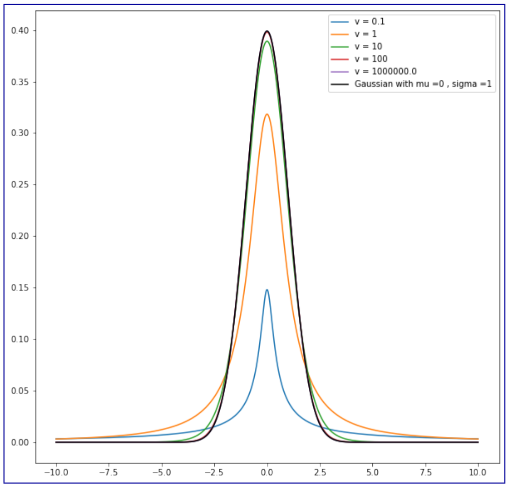
$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mathbf{F}(\boldsymbol{\eta})^{-1} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Also, since in previous part, we proved $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$, we have

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\theta})^{-1}.$$

1 Practice [20 points]

1. [5 Points] Look into the student t's distribution. Let us set the mean and precision to be $\mu = 0$ and $\lambda = 1$. Vary the degree of freedom $\nu = 0.1, 1, 10, 100, 10^6$ and draw the density of the student t's distribution. Also, draw the density of the standard Gaussian distribution $\mathcal{N}(0, 1)$. Please place all the density curves in one figure. Show the legend. What can you observe?
   Answer. As ν goes to infinity, student t's distribution gets closer to the standard Gaussian distribution. Therefore, the standard Gaussian distribution can be seen as the limit of student t's distribution as ν tends to infinity.



2. [5 points] Draw the density plots for Beta distributions: Beta(1,1), Beta(5, 5) and Beta (10, 10). Put the three density curves in one figure. What do you observe? Next draw the density plots for Beta(1, 2), Beta(5,6) and Beta(10, 11). Put the three density curves in another figure. What do you observe?
   Answer. From Figure 1 (left picture), we can see that Beta distribution Beta(a, a) for $a = 1$ is just the uniform distribution over $[0, 1]$ and then it gets more and more bell shaped concentrated around $x = \frac{1}{2}$ as a increases.

   Similarly, from Figure 1 (right picture), we observe that that Beta distribution Beta$(a, a+1)$ for $a = 1$ is a distribution whose density is linearly decreasing on $[0, 1]$ while it starts to get bell shaped and concentrated around $x = \frac{1}{2}$ as a increases.

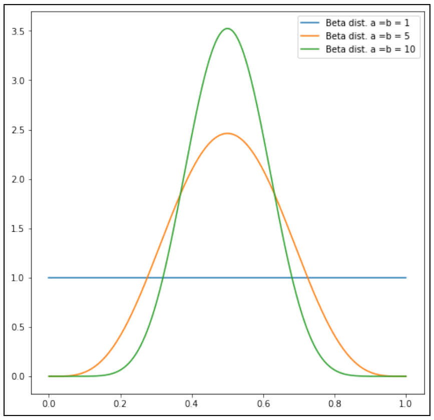


Figure 1: Beta(a, a) for $a \in \{1, 5, 10\}$

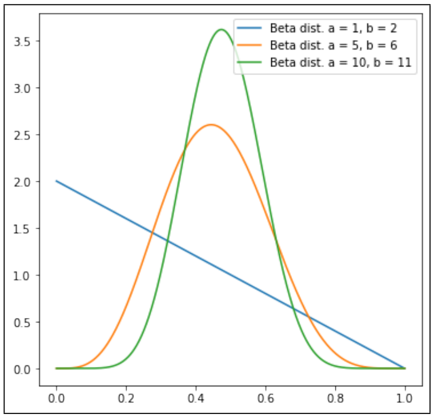


Figure 2: Beta$(a, a+1)$ for $a \in \{1, 5, 10\}$

3. [10 points] Randomly draw 30 samples from a Gaussian distribution $\mathcal{N}(0, 2)$. Use the 30 samples as your observations to find the maximum likelihood estimation (MLE) for a Gaussian distribution and a student t distribution. For both distributions, please use L-BFGS to optimize the parameters. For

student t, you need to estimate the degree of the freedom as well. Draw a plot of the estimated the Gaussian distribution density, student t density and the scatter data points. What do you observe, and why? Next, we inject three noises into the data: we append $\{8, 9, 10\}$ to the 30 samples. Find the MLE for the Gaussian and student t distribution again. Draw the density curves and scatter data points in another figure. What do you observe, and why?

Answer. Comparing Figures 3 and 4, we can see that when there is no noisy points added, the MLE estimation for gaussian and student t's models are resulting pretty the same. But, as the noisy points were added, the MLE estimation for gaussian model was more affected by noise rather than the MLE estimation for student t's model.

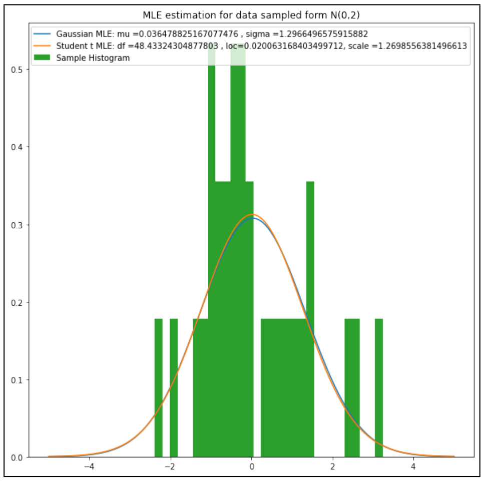


Figure 3: MLE estimation for 30 samples fron $\mathcal{N}(0, 2)$

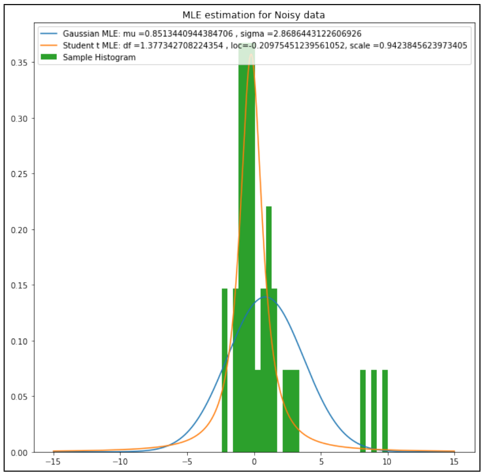


Figure 4: MLE estimation for noisy samples from $\mathcal{N}(0, 2)$