Machine Learning 2 - Homework 1

Andrii Skliar, 11636785

deadline: April 11, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Gabriele Bani Gabriele Cesa Davide Belli Pascal Esser
Gautier Dagan

NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.

Problem 1: Consider two random vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^n$ having Gaussian distribution $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z})$. Consider random vector $\mathbf{y} = \mathbf{x} + \mathbf{z}$. Derive mean and covariance of $p(\mathbf{y})$.

Solution:

Note, that for derivation we have used the fact that for vectors: $\mathbf{x}\mathbf{z}^T = \mathbf{z}\mathbf{x}^T$.

$$
\begin{aligned}
\mathbb{E}[\mathbf{y}] &= \mathbb{E}[\mathbf{x} + \mathbf{z}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{z}] = \boldsymbol{\mu}_x + \boldsymbol{\mu}_z \\
Cov[\mathbf{y}, \mathbf{y}] &= \mathbb{E}[(\mathbf{y} - \mathbb{E}\,\mathbf{y})(\mathbf{y} - \mathbb{E}\,\mathbf{y})^T] \\
&= \mathbb{E}[(\mathbf{x} + \mathbf{z} - \mathbb{E}(\mathbf{x} + \mathbf{z}))(\mathbf{x} + \mathbf{z} - \mathbb{E}(\mathbf{x} + \mathbf{z}))^T] \\
&= \mathbb{E}[(\mathbf{x} + \mathbf{z} - \boldsymbol{\mu}_x - \boldsymbol{\mu}_z)(\mathbf{x} + \mathbf{z} - \boldsymbol{\mu}_x - \boldsymbol{\mu}_z)^T] \\
&= \mathbb{E}[(\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\boldsymbol{\mu}_x^T + \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T) + (\mathbf{z}\mathbf{z}^T - 2\mathbf{z}\boldsymbol{\mu}_z^T + \boldsymbol{\mu}_z\boldsymbol{\mu}_z^T) \\
&+ 2(\mathbf{x}\mathbf{z}^T - \mathbf{x}\boldsymbol{\mu}_z^T - \mathbf{z}\boldsymbol{\mu}_x^T + \boldsymbol{\mu}_x\boldsymbol{\mu}_z^T)] \\
&= \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T] + \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu}_z)(\mathbf{z} - \boldsymbol{\mu}_z)^T] + 2\,\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{z} - \boldsymbol{\mu}_z)^T] \\
&= \mathbb{V}(\mathbf{x}) + \mathbb{V}(\mathbf{z}) + 2 \cdot Cov(\mathbf{x}, \mathbf{z}) \\
&= \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_z + 2Cov(\mathbf{x}, \mathbf{z})
\end{aligned}
$$

If we assume, that $\mathbf{x}$ and $\mathbf{z}$ are independent: $Cov(\mathbf{x}, \mathbf{z}) = 0 \Longrightarrow Cov[\mathbf{y}, \mathbf{y}] = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_z$

Problem 2: Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in which the covariance $\boldsymbol{\Sigma}$ is known. Given a set of $\mathbf{N}$ observations $\mathcal{X} = \{\mathbf{x}_1 \dots \mathbf{x}_N\}$. Assume that $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. [Hint: you may use results from Bishop]

1. Write down the likelihood for the data $p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$;

Solution: **Note**: in following problems we use following properties:

- $\mathbf{x^T S y} = \mathbf{y^T S x}$, when S is symmetric
- $(\mathbf{x} - \mathbf{a})\mathbf{S}(\mathbf{x} - \mathbf{a}) = \mathbf{a^T S a} - \mathbf{2 x^T S a} + \mathbf{x^T S x}$

$$
\begin{aligned}
p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{n=1}^{N} p(x_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} N(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_{n=1}^{N} (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\} \\
&= (2\pi)^{-\frac{ND}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\}
\end{aligned}
$$

2. Write down the posterior $p(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$;

Solution:

$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &= \frac{p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{p(\mathcal{X})} \\
&= \frac{\big(\prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{\int_{-\infty}^{\infty} \big(\prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}', \boldsymbol{\Sigma})\big) p(\boldsymbol{\mu}'|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) d\boldsymbol{\mu}'} \\
&\quad \frac{\big(\prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{\int_{-\infty}^{\infty} \big(\prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}', \boldsymbol{\Sigma})\big) \mathcal{N}(\boldsymbol{\mu}'|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) d\boldsymbol{\mu}'} \\
&= (2\pi)^{-\frac{ND}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\} \\
&\quad \cdot (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\} \\
&= (2\pi)^{-\frac{(N+1)D}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \\
&\quad - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\}
\end{aligned}
$$

3. Show that $p(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ and find the values of $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$.

Solution: To find values of $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ we only need to look at the exponent in the numerator of the posterior distribution.

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\\
&=-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\\
&-\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i+\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i-\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}\\
&=-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i+\sum_{i=1}^{N}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\mu\\
&=-\frac{1}{2}\left[\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{x}_i)+\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\sum_{i=1}^{N}\mathbf{x}_i^T\boldsymbol{\Sigma}^T\mathbf{x}_i\right]\\
&=-\frac{1}{2}\left[\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{x}_i)+const\right]
\end{aligned}
$$

In order for the assumption that $p(\boldsymbol{\mu}|\mathcal{X},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)$ to hold, following equations should be satisfied:

$$
\begin{aligned}
\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}&=\boldsymbol{\mu}^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}\\
\boldsymbol{\Sigma}_N^{-1}&=(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})\\
\boldsymbol{\Sigma}_N&=(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})^{-1}\\
2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{x}_i)&=2\boldsymbol{\mu}^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}\\
\boldsymbol{\mu}_N&=\boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{x}_i)\\
\boldsymbol{\mu}_N&=(\boldsymbol{\Sigma}_0^{-1}+N\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{x}_i)
\end{aligned}
$$

Using these results we can prove that the constant part in the exponent is indeed the last term in exponent of $N(\boldsymbol{\mu}|\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)$. Now we can use found $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ to rewrite formula for posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{X},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)&=\frac{p(\mathcal{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}{p(\mathcal{X})}\\
&=\frac{\cancel{(2\pi)^{-\frac{(N+1)D}{2}}}\cancel{|\boldsymbol{\Sigma}|^{-\frac{N}{2}}}|\boldsymbol{\Sigma}_0|^{-\frac{1}{2}}\exp\{-\frac{1}{2}((\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)\}}{\int_{\boldsymbol{\mu}'}\cancel{(2\pi)^{-\frac{(N+1)D}{2}}}\cancel{|\boldsymbol{\Sigma}|^{-\frac{N}{2}}}|\boldsymbol{\Sigma}_0|^{-\frac{1}{2}}\exp\{-\frac{1}{2}((\boldsymbol{\mu}'-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu}'-\boldsymbol{\mu}_N)\}d\boldsymbol{\mu}'}\\
&=\frac{\exp\{-\frac{1}{2}((\boldsymbol{\mu}-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)\}}{\int_{\boldsymbol{\mu}'}\exp\{-\frac{1}{2}((\boldsymbol{\mu}'-\boldsymbol{\mu}_N)^T\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu}'-\boldsymbol{\mu}_N)\}d\boldsymbol{\mu}'}=N(\boldsymbol{\mu}|\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)
\end{aligned}
$$

4. Derive the maximum of posterior solution for $\boldsymbol{\mu}$;

Solution: The Maximum a Posteriori estimations of μ is the arg max of the posterior distribution: $\boldsymbol{\mu}_{MAP} = \arg\max_{\boldsymbol{\mu}} p(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, which in the case of Gaussian posterior simplifies down to simply taking a mean of the posterior.
Therefore, $\boldsymbol{\mu}_{MAP} = \boldsymbol{\mu}_N$.

Problem 3: Tossing a biased coin with probability that it comes up heasd is μ. [Hint: use Bishop]

1. We toss the coin 3 times and it all comes up with heads. How likely is that in the next toss, the coin comes up with head according to MLE?

Solution: I am using equation 2.7 from Bishop([1]): $\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$.
In our case, that equation tranforms into $\mu = \frac{1}{3}\sum_{i=1}^{3} 1 = 1$. This means that in the next toss, according to MLE, the coin comes up with head with probability 1.

2. Suppose that the prior $\mu \sim Beta(\mu|a, b)$. What is the probability that the coin comes up with head in the 4th toss?

Solution: **Note**: here we are using following formula for Beta distribution: $Beta(x|a, b) = \frac{\Gamma(a+b)}{\Gamma(b)\Gamma(a)} \cdot x^{a-1}(1-x)^{b-1}$ Assuming $\mu \sim \mathcal{B}(\mu|a, b)$, the highest possible probability that the coin comes up with head in the 4th toss can be estimated through MAP estimation:

$$\mu_{MAP} = \arg\max_{\mu} p(\mu|a, b, \mathcal{D}) = \arg\max_{\mu} \log p(\mu|a, b, \mathcal{D}) = \arg\max_{\mu} \log p(\mathcal{D}|\mu)p(\mu|a, b)$$

$$\begin{aligned}
\frac{\partial}{\partial\mu} \log p(\mathcal{D}|\mu)p(\mu|a, b) &= \frac{\partial}{\partial\mu} \log \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{\partial}{\partial\mu} \Bigg[\sum_{i=1}^{N} \Big[x_i \log\mu + (1-x_i)\log(1-\mu) \Big] + \log\Gamma(b+a) - \log\Gamma(b) - \log\Gamma(a) \\
&\quad + (a-1)\log\mu + (b-1)\log(1-\mu) \Bigg] \\
&= \sum_{i=1}^{N} \Big[\frac{x_i}{\mu} - \frac{1-x_i}{1-\mu} \Big] + \frac{a-1}{\mu} - \frac{b-1}{1-\mu} \\
&= \sum_{i=1}^{N} \frac{x_i - x_i\mu - \mu + \mu x_i}{\mu(1-\mu)} + \frac{a-1-a\mu+\mu-b\mu+\mu}{\mu(1-\mu)} \\
&= \frac{1}{\mu(1-\mu)} \Big[\sum_{i=1}^{N} x_i - N\mu + a - 1 - a\mu - b\mu + 2\mu \Big]
\end{aligned}$$

To find MAP estimation, we need to set $\frac{\partial}{\partial \mu} \log p(\mathcal{D}|\mu)p(\mu|a,b)$ to 0:

$$\frac{1}{\mu(1-\mu)}\Big[\Big(\sum_{i=1}^{N} x_i\Big) - N\mu + a - 1 - a\mu - b\mu + 2\mu\Big] = 0$$

$$\sum_{i=1}^{N} x_i - N\mu + a - 1 - a\mu - b\mu + 2\mu = 0$$

$$\sum_{i=1}^{N} x_i - \mu(N + a + b - 2) + a - 1 = 0$$

$$\mu(N + a + b - 2) = \sum_{i=1}^{N} x_i + a - 1$$

$$\mu_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{N + a + b - 2}$$

In our case, $p(x = 1|\mu_{MAP}, \mathcal{D}) = \mu_M AP = \frac{\sum_{i=1}^{N} x_i + a - 1}{N+a+b-2} = \frac{3+a-1}{3+a+b-2} = \frac{2+a}{1+a+b}$

However, we can write full posterior distribution for $p(\mu|a, b, \mathcal{D})$ as this allows us to use more Bayesian approach and not only use mode of the posterior (which is equal to MAP), but also, for example, use mean of the posterior for predicting the probability of heads in the next toss if we decide to do so.

Note: here we use Bishop notation, where m is the number times that the coin lands heads and l is the number of times that it lands tails.

$$p(\mu|\mathcal{D}) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}$$

$$= \mu^m \cdot (1-\mu)^l$$

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu) \cdot p(\mu|a,b)}{p(\mathcal{D})}$$

$$= \frac{\mu^m \cdot (1-\mu)^l \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}\mu^{a-1}(1-\mu)^{b-1}}{\int_0^1 \left((\mu')^m \cdot (1-\mu')^l \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}\mu'^{a-1}(1-\mu')^{b-1}\right) d\mu'}$$

Note: here $\mathcal{B}(\theta_0, \theta_1)$ means Beta-function with θ_0 and θ_1 as parameters.
First, we should calculate an integral in the denominator to make further calculations easier and less cumbersome.

$$\int_0^1 \left((\mu')^m \cdot (1-\mu')^l \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}(\mu')^{a-1}(1-\mu')^{b-1}\right) d\mu' = \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \int_0^1 \left((\mu')^{m+a-1}(1-\mu')^{l+b-1}\right) d\mu'$$

$$= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \mathcal{B}(m+a, l+b)$$

$$= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \frac{\Gamma(m+a)\Gamma(l+b)}{\Gamma(N+a+b)}$$

Using previously calculated integral, we get following:

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \\
&= \frac{\mu^m \cdot (1-\mu)^l \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}\mu^{a-1}(1-\mu)^{b-1}}{\int_0^1 \left((\mu')^m \cdot (1-\mu')^l \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}\mu'^{a-1}(1-\mu')^{b-1}\right) d\mu'} \\
&= \frac{\mu^m \cdot (1-\mu)^l \cdot \cancel{\frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}}\mu^{a-1}(1-\mu)^{b-1}}{\cancel{\frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}} \cdot \frac{\Gamma(m+a)\Gamma(l+b)}{\Gamma(N+a+b)}} \\
&= \frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(l+b)} \cdot \mu^{m+a-1}(1-\mu)^{l+b-1} \\
&= Beta(\mu|m+a, l+b)
\end{aligned}
$$

Note: here the last $Beta(\mu|m+a, l+b)$ means that $p(\mu|\mathcal{D})$ is distributed by Beta distribution with parameters $m+a$ *and* $l+b$.

This allows us to calculate mean of the posterior, which will be the probability of having head in the next coin toss as proved in Bishop , equation (2.19):
$\mathbb{E}[\mu|\mathcal{D}] = \frac{m+a}{N+a+b} = \{\text{with plugged-in numbers}\} = \frac{3+a}{3+a+b}$.

3. Suppose that we observe m times that the coin lands heads and l times that it lands tails. Show that the posterior mean lies between the prior mean and μ_{MLE}.

Solution: Given that $\mu \sim \mathcal{B}(\mu|a, b)$, prior mean can be calculated as mean of Beta distribution: $\mathbb{E}(\mu|a, b) = \frac{a}{a+b}$. We also know, that $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{m}{N}$ and $\mathbb{E}(\mu|\mathcal{D}) = \frac{m+a}{m+l+a+b}$. To prove that the posterior mean lies between the prior mean and μ_{MLE}, I will prove that posterior mean can be expressed as a convex combination of the prior mean and μ_{MLE}:

$$
\begin{aligned}
\frac{m+a}{m+l+a+b} &= \frac{m}{m+l+a+b} + \frac{a}{m+l+a+b} \\
&= \frac{m+l}{m+l}\frac{m}{m+l+a+b} + \frac{a+b}{a+b}\frac{a}{m+l+a+b} \\
&= \frac{m}{m+l}\frac{m+l}{m+l+a+b} + \frac{a}{a+b}\frac{a+b}{m+l+a+b} \\
&= \mu_{MLE}\frac{m+l}{m+l+a+b} + \mathbb{E}(\mu|a, b)\frac{a+b}{m+l+a+b} \\
&= \alpha_1\mu_{MLE} + \alpha_2\,\mathbb{E}(\mu|a, b), \text{ where } \alpha_1 + \alpha_2 = 1
\end{aligned}
$$

From the definition, of convex combination, follows that that the posterior mean lies between the prior mean and μ_{MLE}.

Problem 4:

- $Pois(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $Gamma(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}$
- $Cauchy(x|\gamma, \mu) = \frac{1}{\pi\gamma} \frac{1}{1+(\frac{x-\mu}{\gamma})^2}$
- $vonMises(x|\kappa, \mu) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa cos(x-\mu)}$

1. Are the above distributions members of an exponential family. If yes, then (a) cast them in exponential form (Bishop eq. 2.194) with a minimum numbers of parameters, (b) derive their sucient statistics.

Solution: Note: using Bishop notation for exponential family. $u(\cdot)$ is the *sufcent statistic*.

(a) $Pois(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1}{k!} e^{k \log \lambda} e^{-\lambda}$
This results in following parameters for exponential family:

$$\begin{aligned} h(x) &= \frac{1}{x!} \\ \eta &= \log \lambda \Longrightarrow \lambda = e^{\eta} \\ g(\eta) &= e^{-\lambda} = e^{-e^{\eta}} \\ u(x) &= x \end{aligned}$$

(b)

$$\begin{aligned} Gamma(\tau|a, b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} \\ &= \frac{b^a}{\Gamma(a)} e^{(a-1 \log \tau - b\tau)} \\ &= \frac{b^a}{\Gamma(a)} \exp\left\{ \begin{pmatrix} a-1 \\ b \end{pmatrix}^T \begin{pmatrix} \log \tau \\ \tau \end{pmatrix} \right\} \end{aligned}$$

This results in following parameters for exponential family:

$$\begin{aligned} h(x) &= 1 \\ \boldsymbol{\eta} &= \begin{pmatrix} a-1 \\ b \end{pmatrix} \Longrightarrow \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} a-1 \\ -b \end{pmatrix} \Longrightarrow \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \eta_1 + 1 \\ -\eta_2 \end{pmatrix} \\ g(\boldsymbol{\eta}) &= \frac{b^a}{\Gamma(a)} = \frac{\eta_2^{\eta_1+1}}{\Gamma(\eta_1+1)} \\ \mathbf{u}(x) &= \begin{pmatrix} \log x \\ x \end{pmatrix} \end{aligned}$$

(c) $Cauchy(x|\gamma, \mu) = \frac{1}{\pi\gamma} \frac{1}{1+(\frac{x-\mu}{\gamma})^2}$
From the properties of the exponential family, it follows that any distribution, that belongs to the exponential family, needs to have finite expectation value. Therefore, it is enough to prove that for at least one set of parameters for a

specific distribution expectation value is infinite (for general exponential family). To simplify integral calculation, I would like to calculate expectation value of $Cauchy(x|1,0)$.

$$\begin{aligned}\mathbb{E}(x|1,0) &= \int_{-\infty}^{+\infty} \frac{x}{(1+x^2)\pi} \\ &= \frac{2}{\pi}\int_{0}^{+\infty} \frac{x}{(1+x^2)} \\ &= \frac{1}{\pi}\log(1+x^2)\Big|_0^{+\infty} \\ &= +\infty\end{aligned}$$

Therefore, expectation value is infinity, which means, that Cauchy distribution is not member of the general exponential family. Although parametric distributions can be not members of general exponential family, they can become members of it once one of the parameters is fixed. We don't consider that case. However, if we don't fix the parameters and look at the general form of the Cauchy distribution, we also find that it doesn't have finite mean, therefore, doesn't belong to the exponential family as well.

(d)

$$\begin{aligned}vonMises(x|\kappa,\mu) &= \frac{1}{2\pi I_0(\kappa)} e^{\kappa cos(x-\mu)} \\ &= \frac{1}{2\pi I_0(\kappa)} e^{\kappa(\cos x\cos\mu+\sin x\sin\mu)} \\ &= \frac{1}{2\pi I_0(\kappa)} e^{\begin{pmatrix}\kappa\cos\mu\\ \kappa\sin\mu\end{pmatrix}^T\begin{pmatrix}\cos x\\ \sin x\end{pmatrix}}\end{aligned}$$

This results in following parameters for exponential family:

$$\begin{aligned}h(x) &= \frac{1}{2\pi} \\ \boldsymbol{\eta} &= \begin{pmatrix}\kappa\cos\mu\\ \kappa\sin\mu\end{pmatrix} \Longrightarrow \kappa = \sqrt{\eta_1^2+\eta_2^2} \\ g(\boldsymbol{\eta}) &= \frac{1}{I_0(\kappa)} = \frac{1}{I_0(\sqrt{\eta_1^2+\eta_2^2})} \\ \mathbf{u}(x) &= \begin{pmatrix}\cos x\\ \sin x\end{pmatrix}\end{aligned}$$

2. Derive the first and second moment of the distributions (i) and (ii).

Solution: Moments can be both raw and central, however, from the general point of the task, I have concluded that we were asked to calculate central moments.

(a) $Pois(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

$$\mathbb{E}[k] = \mathbb{E}[u(x)] = -\frac{\partial}{\partial \eta} \log e^{-e^{\eta}} = -\frac{\partial}{\partial \eta}(-e^{\eta}) = e^{\eta} = e^{\log \lambda} = \lambda$$
$$\mathbb{E}[k^2] = \mathbb{E}[u^2(x)] = -\frac{\partial^2}{\partial \eta^2} \log e^{-e^{\eta}} = e^{\eta} = e^{\log \lambda} = \lambda$$

(b) $Gamma(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}$

$$\begin{aligned}\mathbb{E}[\tau] = \mathbb{E}[u_2(x)] &= -\frac{\partial}{\partial \eta_2} \log \frac{-\eta_2^{\eta_1+1}}{\Gamma(\eta_1+1)} = -\frac{\partial}{\partial \eta_2}(\eta_1+1)\log(-\eta_2) - \log\Gamma(\eta_1+1) \\ &= (\eta_1+1)[-\frac{\partial}{\partial \eta_2}\log(-\eta_2)] = -\frac{\eta_1+1}{\eta_2} = -\frac{a-1+1}{-b} = \frac{a}{b} \\ \mathbb{E}[\tau^2] = \mathbb{E}[u_2^2(x)] &= (\eta_1+1)[-\frac{\partial^2}{\partial \eta_2^2}\log(-\eta_2)] = (\eta_1+1)[-(-\frac{1}{\eta_2^2})] \\ &= \frac{\eta_1+1}{\eta_2^2} = \frac{a}{b^2}\end{aligned}$$

3. Does the Poisson distribution have a conjugate prior? Derive the conjugate prior, if the answer is yes.

Solution: Yes, it does. I would like to derive it using that conjugate prior has Gamma distribution, so I only need to prove that posterior distribution will also be distributed as Gamma distribution.

Note: here following notation is used for convenience: $(d_1, d_2, \dots d_T)$ is a set of observations, $n = \sum_{t=1}^{T} d_t$ and $(d_1, d_2, \dots d_T) = \mathbf{d}$

$$\begin{aligned}p(x|\lambda) &= \frac{\lambda^x}{x!}e^{-\lambda} \\ p(\mathbf{d}|\lambda) &= \prod_{t=1}^{T}\frac{\lambda^{d_t}}{d_t!}e^{-\lambda} \\ &= \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t!}e^{-T\lambda} \\ &= \frac{\lambda^n}{\prod_{t=1}^{T} d_t!}e^{-T\lambda}\end{aligned}$$

Note: here $\int_\lambda$ means integrating over set of all possible values of λ.

$$
\begin{aligned}
p(\lambda|\mathbf{d}) &= \frac{p(\mathbf{d}|\lambda)p(\lambda|a,b)}{\int_\lambda p(\mathbf{d}|\lambda)p(\lambda|a,b)d\lambda} \\
&= \frac{\prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)}{\int_\lambda \prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)d\lambda} \\
&= \frac{\prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)}{\int_0^{+\infty} \prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)d\lambda} \\
&= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{\int_0^{+\infty} \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda}
\end{aligned}
$$

Note: here $\Gamma(\theta)$ means Gamma-function with θ as a parameter.
First, we should calculate an integral in the denominator to make further calculations easier and less cumbersome.

$$
\begin{aligned}
\int_0^{+\infty} \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!} \int_0^{+\infty} \lambda^{\sum_{t=1}^{T} d_t} e^{-T\lambda - b\lambda} \lambda^{a-1} d\lambda \\
&= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!} \int_0^{+\infty} \lambda^{n+a-1} e^{-(T+b)\lambda} d\lambda \\
&= \left\{ \lambda' = (T+b)\lambda;\ d\lambda' = (T+b)d\lambda;\ \lambda = \frac{\lambda'}{T+b};\ d\lambda = \frac{d\lambda'}{T+b} \right\} \\
&= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!} \int_0^{+\infty} \left(\frac{\lambda'}{T+b} \right)^{n+a-1} e^{-\lambda'} \frac{1}{T+b} d\lambda' \\
&= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!} \cdot \frac{1}{(T+b)^{n+a}} \int_0^{+\infty} {\lambda'}^{n+a-1} e^{-\lambda'} d\lambda' \\
&= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!} \cdot \frac{1}{(T+b)^{n+a}} \Gamma(n+a)
\end{aligned}
$$

Using previously calculated integral, we get following:

$$
\begin{aligned}
p(\lambda|\mathbf{d}) &= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{\int_0^{+\infty} \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda} \\
&= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{\frac{b^a}{\Gamma(a) \prod_{t=1}^{T} d_t!} \cdot \frac{1}{(T+b)^{n+a}} \Gamma(n+a)} \\
&= \frac{\cancel{\frac{\lambda^n}{\prod_{t=1}^{T} d_t}} e^{-T\lambda} \cdot \cancel{\frac{b^a}{\Gamma(a)}} \lambda^{a-1} e^{-b\lambda}}{\cancel{\frac{b^a}{\Gamma(a) \prod_{t=1}^{T} d_t!}} \cdot \frac{1}{(T+b)^{n+a}} \Gamma(n+a)} \\
&= \frac{\lambda^n e^{-T\lambda} \lambda^{a-1} e^{-b\lambda} (T-b)^{n+a}}{\Gamma(n+a)} \\
&= \frac{\lambda^{n+a-1} e^{-(T+b)\lambda} (T+b)^{n+a}}{\Gamma(n+a)} \\
&= \frac{(T+b)^{n+a}}{\Gamma(n+a)} \lambda^{(n+a)-1} e^{-(T+b)\lambda} \\
&= Gamma(\lambda|n+a, T+b)
\end{aligned}
$$

Note: here the last $Gamma(\lambda|n+a, T+b)$ means that $p(\lambda|\mathbf{d})$ is distributed by Gamma distribution with parameters $n+a$ *and* $T+b$. So, as you can see, Poisson distribution indeed has conjugate prior, which is distributed using Gamma distribution.

Problem 5: * Derive mean, covariance, and mode of multivariate Student's t-distribution.

Solution: Mean:

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}-\boldsymbol{\mu}] &= \int_{-\infty}^{+\infty} \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{U} (\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{D+\nu}{2}} (\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x} \\
&= const \cdot \int_{-\infty}^{+\infty} \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{U} (\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{D+\nu}{2}} (\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x} \\
&= \{\text{using change of variables:} (\mathbf{x}-\boldsymbol{\mu}) = \mathbf{y}; d\mathbf{x} = d\mathbf{y}\} \\
&= const \cdot \int_{-\infty}^{+\infty} \left[1 + \frac{1}{\nu}\mathbf{y}^T \mathbf{U} \mathbf{y}\right]^{-\frac{D+\nu}{2}} \mathbf{y} d\mathbf{y} = 0
\end{aligned}
$$

This equation can be easily proved as the function over which we integrate is an odd function, therefore, its integral over a symmetric interval will be 0.
Therefore: $\mathbb{E}[\mathbf{x}-\boldsymbol{\mu}] = \mathbb{E}[\mathbf{x}] - \mathbb{E}[\boldsymbol{\mu}] = 0 \Longrightarrow \mathbb{E}[\mathbf{x}] = \mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}$
Mode:
If we take a look at $\left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{U} (\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{D+\nu}{2}}$, we can notice, that $(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{U} (\mathbf{x}-\boldsymbol{\mu})$ is even function, which has its minimum at $\boldsymbol{\mu}$ as $\mathbf{U}$ is positive-definite.

Therefore, $\left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{U}(\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{D+\nu}{2}}$, has its maximum value for $\boldsymbol{\mu}$, thus, $\boldsymbol{\mu}$ is mode.
Variance:
We will use the following property of multivariate Student distribution [1]: If $\mathbf{x} \sim T(\mathbf{x}|\boldsymbol{\mu}, \mathbf{U}, \nu)$, $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{Z}$, where $\mathbf{Z}$ is a Dx1 vector having a standard multivariate Student t distribution with ν degrees of freedom and $\boldsymbol{\Sigma}$ is a DxD invertible matrix such that $\mathbf{U} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$.
Then, calculating variance is very simple:

$$\mathbb{V}[\mathbf{x}] = \mathbb{V}[\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{Z}] = \boldsymbol{\Sigma}\,\mathbb{V}[\mathbf{Z}]\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}\frac{n}{n-2}\mathbf{I}\boldsymbol{\Sigma}^T = \frac{n}{n-2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \frac{n}{n-2}\mathbf{U}$$

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.