CS 6190: Probabilistic Modelling Spring 2019

Homework 1
Aishwarya Gupta(01266423)

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

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.

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 $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}{2}\right) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp(K)$$

Let $\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$ and K be defined as :

$$\begin{aligned}
K &= -\frac{1}{2}(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) \\
&= -\frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - \frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{12}(\mathbf{x}_2-\mu_2) \\
&\quad - \frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \mathbf{V}_{21}(\mathbf{x}_1-\mu_1) - \frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2) \\
&= -\frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - (\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{12}(\mathbf{x}_2-\mu_2) - \frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2) \\
&= -[\frac{1}{2}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1)) \\
&\quad + \frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - \frac{1}{2}(\mathbf{V}_{12}(\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{22}^{-1}(\mathbf{V}_{12}(\mathbf{x}_1-\mu_1)] \qquad (1)
\end{aligned}$$

Using eq (1), we get :

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}\left((\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - (\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}^{-1}(\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))\right)\right)$$
$$exp(-\frac{1}{2}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1)))$$
$$= K_1 exp(-\frac{1}{2}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))) \quad (2)$$

where $K_1 = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}\left((\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - (\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}^{-1}(\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))\right)\right)$

$$p(\mathbf{x}_1) = \int p(\mathbf{x}) dx_2$$
$$= K_1 \int exp(-\frac{1}{2}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu_2+\mathbf{V}_{22}^{-1}\mathbf{V}_{12}(\mathbf{x}_1-\mu_1)))dx_2$$
$$= K_1 \int exp(-\frac{1}{2}(\mathbf{x}_2-\mu')^\top \mathbf{V}_{22}(\mathbf{x}_2-\mu'))dx_2 = K_1(2\pi)^{\frac{1}{2}(\frac{d}{2})}|\mathbf{V}_{22}^{-1}|^{1/2} \quad (3)$$

$$\mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{bmatrix}^{-1} = \begin{bmatrix}(\mathbf{V}_{11}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21})^{-1} & \mathbf{V}'_{12}\\ \mathbf{V}'_{21} & \mathbf{V}'_{22}\end{bmatrix} \quad (4)$$

$$K_1 = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}\left((\mathbf{x}_1-\mu_1)^\top \mathbf{V}_{11}(\mathbf{x}_1-\mu_1) - (\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))^\top \mathbf{V}_{22}^{-1}(\mathbf{V}_{12}(\mathbf{x}_1-\mu_1))\right)\right)$$
$$= \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}_1-\mu_1)^\top (\mathbf{V}_{11}-\mathbf{V}_{12}^\top \mathbf{V}_{22}^{-1}\mathbf{V}_{12})(\mathbf{x}_1-\mu_1)\right)$$
$$= \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_1-\mu_1)\right) \quad \text{from eq 4}$$

$$p(\mathbf{x}_1) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}(2\pi)^{d/4}|\mathbf{V}_{22}^{-1}|^{1/2} exp\left(-\frac{1}{2}(\mathbf{x}_1-\mu_1)^\top \mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_1-\mu_1)\right)$$

[Help is taken from this link]

2. [8 points] Given a Gaussian random vector, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$. 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 $\mathbf{x}$, $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{\Lambda})$. Show $\mathbf{y}$ follows Gaussian distribution as well, and derive its form.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{z} = \mathbf{A}\mathbf{x} + (\mathbf{b} + \mathbf{z}) = \mathbf{u} + \mathbf{v}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{u}|\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$ and $\mathbf{v} \sim \mathcal{N}(\mathbf{v}|\mathbf{b}, \mathbf{\Lambda})$. Since $\mathbf{u}$ and $\mathbf{v}$ are independent, the characteristic function of $\mathbf{y}$ is equal to the product of the characteristic functions of $\mathbf{u}$ and $\mathbf{v}$

$$\phi_{\mathbf{y}}(t) = \phi_{\mathbf{u}}(t)\phi_{\mathbf{v}}(t) = exp\left(it^\top(\mathbf{A}\boldsymbol{\mu}) - \frac{t^\top(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)t}{2}\right) exp\left(it^\top(\mathbf{b}) - \frac{t^\top \mathbf{\Lambda} t}{2}\right)$$
$$= exp\left(it^\top(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}) - \frac{t^\top(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top+\mathbf{\Lambda})t}{2}\right) \quad (1)$$

Since the characteristic function of $\mathbf{y}$ is same as the characteristic function of Gaussian distribution with mean $= \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and covariance $= \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top + \mathbf{\Lambda}$, we can say that $\mathbf{y}$ also follows a Gaussian distribution i.e. $\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top + \mathbf{\Lambda})$

Help is taken from https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

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 $\mathbf{x}$.

$$\begin{aligned}
H[\mathbf{x}] &= -\int p(\mathbf{x}) \log p(\mathbf{x}) dx \\
&= -\int p(\mathbf{x}) \log \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}{2} \right) \right) dx \\
&= \int \left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \right) p(\mathbf{x}) dx + \frac{1}{2}\int (\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\mathbf{p}(\mathbf{x}) dx \\
&= \left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \right) + \int p(\mathbf{x}) dx + \frac{1}{2}\mathbf{E}\left[(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) \right] \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbf{E}\left[tr((\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)) \right] \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\mathbf{E}\left[tr(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top) \right] \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1}\mathbf{E}\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top \right]) \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2} tr(\mathbf{I}) \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2} = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}(1 + \log 2\pi)
\end{aligned}$$

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

$$\begin{aligned}
KL(q||p) &= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} dx \\
&= \int q(\mathbf{x}) \log q(\mathbf{x}) dx - \int q(\mathbf{x}) \log p(\mathbf{x}) dx \\
&= -H_q[\mathbf{x}] - \int q(\mathbf{x}) \log \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right) dx + \frac{1}{2}\int (\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) q(\mathbf{x}) dx \\
&= -\frac{1}{2}\log|\Lambda| - \frac{d}{2}(1 + \log 2\pi) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\int (\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) q(\mathbf{x}) dx \\
&= \log(\sqrt{|\boldsymbol{\Sigma}|/|\Lambda|}) - \frac{d}{2} + \mathbf{Y} \qquad (1)
\end{aligned}$$

$$
\begin{aligned}
\mathbf{Y} &= \frac{1}{2}\int (\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)q(\mathbf{x})dx \\
&= \frac{1}{2}\mathbb{E}_q\left((\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right) \\
&= \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}_q\left((\mathbf{x}-\mu)^\top(\mathbf{x}-\mu)\right)\right) \\
&= \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}_q\left((\mathbf{x}-\mathbf{m}+\mathbf{m}-\mu)^\top(\mathbf{x}-\mathbf{m}+\mathbf{m}-\mu)\right)\right) \\
&= \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}_q\left((\mathbf{x}-\mathbf{m})^\top(\mathbf{x}-\mathbf{m})\right) + 2\mathbb{E}_q\left((\mathbf{m}-\mu)^\top\mathbf{x}\right) - \mathbb{E}_q\left((\mathbf{m}-\mu)^\top(\mathbf{m}+\mu)\right)\right) \\
&= \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\left(\Lambda + 2(\mathbf{m}-\mu)^\top\mathbf{m} - (\mathbf{m}-\mu)^\top(\mathbf{m}+\mu)\right)\right) \\
&= \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\left(\Lambda + (\mathbf{m}-\mu)^\top(\mathbf{m}-\mu)\right)\right) \qquad (1)
\end{aligned}
$$

Substituting $\mathbf{Y}$ from equation (2) in equation(1), we get :

$$KL(q||p) = \log(\sqrt{|\boldsymbol{\Sigma}|/|\Lambda|}) - \frac{d}{2} + \frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\left(\Lambda + (\mathbf{m}-\mu)^\top(\mathbf{m}-\mu)\right)\right)$$

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.

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

$$\mathbf{Z}(\boldsymbol{\eta}) = \int h(\mathbf{x})exp(-u(\mathbf{x})^\top\boldsymbol{\eta})dx \qquad (2)$$

$$
\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\eta}^2}\log\mathbf{Z}(\boldsymbol{\eta}) &= \frac{\partial}{\partial\boldsymbol{\eta}}\left(\frac{\partial}{\partial\boldsymbol{\eta}}\log\mathbf{Z}(\boldsymbol{\eta})\right) \\
&= \frac{\partial}{\partial\boldsymbol{\eta}}\left(\frac{1}{\mathbf{Z}(\boldsymbol{\eta})}\frac{\partial\mathbf{Z}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right) \\
&= \frac{1}{\mathbf{Z}(\boldsymbol{\eta})}\frac{\partial^2\mathbf{Z}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^2} - \frac{1}{\mathbf{Z}(\boldsymbol{\eta})^2}\left(\frac{\partial\mathbf{Z}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right)^2 \qquad (3)
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial \mathbf{Z}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\partial}{\partial \boldsymbol{\eta}} \int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) dx \quad \text{from eq (2)} \\
&= -\int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) dx \qquad (4) \\
\frac{1}{\mathbf{Z}(\boldsymbol{\eta})} \left(\frac{\partial \mathbf{Z}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right) &= -\frac{1}{\mathbf{Z}(\boldsymbol{\eta})} \int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) dx \\
&= -\int \frac{h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) dx}{\mathbf{Z}(\boldsymbol{\eta})} \\
&= -\int u(\mathbf{x}) p(\mathbf{x}|\boldsymbol{\eta}) dx \quad \text{from eq(1)} \\
&= -\mathbb{E}(u(\mathbf{x})) \\
\left(\frac{1}{\mathbf{Z}(\boldsymbol{\eta})} \left(\frac{\partial \mathbf{Z}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \right) \right)^2 &= [\mathbb{E}(u(\mathbf{x}))]^2 \qquad (5)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \mathbf{Z}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} &= \frac{\partial}{\partial \boldsymbol{\eta}} \int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) dx \quad \text{from eq (3)} \\
&= \int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) u(\mathbf{x})^\top dx \\
\frac{1}{\mathbf{Z}(\boldsymbol{\eta})} \frac{\partial^2 \mathbf{Z}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} &= \frac{1}{\mathbf{Z}(\boldsymbol{\eta})} \int h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) u(\mathbf{x})^\top dx \\
&= \int \frac{1}{\mathbf{Z}(\boldsymbol{\eta})} h(\mathbf{x}) exp(-u(\mathbf{x})^\top \boldsymbol{\eta}) u(\mathbf{x}) u(\mathbf{x})^\top dx \\
&= \int u(\mathbf{x}) u(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta}) dx \quad \text{from eq (1)} \\
&= \mathbb{E}[u(\mathbf{x}) u(\mathbf{x})^\top] \qquad (6)
\end{aligned}$$

From eq (3), (5) and (6), we get :

$$\begin{aligned}
\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} &= \mathbb{E}[u(\mathbf{x}) u(\mathbf{x})^\top] - [\mathbb{E}(u(\mathbf{x}))]^2 \\
&= cov(u(\mathbf{x})) \qquad (2)
\end{aligned}$$

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

$$\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = \text{cov}(\mathbf{u}(\mathbf{x})) \geq 0 \qquad (3)$$

Thus, $\log Z(\boldsymbol{\eta})$ 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.

$$
\begin{aligned}
I(\mathbf{x},\mathbf{y}) &= \int\int p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} dxdy \\
&= \int\int p(\mathbf{x},\mathbf{y})(\log(p(\mathbf{x},\mathbf{y})) - \log(p(\mathbf{x})) - \log(p(\mathbf{y})))dxdy \\
&= \int\int p(\mathbf{x},\mathbf{y}) \log(p(\mathbf{x},\mathbf{y}))dxdy - \int \log(p(\mathbf{x})) \left(\int p(\mathbf{x},\mathbf{y})dy \right) dx - \int \log(p(\mathbf{y})) \left(\int p(\mathbf{x},\mathbf{y})dx \right) dy \\
&= -H[\mathbf{x},\mathbf{y}] - \int \log(p(\mathbf{x}))p(\mathbf{x})dx - \int \log(p(\mathbf{y}))p(\mathbf{y})dy \\
&= -H[\mathbf{x},\mathbf{y}] + H[\mathbf{x}] + H[\mathbf{y}] \\
&= -(H[\mathbf{y}] + H[\mathbf{x}|\mathbf{y}]) + H[\mathbf{x}] + H[\mathbf{y}] \\
&= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}]
\end{aligned}
$$

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

$$
\begin{aligned}
p(\mathbf{x}|\alpha_1,\alpha_2,...,\alpha_K) &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \prod_{k=1}^K x_k^{\alpha_k - 1} \\
&= exp\left(\log_e \left(\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \prod_{k=1}^K x_k^{\alpha_k - 1} \right) \right) \\
&= exp\left(\log_e \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} + \sum_{k=0}^K (\alpha_k - 1) \log_e x_k \right) \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} exp[(\log \mathbf{x})^\top (\alpha - 1)]
\end{aligned}
$$

   This gives us $h(\mathbf{x}) = 1$, $u(\mathbf{x}) = \log \mathbf{x}$, $\boldsymbol{\eta} = (\alpha - 1)$ and $Z(\boldsymbol{\eta}) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)}{\Gamma(\sum_{k=1}^K \alpha_k)}$

   - Gamma distribution

$$
\begin{aligned}
Gamma(\mathbf{x}|a,b) &= \frac{b^a \mathbf{x}^{a-1}}{\Gamma(a)} exp(-b\mathbf{x}) \\
&= \frac{b^a}{\Gamma(a)} exp((a-1)\log \mathbf{x} - b\mathbf{x}) \\
&= \frac{b^a}{\Gamma(a)} exp(u(\mathbf{x})^\top \boldsymbol{\eta}) \qquad (4)
\end{aligned}
$$

   This gives : $h(\mathbf{x}) = 1$, $u(\mathbf{x}) = \begin{bmatrix} \log \mathbf{x} \\ \mathbf{x} \end{bmatrix}$, $\boldsymbol{\eta} = \begin{bmatrix} (a-1) \\ -b \end{bmatrix} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}$

$$
Z(\boldsymbol{\eta}) = \frac{\Gamma(a)}{b^a} = \frac{\Gamma(\boldsymbol{\eta}_1 + 1)}{(-\boldsymbol{\eta}_2)^{\boldsymbol{\eta}_1 + 1}} \qquad (5)
$$

- Wishart distribution

$$Wishart(\mathbf{X}|\mathbf{W},v) = \frac{1}{2^{\frac{dv}{2}}|\mathbf{W}|^{v/2}\Gamma_d(v/2)}|\mathbf{X}|^{(v-d-1)/2}exp\left(-\frac{1}{2}tr(\mathbf{W}^{-1}\mathbf{X})\right)$$
$$= \frac{1}{2^{\frac{dv}{2}}|\mathbf{W}|^{v/2}\Gamma_d(v/2)}exp\left(-\frac{1}{2}tr(\mathbf{W}^{-1}\mathbf{X}) + \frac{v-d-1}{2}\log|\mathbf{X}|\right) = \frac{1}{2^{\frac{dv}{2}}|\mathbf{W}|^{v/2}\Gamma_d(v/2)}exp(u(\mathbf{X}$$

This gives : $h(\mathbf{x}) = 1$, $u(\mathbf{x}) = \begin{bmatrix}\mathbf{X}\\ \log|\mathbf{X}|\end{bmatrix}$, $\boldsymbol{\eta} = \begin{bmatrix}-\frac{1}{2}\mathbf{W}^{-1}\\ \frac{v-d-1}{2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix}$ and $z(\boldsymbol{\eta}) = 2^{\frac{dv}{2}}|\mathbf{W}|^{v/2}\Gamma_d(v/2)$

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

$$t(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-v/2-1/2}$$
$$= exp\left(\log\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2} + \log\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-v/2-1/2}\right)$$
$$= exp\left(\log\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\right)exp\left(\log\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-v/2-1/2}\right)$$
$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}exp\left((-v/2-1/2)\log\left(\nu+\lambda(x-\mu)^2\right) + (v/2+1/2)\log\nu\right)$$
$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\nu^{\nu/2+1/2}exp\left((-v/2-1/2)\log\left(\nu+\lambda(x-\mu)^2\right)\right)$$
$$= \phi(\nu,\lambda)exp\left((-v/2-1/2)\log\left(\nu+\lambda(x-\mu)^2\right)\right) \quad (6)$$

The expression inside the exponent cannot be written as $(u(x)^\top\eta)$ and thus the t-distribution does not belong to exponential family. For vector case,

$$t(\mathbf{x}|\mu,\boldsymbol{\Lambda},\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{|\boldsymbol{\Lambda}|}{\pi\nu}\right)^{1/2}\nu^{\nu/2+1/2}exp\left((-v/2-1/2)\log\left(\nu+(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\right) \quad (7)$$

Again, the expression inside the exponent cannot be written as $(u(\mathbf{x})^\top\boldsymbol{\eta})$ and thus the t-distribution for vector case also does not belong to the exponential family.

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

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

$$p(\mathbf{x}) = \frac{1}{2(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}exp((\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) + \frac{1}{2(2\pi)^{d/2}|\boldsymbol{\Lambda}|^{1/2}}exp((\mathbf{x}-\mathbf{m})^\top\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\mathbf{m}))$$

This cannot be reduced in the form of $h(\mathbf{x})exp(u(\mathbf{x})^\top\boldsymbol{\eta} - \log z(\boldsymbol{\eta}))$. Thus, the mixture of Gaussian distributions 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 $\mathbf{F}(\boldsymbol{\eta}) = \text{cov}\big(\mathbf{u}(\mathbf{x})\big)$.

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\eta}) &= \frac{h(\mathbf{x})}{z(\boldsymbol{\eta})} exp(u^\top(\mathbf{x})\boldsymbol{\eta}) \\
\log p(\mathbf{x}|\boldsymbol{\eta}) &= \log h(\mathbf{x}) - \log z(\boldsymbol{\eta}) + u^\top(\mathbf{x})\boldsymbol{\eta} \\
\frac{\partial \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= u(\mathbf{x}) - \frac{\partial \log z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \\
\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} &= -\frac{\partial^2 \log z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = -cov(u(\mathbf{x})) \quad \text{using the result of the proof written in Q.8} \\
-\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}] &= -\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}[-cov(u(\mathbf{x}))] = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}[cov(u(\mathbf{x}))] = cov(u(\mathbf{x})) = \mathbf{F}(\boldsymbol{\eta})
\end{aligned}
$$

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

$$
\begin{aligned}
\theta = \mathbb{E}[u(\mathbf{x})] &= \int \frac{u(\mathbf{x})h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})} dx \\
\frac{\partial \theta}{\partial \boldsymbol{\eta}} &= \int \frac{u(\mathbf{x})u(\mathbf{x})^\top h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})} dx - \int \frac{u(\mathbf{x})h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})^2} dx \frac{\partial z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \\
&= \int \frac{u(\mathbf{x})u(\mathbf{x})^\top h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})} dx - \int \frac{u(\mathbf{x})h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})^2} dx \int u(\mathbf{x})h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})dx \\
&= \mathbb{E}[u(\mathbf{x})u(\mathbf{x})^\top] - \left(\int \frac{u(\mathbf{x})h(x)exp(u(\mathbf{x})^\top\boldsymbol{\eta})}{z(\boldsymbol{\eta})} dx\right)^2 \\
&= \mathbb{E}[u(\mathbf{x})u(\mathbf{x})^\top] - \mathbb{E}[u(\mathbf{x})]\mathbb{E}[u(\mathbf{x})]^\top = cov(u(\mathbf{x})) = \mathbf{F}(\boldsymbol{\eta}) \qquad (8)
\end{aligned}
$$

(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, $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{F}^{-1}(\boldsymbol{\eta})$.

$$
\begin{aligned}
\mathbf{F}(\theta) &= -\mathbb{E}_{p(\mathbf{x}|\theta)}\left[\frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta^2}\right] = -\mathbb{E}_{p(\mathbf{x}|\theta)}\left[\frac{\partial}{\partial \theta}\left(\frac{\partial \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\frac{\partial \boldsymbol{\eta}}{\partial \theta}\right)\right] \\
&= -\mathbb{E}_{p(\mathbf{x}|\theta)}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}\left(\frac{\partial \boldsymbol{\eta}}{\partial \theta}\right)^2 + \frac{\partial \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\frac{\partial^2 \boldsymbol{\eta}}{\partial \theta^2}\right] \\
\frac{\partial \theta}{\partial \boldsymbol{\eta}} &= \mathbf{F}(\boldsymbol{\eta}) = cov(u(\mathbf{x})) \quad \text{using the result of Q.11 (a)} \qquad (1) \\
\frac{\partial^2 \theta}{\partial \boldsymbol{\eta}^2} &= \frac{\partial}{\partial \boldsymbol{\eta}} cov(u(\mathbf{x})) = 0 \qquad (2)
\end{aligned}
$$

Using eq. (1) and (2), we get :

$$
\begin{aligned}
\mathbf{F}(\theta) &= -\mathbb{E}_{p(\mathbf{x}|\theta)}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}\left(\frac{\partial \theta}{\partial \boldsymbol{\eta}}\right)^{-2}\right] = -\mathbb{E}_{p(\mathbf{x}|\theta)}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}\right](\mathbf{F}(\boldsymbol{\eta})\mathbf{F}(\boldsymbol{\eta})^\top)^{-1} \\
&= \mathbf{F}(\boldsymbol{\eta})\mathbf{F}(\boldsymbol{\eta})^{\top -1}(\mathbf{F}(\boldsymbol{\eta})^{-1} = \mathbf{F}(\boldsymbol{\eta})^{-1} \quad \text{[since } \mathbf{F}(\boldsymbol{\eta}) = cov(u(\mathbf{x})) \text{ and thus is symmetric in nature]}
\end{aligned}
$$

(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

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

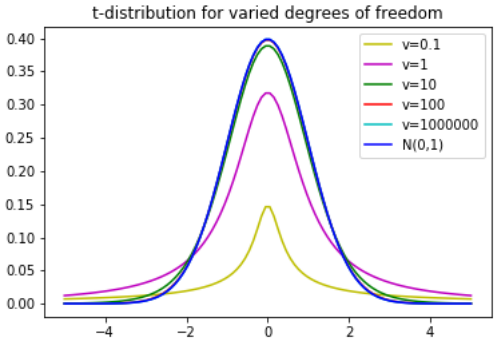
Note that I choose the orientation of the gradient vector to be consistent with Jacobian. So, in this case, the gradient vector is a row vector (rather than a column vector). If you want to use a column vector to represent the gradient, you can move the information matrix to the left. It does not influence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

$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\eta}}{\partial \theta} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \left(\frac{\partial \theta}{\partial \boldsymbol{\eta}} \right)^{-1} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \left(\mathbf{F}(\boldsymbol{\eta}) \right)^{-1} \quad \text{using the result of Q.11 (b)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{F}(\boldsymbol{\theta})^{-1} &= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \mathbf{F}(\boldsymbol{\eta})^{-1} \mathbf{F}(\boldsymbol{\theta})^{-1} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \left(\mathbf{F}(\boldsymbol{\theta}) \mathbf{F}(\boldsymbol{\eta}) \right)^{-1} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \left(\mathbf{F}(\boldsymbol{\eta})^{-1} \mathbf{F}(\boldsymbol{\eta}) \right)^{-1} \quad \text{using the result of Q. 11(c)} \\
&= \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \mathbf{I} = \frac{\partial \log p(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\eta}} \quad \mathbf{I} \text{ is the identity matrix}
\end{aligned}$$

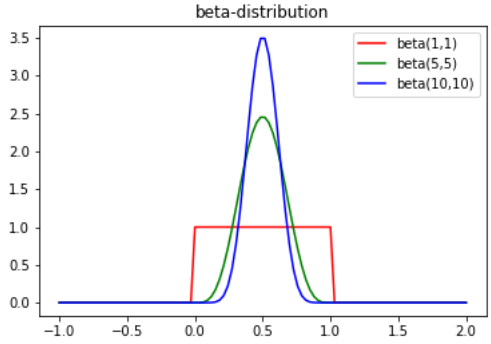
Programming Practice [20 points]

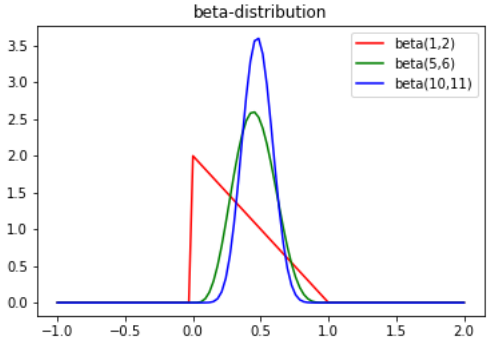
1. [5 Points] Look into the student t's distribution. Let us set the mean and precision to be $\mu = 0$ and $\lambda = 1$. Vary the degree of freedom $\nu = 0.1, 1, 10, 100, 10^6$ and draw the density of the student t's distribution. Also, draw the density of the standard Gaussian distribution $\mathcal{N}(0,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



As the degree of freedom increases, the t-distibution plot becomes similar to gaussian distribution with $\mu = 0$ and $\sigma = 1$

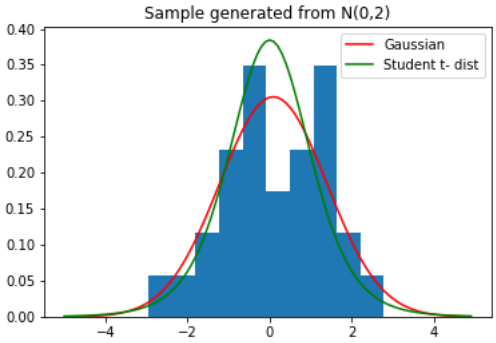
2. [5 points] Draw the density plots for Beta distributions: Beta(1,1), Beta(5, 5) and Beta (10, 10). Put the three density curves in one figure. What do you observe? Next draw the density plots for Beta(1, 2), Beta(5,6) and Beta(10, 11). Put the three density curves in another figure. What do you observe?

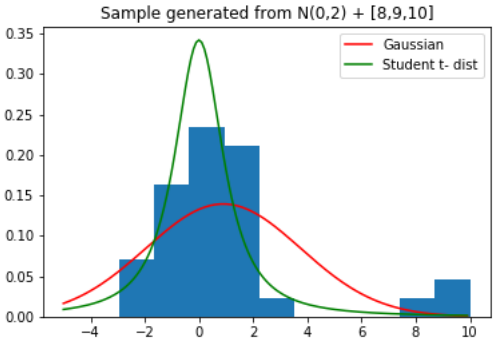




When $a = b$, the beta distribution is symmetric about 0.5 and as the value of a increases, the variance of the beta distribution decreases. When $a! = b$, the beta distribution is no more symmetric and shifts about left (as $a < b$). However, with the increase in a and b, the variance of the beta distribution decreases.

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?





We observed that the t-distribution is not affected much by the outliers but the Gaussian distribution got significantly affected.