CS 6190: Probabilistic Modelling Spring 2019

Homework 1
Abhinav Kumar (u1209853)

Handed out: 9 Sep, 2019
Due: 11:59pm, 23 Sep, 2019

Analytical problems [80 points + 20 bonus]

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

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

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

We first carry out the transformation of variables $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix}$. Then $\mathbf{y}$ follows the following multivariate Gaussian distribution

$$p(\mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} | \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

The term in the exponential of the Gaussian distribution is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \tag{1}$$

where,

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix}$$

This gives the following equations -

$$\boldsymbol{\Sigma}_{11}\mathbf{V}_{11} + \boldsymbol{\Sigma}_{12}\mathbf{V}_{21} = \mathbf{I}_1 \tag{2}$$
$$\boldsymbol{\Sigma}_{11}\mathbf{V}_{12} + \boldsymbol{\Sigma}_{12}\mathbf{V}_{22} = \mathbf{0} \tag{3}$$
$$\boldsymbol{\Sigma}_{21}\mathbf{V}_{11} + \boldsymbol{\Sigma}_{22}\mathbf{V}_{21} = \mathbf{0} \tag{4}$$
$$\boldsymbol{\Sigma}_{21}\mathbf{V}_{12} + \boldsymbol{\Sigma}_{22}\mathbf{V}_{22} = \mathbf{I}_2 \tag{5}$$

Using (1), we have, $\mathbf{y}_1^T\mathbf{V}_{11}\mathbf{y}_1 + \mathbf{y}_1^T\mathbf{V}_{12}\mathbf{y}_2 + \mathbf{y}_2^T\mathbf{V}_{21}\mathbf{y}_1 + \mathbf{y}_2^T\mathbf{V}_{22}\mathbf{y}_2$.
Now, we use the trick of completing the squares

$$(\mathbf{y}_2 - \mathbf{m})^T\mathbf{M}(\mathbf{y}_2 - \mathbf{m}) + c \tag{6}$$

We need to determine $\mathbf{m}, \mathbf{M}$ and c by comparing these two equations.
The easiest one is $\mathbf{M}$ since it only appears in term involving $\mathbf{y}_2^T$ and $\mathbf{y}_2$ and clearly,

$$\mathbf{M} = \mathbf{V}_{22} \tag{7}$$

Next, we have

$$-\mathbf{y}_2^T\mathbf{M}\mathbf{m} = \mathbf{y}_2^T\mathbf{V}_{21}\mathbf{y}_1 \tag{8}$$

$$\begin{aligned}
&\implies -\mathbf{M}\mathbf{m} = \mathbf{V}_{21}\mathbf{y}_1 \\
&\qquad \implies \mathbf{m} = -\mathbf{M}^{-1}\mathbf{V}_{21}\mathbf{y}_1 \\
&\qquad \implies \mathbf{m} = -\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{y}_1 \quad \text{Using (7)}
\end{aligned} \tag{9}$$

Computing the constant is a bit more involved since

$$\begin{aligned}
&\mathbf{m}^T\mathbf{M}\mathbf{m} + c = \mathbf{y}_1^T\mathbf{V}_{11}\mathbf{y}_1 \\
&\qquad \implies c = \mathbf{y}_1^T\mathbf{V}_{11}\mathbf{y}_1 - \mathbf{m}^T\mathbf{M}\mathbf{m} \\
&\qquad \implies c = \mathbf{y}_1^T\mathbf{V}_{11}\mathbf{y}_1 - \mathbf{y}_1^T\mathbf{V}_{21}^T\mathbf{V}_{22}^{-T}\mathbf{V}_{22}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{y}_1 \quad \text{Using (7) and (9)} \\
&\qquad \implies c = \mathbf{y}_1^T(\mathbf{V}_{11} - \mathbf{V}_{21}^T\mathbf{V}_{22}^{-1}\mathbf{V}_{21})\mathbf{y}_1
\end{aligned} \tag{10}$$

Hence, if we want to marginalize $\mathbf{y}_1$ by integrating out $\mathbf{y}_2$, the term involving $\mathbf{y}_2$ of (6) disappears. The only term left is c which clearly shows that

$$\mathbf{y}_1 \sim \mathcal{N}\left(\mathbf{0}, (\mathbf{V}_{11} - \mathbf{V}_{21}^T\mathbf{V}_{22}^{-1}\mathbf{V}_{21})^{-1}\right) \tag{11}$$

Clearly, from (3), we have $\boldsymbol{\Sigma}_{12} = -\boldsymbol{\Sigma}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}$ and susbstituting this in (2), we have

$$\begin{aligned}
&\boldsymbol{\Sigma}_{11}\mathbf{V}_{11} - \boldsymbol{\Sigma}_{11}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} = \mathbf{I}_1 \\
&\qquad \implies \boldsymbol{\Sigma}_{11} = \left(\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right)^{-1} \\
&\qquad \implies \boldsymbol{\Sigma}_{11} = \left(\mathbf{V}_{11} - \mathbf{V}_{21}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right)^{-1} \quad (\mathbf{V} \text{ is symmetric as well})
\end{aligned} \tag{12}$$

Substituting (12) in (11), we have

$$\mathbf{y}_1 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{11}\right) \tag{13}$$

Since $\mathbf{y}_1 = \mathbf{x}_1 - \boldsymbol{\mu}_1$ or $\mathbf{x}_1 = \mathbf{y}_1 + \boldsymbol{\mu}_1$, we have

$$\mathbf{x}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}\right) \tag{14}$$

Reference: Imperial College Handouts - `http://wwwf.imperial.ac.uk/~das01/MyWeb/M3S3/Handouts/MVN.pdf`

2. [8 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.

   The Moment Generating Function (MGF) of a Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\phi_{\mathbf{x}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}^T\mathbf{x}}) = e^{\mathbf{t}^t\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^t\boldsymbol{\Sigma}\mathbf{t}} \tag{15}$$

The MGF of the new variable $\mathbf{y}$ is given by

$$\begin{aligned}
\phi_{\mathbf{y}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}^T\mathbf{y}}) &= \mathbb{E}\left(e^{\mathbf{t}^T(\mathbf{A}\mathbf{x}+\mathbf{b}+\mathbf{z})}\right) \\
&= \mathbb{E}\left(e^{(\mathbf{A}^T\mathbf{t})^T\mathbf{x}+\mathbf{t}^T\mathbf{b}+\mathbf{t}^T\mathbf{z}}\right) \\
&= \mathbb{E}\left(e^{(\mathbf{A}^T\mathbf{t})^T\mathbf{x}}e^{\mathbf{t}^T\mathbf{b}}e^{\mathbf{t}^T\mathbf{z}}\right) \\
&= \mathbb{E}\left(e^{(\mathbf{A}^T\mathbf{t})^T\mathbf{x}}\right)\mathbb{E}\left(e^{\mathbf{t}^T\mathbf{b}}\right)\mathbb{E}\left(e^{\mathbf{t}^T\mathbf{z}}\right) \quad \mathbf{x}, \mathbf{z} \text{ are independent and using linearity of expectation} \\
&= e^{\mathbf{t}^T\mathbf{A}\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^T\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\mathbf{t}} \quad e^{\mathbf{t}^T\mathbf{b}} \quad e^{\mathbf{t}^T\mathbf{0}+\frac{1}{2}\mathbf{t}^T\boldsymbol{\Lambda}\mathbf{t}} \quad \text{Using (15)} \\
&= e^{\mathbf{t}^T(\mathbf{A}\boldsymbol{\mu}+\mathbf{b})+\frac{1}{2}\mathbf{t}^t(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T+\boldsymbol{\Lambda})\mathbf{t}}
\end{aligned} \tag{16}$$

Since MGF uniquely identifies the random variable and since (16) is in the form of MGF of Gaussian Random vector, we conclude that $\mathbf{y}$ is a Gaussian random vector. Clearly, $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T + \boldsymbol{\Lambda})$

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 $\mathbf{x}$.

$$\begin{aligned}
\mathrm{H}[\mathbf{x}] &= -\int_{-\infty}^{+\infty} N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\ln(N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}))dx \\
&= -\mathbb{E}[\ln(N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}))] \\
&= -\mathbb{E}\left[\ln\left((2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right)\right] \text{ definition of multivariate Gaussian} \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}\left[tr\left[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}\left[tr\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\mathbb{E}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\mathbf{I}\right] \\
&= \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{d}{2} \\
&= \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{d}{2}(1+\ln 2\pi) \qquad (17)
\end{aligned}$$

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., $\mathrm{KL}(q||p)$.

$$\begin{aligned}
\mathrm{KL}(q||p) &= \int q(\mathbf{x})\ln\frac{q(\mathbf{x})}{p(\mathbf{x})}d\mathbf{x} \\
&= \int q(\mathbf{x})\ln q(\mathbf{x})d\mathbf{x} - \int q(\mathbf{x})\log p(\mathbf{x})d\mathbf{x} \\
&= -\mathrm{H}_{\mathbf{q}}[\mathbf{x}] + -\mathbb{E}_{\mathbf{q}}[\ln(N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}))] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2}(1+\ln 2\pi) - \mathbb{E}_{\mathbf{q}}\left[\ln\left((2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right)\right] \quad \text{Using question 3}
\end{aligned}$$

This could be further simplified to

$$\begin{aligned}
\mathrm{KL}(q||p) &= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2}(1+\ln 2\pi) + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbb{E}_{\mathbf{q}}\left[tr\left[(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2}(1+\ln 2\pi) + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\mathbf{q}}\left[(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\mathbf{q}}\left[(\mathbf{x}-\mathbf{m}+\mathbf{m}-\boldsymbol{\mu})(\mathbf{x}-\mathbf{m}+\mathbf{m}-\boldsymbol{\mu})^T\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\mathbf{q}}\left[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T + (\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^T + 2(\mathbf{x}-\mathbf{m})(\mathbf{m}-\boldsymbol{\mu})^T\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\left[\Lambda + (\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^T + 2(\mathbf{m}-\mathbf{m})(\mathbf{m}-\boldsymbol{\mu})^T\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\left[\Lambda + (\mathbf{m}-\boldsymbol{\mu})(\mathbf{m}-\boldsymbol{\mu})^T\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\left[\Lambda + (\boldsymbol{\mu}-\mathbf{m})(\boldsymbol{\mu}-\mathbf{m})^T\right]\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\Lambda + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m})(\boldsymbol{\mu}-\mathbf{m})^T\right] \\
&= -\frac{1}{2}\ln|\Lambda| - \frac{d}{2} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\Lambda\right] + \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}) \\
&= \frac{1}{2}\left[tr\left(\boldsymbol{\Sigma}^{-1}\Lambda\right) + (\boldsymbol{\mu}-\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m}) - d + \ln\frac{|\boldsymbol{\Sigma}|}{|\Lambda|}\right]
\end{aligned}
$$

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

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

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.

$$Z(\boldsymbol{\eta}) = \int_{\mathbf{x}} h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)d\mathbf{x} \tag{18}$$

Now,

$$\frac{\partial Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \int_x h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)(\mathbf{u}(\mathbf{x}))d\mathbf{x} \tag{19}$$

Doing the second derivative,

$$
\begin{aligned}
\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2} &= \frac{\partial}{\partial\boldsymbol{\eta}}\cdot\frac{\partial^T}{\partial\boldsymbol{\eta}}\log Z(\boldsymbol{\eta}) \\
&= \frac{\partial}{\partial\boldsymbol{\eta}}\cdot\frac{1}{Z(\boldsymbol{\eta})}\int_{\mathbf{x}} h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)(\mathbf{u}(\mathbf{x})^T)d\mathbf{x} \\
&= \frac{-1}{Z(\boldsymbol{\eta})^2}\int_{\mathbf{x}} h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)(\mathbf{u}(\mathbf{x}))d\mathbf{x}\int_x h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)(\mathbf{u}(\mathbf{x})^T)d\mathbf{x} \\
&\quad + \frac{1}{Z(\boldsymbol{\eta})}\int_{\mathbf{x}} h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)(\mathbf{u}(\mathbf{x}))(\mathbf{u}(\mathbf{x})^T)d\mathbf{x}
\end{aligned} \tag{20}
$$

Observe that the function $p(\mathbf{x}|\boldsymbol{\eta}) = \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})}\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)$ is a proper probability distribution over $\mathbf{x}$ as $h(\mathbf{x})$ is non-negative by definition and other terms are non-negative as well. Moreover,

$$\int_x \frac{1}{Z(\boldsymbol{\eta})}h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)d\mathbf{x} = \frac{1}{Z(\boldsymbol{\eta})}Z(\boldsymbol{\eta}) = 1$$

.

Hence, substituting this in (20), we get

$$
\begin{aligned}
\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2} &= -\int_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\eta})u(\mathbf{x})d\mathbf{x}\int_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\eta})u(\mathbf{x})^T d\mathbf{x} + \int_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\eta})u(\mathbf{x})u(\mathbf{x})^T d\mathbf{x} \\
&= -\mathbb{E}(u(\mathbf{x}))\mathbb{E}(u(\mathbf{x})^T) + \mathbb{E}(u(\mathbf{x})u(\mathbf{x})^T) \\
&= Cov(u(\mathbf{x}))
\end{aligned} \tag{21}
$$

6. [2 points] Is $\log Z(\boldsymbol{\eta})$ convex or nonconvex? Why?

Since covariance matrices are positive semi-definite, hence $\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2}$ is also positive semi-definite. Hence, $\log Z(\boldsymbol{\eta})$ is convex

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.

$$\begin{aligned}
I(\mathbf{x},\mathbf{y}) &= \sum_x \sum_y p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_x \sum_y p(x,y) \ln \frac{p(x|y)p(y)}{p(x)p(y)} \\
&= \sum_x \sum_y p(x,y) \left[\ln p(x|y) - \ln p(x)\right] \\
&= \sum_x \sum_y p(x,y) \ln p(x|y) - \sum_x \ln p(x) \sum_y p(x,y) \\
&= \sum_x \sum_y p(x,y) \ln p(x|y) - \sum_x p(x) \ln p(x) \\
&= \sum_y p(y) \sum_x p(x|y) \ln p(x|y) + H[\mathbf{x}] \\
&= -\sum_y p(y) H(\mathbf{x}|\mathbf{y}=y) + H[\mathbf{x}] \\
&= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \qquad (22)
\end{aligned}$$

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

The exponential family consists of distributions which can be written in the following form

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})} \exp\left(-\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right) \qquad (23)$$

. The interaction between the parameters and the variables should only occur in the exponential part.

- Dirichlet Distribution: It is a generalization of Binomial Distribution and can be written as

$$p(x_1,...,x_K|\alpha_1,...,\alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \qquad (24)$$

and where $\sum x_i = 1$ and $x_i \geq 0$

$$\begin{aligned}
p(x_1,...,x_K|\alpha_1,...,\alpha_K) &= \frac{1}{B(\alpha)} \frac{\prod \exp(\ln(x_i)\alpha_i)}{\prod x_i} \\
&= \frac{1}{B(\alpha)} \frac{\exp(\sum \ln(x_i)\alpha_i)}{\prod x_i} \\
&= \frac{\frac{1}{\prod x_i}}{B(\alpha)} \exp(\sum \ln(x_i)\alpha_i) \qquad (25)
\end{aligned}$$

Comparing (25) with (23), we get

$$\boldsymbol{\eta} = \begin{bmatrix} -\alpha_1 \\ \dots \\ -\alpha_K \end{bmatrix}$$

$$\begin{aligned}
\mathbf{u}(\mathbf{x}) &= \begin{bmatrix} \ln x_1 \\ \dots \\ \ln x_K \end{bmatrix} \\
h(\mathbf{x}) &= \frac{1}{\prod x_i} \\
Z(\boldsymbol{\eta}) &= B(\alpha)
\end{aligned} \tag{26}$$

- Gamma Distribution:

$$\begin{aligned}
p(x|\alpha,\beta) &= \frac{\beta^\alpha}{\tau(\alpha)} x^{\alpha-1}\exp(-\beta x) \\
&= \frac{\frac{1}{x}}{\beta^{-\alpha}\tau(\alpha)} x^{\alpha}\exp(-\beta x) \\
&= \frac{\frac{1}{x}}{\beta^{-\alpha}\tau(\alpha)} \exp(\alpha\ln x)\exp(-\beta x) \\
&= \frac{\frac{1}{x}}{\beta^{-\alpha}\tau(\alpha)} \exp(\alpha\ln x - \beta x)
\end{aligned} \tag{27}$$

Comparing (25) with (23), we get

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \\
\mathbf{u}(\mathbf{x}) &= \begin{bmatrix} -\ln x \\ x \end{bmatrix} \\
h(\mathbf{x}) &= \frac{1}{x} \\
Z(\boldsymbol{\eta}) &= \beta^{-\alpha}\tau(\alpha)
\end{aligned} \tag{28}$$

- Wishart Distribution:

$$\begin{aligned}
f(\mathbf{x}|p,\mathbf{V}) &= \frac{1}{2^{np/2}|\mathbf{V}|^{n/2}\tau_p(\frac{n}{2})}|\mathbf{x}|^{\frac{n-p-1}{2}}e^{\frac{-1}{2}tr(\mathbf{V}^{-1}\mathbf{x})} \\
&= \frac{|\mathbf{x}|^{\frac{n-1}{2}}}{2^{np/2}|\mathbf{V}|^{n/2}\tau_p(\frac{n}{2})}|\mathbf{x}|^{\frac{-p}{2}}e^{\frac{-1}{2}tr(\mathbf{V}^{-1}\mathbf{x})} \\
&= \frac{|\mathbf{x}|^{\frac{n-1}{2}}}{2^{np/2}|\mathbf{V}|^{n/2}\tau_p(\frac{n}{2})}\exp\left(-\frac{p}{2}\ln|\mathbf{x}| - tr(\mathbf{V}^{-1}\frac{\mathbf{x}}{2})\right)
\end{aligned} \tag{29}$$

Comparing (29) with (23), we get

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{bmatrix} \frac{p}{2} \\ \mathbf{V}^{-1} \end{bmatrix} \\
\mathbf{u}(\mathbf{x}) &= \begin{bmatrix} \ln|\mathbf{x}| \\ \frac{\mathbf{x}}{2} \end{bmatrix} \\
h(\mathbf{x}) &= |\mathbf{x}|^{\frac{n-1}{2}} \\
Z(\boldsymbol{\eta}) &= 2^{np/2}|\mathbf{V}|^{n/2}\tau_p(\frac{n}{2})
\end{aligned} \tag{30}$$

9. [4 points] Does student t distribution (including both the scalar and vector cases) belong to the exponential family? Why?

- The pdf of the scalar student t distribution is given by

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \nu > 0 \tag{31}$$

Let us assume we have some $u(x)$ such that $\exp(-u(x)\eta) = (\nu+x^2)^{-\frac{\nu+1}{2}}$. Choose $\nu = 1$, we have $\exp(-u(x)\eta|_{\nu=1}) = (1+x^2)^{-1}$ and so, we have $u(x) = c\ln(1+x^2)$ for some constant c. Now, substituting back, we have

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

$$\implies \eta = \frac{\nu+1}{2c}\frac{\ln(\nu+x^2)}{\ln(1+x^2)} \tag{32}$$

The above statement should have been independent of x but it contains x. Hence, our assumption that the scalar student t distribution of the exponential family is wrong. Hence, scalar student t distribution does not belong to the exponential family.

- The pdf of vector t distribution is given by

$$p(\mathbf{x}) = \frac{\Gamma\left[(\nu+p)/2\right]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}^{1/2}|}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2} \tag{33}$$

We will prove this by induction over the number of dimensions n. The base case for $n = 1$ (scalar case) fails as shown above. Hence, our hypothesis that student t distribution belongs to the exponential case is wrong.

10. [2 points] Does the mixture of Gaussian distribution belong to the exponential family? Why?

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

The mixture of Gaussians can not be written in the form required for exponential family and hence it does not 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 $\mathbf{F}(\boldsymbol{\eta}) = \mathrm{cov}\big(\mathbf{u}(\mathbf{x})\big)$.

We have, $\mathbf{F}(\boldsymbol{\eta}) = -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left[\frac{\partial^2\ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial\boldsymbol{\eta}^2}\right]$ The exponential family is given by

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\eta}) &= \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})}\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)\\
\implies \ln p(\mathbf{x}|\boldsymbol{\eta}) &= -\ln Z(\boldsymbol{\eta}) + \mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta} + \ln h(\mathbf{x})\\
\implies \frac{\partial\ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial\boldsymbol{\eta}} &= -\frac{\partial\ln Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} + \mathbf{u}(\mathbf{x})\\
\implies \frac{\partial^2\ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial\boldsymbol{\eta}^2} &= -\frac{\partial^2\ln Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2}\\
\implies -\mathbb{E}\left[\frac{\partial^2\ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial\boldsymbol{\eta}^2}\right] &= \mathbb{E}\left[\frac{\partial^2\ln Z(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2}\right] = \mathrm{cov}(u(\mathbf{x})) \quad \text{Using the result of question 5}\\
\implies \mathbf{F}(\boldsymbol{\eta}) &= \mathrm{cov}\big(\mathbf{u}(\mathbf{x})\big)
\end{aligned} \tag{34}$$

(b) [5 points] Show that $\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\eta}} = \mathbf{F}(\boldsymbol{\eta})$.

$$\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\eta}} = \frac{\partial\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\mathbf{u}(\mathbf{x})}{\partial\boldsymbol{\eta}} \tag{35}$$

Now, $p(\mathbf{x}|\boldsymbol{\eta}) = \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})}\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)$ since the terms $h(\mathbf{x})$ and $Z(\boldsymbol{\eta})$ are all non-negative and they all sum up to 1. Hence. we have

$$\begin{aligned}
\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\eta}} &= \frac{\partial}{\partial\boldsymbol{\eta}}\int\frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})}\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)\mathbf{u}(\mathbf{x})d\mathbf{x}\\
&= \frac{\partial\frac{1}{Z(\boldsymbol{\eta})}}{\partial\boldsymbol{\eta}}\int h(\mathbf{x})\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)\mathbf{u}(\mathbf{x})d\mathbf{x} + \int\frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})}\exp\left(\mathbf{u}(\mathbf{x})^\top\boldsymbol{\eta}\right)\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^T d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
&= \frac{-1}{Z(\boldsymbol{\eta})^2} \int_x h(\mathbf{x}) \exp\big(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) dx \int h(\mathbf{x}) \exp\big(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) d\mathbf{x} \\
&\quad + \int \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})} \exp\big(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\big) \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^T d\mathbf{x} \quad \text{Using equation (19)} \\
&= -\mathbb{E}(u(\mathbf{x}))\mathbb{E}(u(\mathbf{x})^T) + \mathbb{E}(u(\mathbf{x})u(\mathbf{x})^T) \\
&= \text{cov}\big(\mathbf{u}(\mathbf{x})\big) \\
&= \mathbf{F}(\boldsymbol{\eta})
\end{aligned} \tag{36}$$

(c) [5 points] Show that the Fisher information matrix in terms of the expectation parameters is the inverse of that in terms of the natural parameters, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$.

$$\begin{aligned}
\mathbf{F}(\boldsymbol{\theta}(\boldsymbol{\eta})) &= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial}{\partial \theta} \cdot \left[\frac{\partial^T \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right]\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \cdot \left[\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}^T \frac{\partial^T \ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right]\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial^2 \boldsymbol{\eta}}{\partial \boldsymbol{\theta}^2} \frac{\partial^T \ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} + \frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\theta}} \frac{\partial}{\partial \boldsymbol{\theta}} \cdot \frac{\partial^T \ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}}\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial^2 \boldsymbol{\eta}}{\partial \boldsymbol{\theta}^2} \frac{\partial^T \ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} + \frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\theta}} \frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\eta}^2} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\theta}} \frac{\partial^2 \ln(p(\mathbf{x}|\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}^2} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}}\right] \\
&= \frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\eta}) \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \\
&= \mathbf{F}(\boldsymbol{\eta})^{-T} \mathbf{F}(\boldsymbol{\eta}) \mathbf{F}(\boldsymbol{\eta})^{-1} \quad \text{Using Q 10(b)} \\
&= \mathbf{F}(\boldsymbol{\eta})^{-T} \\
&= \mathbf{F}(\boldsymbol{\eta})^{-1} \quad \text{Since } \mathbf{F}(\boldsymbol{\eta}) = cov(\mathbf{u}(\mathbf{x})) \text{is a covariance matrix and therefore a symmetric matrix.}
\end{aligned} \tag{37}$$

(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

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

Note that I choose the orientation of the gradient vector to be consistent with Jacobian. So, in this case, the gradient vector is a row vector (rather than a column vector). If you want to use a column vector to represent the gradient, you can move the information matrix to the left. It does not influence the conclusion.

-
$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mathbf{F}(\boldsymbol{\eta})^{-1}
\end{aligned} \tag{38}$$

- Now, using (38),we have

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\eta}) \tag{39}$$

But from question 11 (c), we have $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$ or in other words, $\mathbf{F}(\boldsymbol{\eta}) = \mathbf{F}(\boldsymbol{\theta})^{-1}$. Substituting this in the above equation, we have

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\theta})^{-1} \tag{40}$$

Programming 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?

   Figure 1 shows the plot. As $\nu \to \infty$, student's t approaches Gaussian distribution.

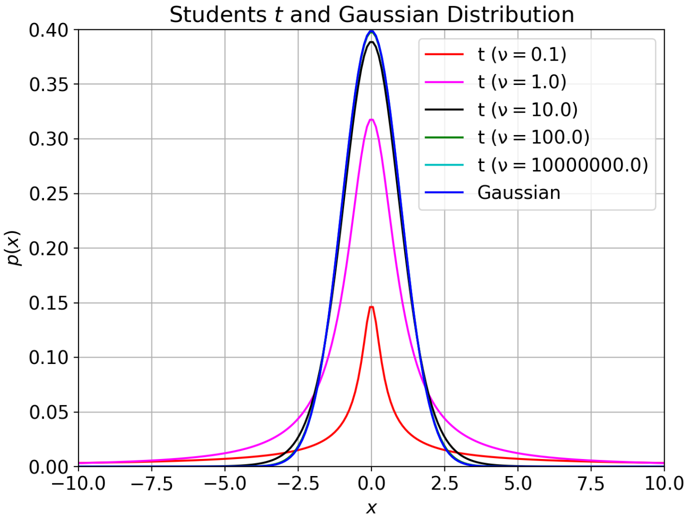


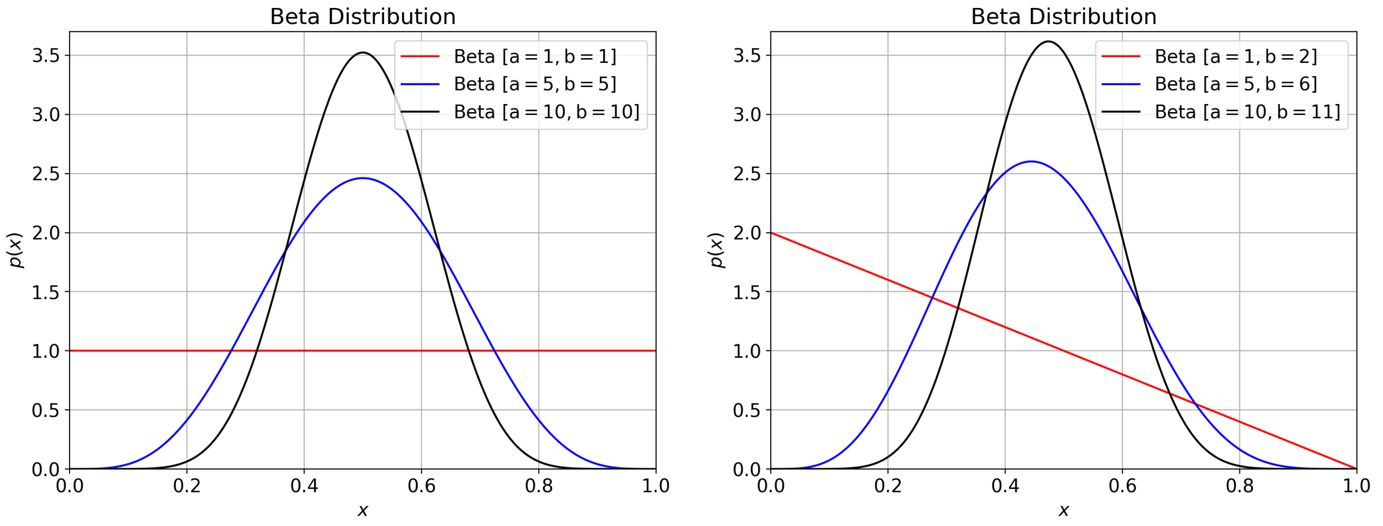
Figure 1: Student t and Gaussian Plot

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?

   Figure 2a shows the first plot. As a increases, Beta becomes more and more bell shaped. Figure 2b shows the second plot. As a and b both increases, Beta modifies itself from Triangular to Bell shaped.

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?

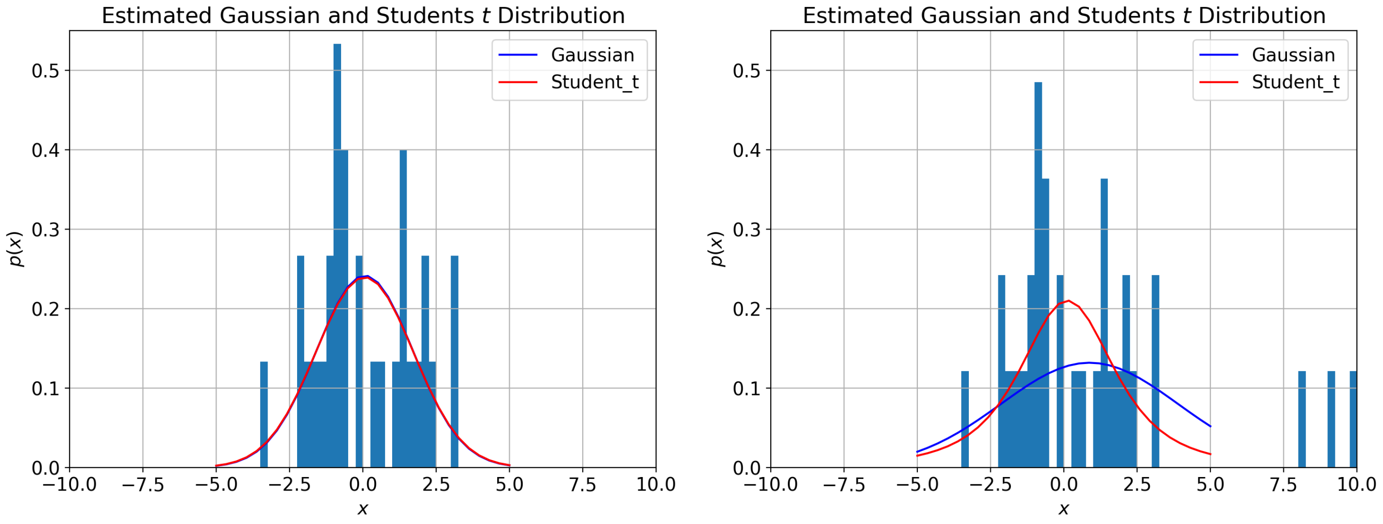
   Figure 3a shows the first plot. We see that the estimated Gaussian and Student's t distribution are in close agreement to each other. Figure 3b shows the second plot when noisy samples are added. We see that the estimated Gaussian and Student's t distribution are no longer close agreement to each other. Table 1 shows the estimated parameters in both the cases. It seems that the MLE of parameters of Student's t distribution are not affected by outliers while the the MLE of parameters of Gaussian distribution are affected by outliers.



(a) Beta Distribution Plots: Tied a and b

(b) Beta Distribution Plots: $b = a + 1$

Figure 2: Beta Distribution Plots



(a) Estimated Gaussian and Student's t Distribution when only normally drawn samples are there.

(b) Estimated Gaussian and Student's t Distribution when noisy samples are also added.

Figure 3: Estimated Distributions and Scatter data points

Case	Gaussian		Student's t		
	μ	σ	μ	σ	ν
No outlier	0.063	1.66	0.06	1.66	629.47
With outliers	0.88	3.02	0.12	1.72	2.46

Table 1: MLE Parameters. The true value of $\mu = 0$ and $\sigma = 1.41$