Assignment 5

Marion Rosec

Decembre 2023

Exercise 1

We can first find the MLE by taking the product of the PMF for each of the X value $(X_1, X_2, ..., X_n)$:

$$L(\theta) = \prod_{i=1}^{n} (1-\theta)^{x_i - 1}\theta$$

We can then take the log of the MLE expression :

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} [\log(1-\theta)(x_i - 1) + \log \theta]$$

To compute the $\hat{\theta}_n$ that optimized the probability of the observed data $X_1, X_2, ..., X_n$, we need to solve the following equation for θ :

$$\begin{aligned}
&\frac{\partial l}{\partial \theta} = 0 \\
\Leftrightarrow &\sum_{i=1}^{n} [-\frac{x_i - 1}{1-\theta} + \frac{1}{\theta}] = 0 \\
\Leftrightarrow &\sum_{i=1}^{n} [-\frac{\theta(1-\theta)(x_i - 1)}{1-\theta} + \frac{\theta(1-\theta)}{\theta}] = 0 \\
\Leftrightarrow &-\sum_{i=1}^{n} [\theta x_i + \theta + 1 - \theta] = 0 \\
\Leftrightarrow &-\theta \sum_{i=1}^{n} (x_i) + n = 0 \\
\Leftrightarrow &\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}$$

To determine whether this is a maximum or not, we need to compute the second derivative :

$$\begin{aligned}
\frac{\partial l}{\partial \theta} &= \sum_{i=1}^{n} [\frac{1}{\theta} - \frac{x_i - 1}{1-\theta}] \\
\frac{\partial^2 l}{\partial^2 \theta} &= \sum_{i=1}^{n} [-\frac{1}{\theta^2} - \frac{x_i - 1}{(1-\theta)^2}] \\
\frac{\partial^2 l}{\partial^2 \theta} &= -\frac{n}{\theta^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1-\theta)^2}
\end{aligned}$$

n is positive (or equal to zero), and we also know that $x\mathbf{Z}_+$ (positive, zero excluded). Therefore, $\sum_{i=1}^{n} x_i \geq n$ and we can say that $\frac{\partial^2 l}{\partial^2 \theta} < 0$. The estimator is a maximum.

Exercise 2

a)

We know that the integral of a pdf is equal to 1 so we can compute c using this property :

$$\int_{x_{min}}^{x_{max}} \int_{y_{min}}^{y_{max}} f(x,y)dxdy = 1$$

$$\Leftrightarrow \int_{x_{min}}^{x_{max}} \int_{y_{min}}^{y_{max}} c\ dxdy = 1$$

$$\Leftrightarrow [cx]_{x_{min}}^{x_{max}} [cy]_{y_{min}}^{y_{max}} = 1$$

$$\Leftrightarrow c(x_{max} - x_{min})(y_{max} - y_{min}) = 1$$
$$\Leftrightarrow c = \frac{1}{(x_{max} - x_{min})(y_{max} - y_{min})}$$

b)

In both cases the four points are in the window defined by the theta parameter, so the likelihoods are :

$$L_{\theta_1} = f_{\theta_1}(0,0)f_{\theta_1}(0,1)f_{\theta_1}(1,1)f_{\theta_1}(2,2) = c^4 = (\frac{1}{(4+1)(3+1)})^4 = \frac{1}{20^4}$$
$$L_{\theta_2} = f_{\theta_2}(0,0)f_{\theta_2}(0,1)f_{\theta_2}(1,1)f_{\theta_2}(2,2) = c^4 = (\frac{1}{(5+2)(6+3)})^4 = \frac{1}{63^4}$$

c)

In order to find the maximum likelihood estimator, we need to find the smallest window in which all the points are included. For that we need $\hat{x_{min}}$ and $\hat{y_{min}}$ to be equal to 0, and $\hat{x_{max}}$ and $\hat{y_{max}}$ to be equal to 2. We then get the following MLE :

$$\theta^{\hat{ML}} = (0,2,0,2)$$

With this estimator, the value of c would be :

$$c = \frac{1}{(2-0)(2-0)}$$
$$c = \frac{1}{4}$$

Exercise 3

For all a, b and c questions, we know that the prior is a beta distribution (with different parameters alpha and beta) and the likelihood is normally distributed. Therefore, for all the questions, the posterior distribution will be :

$$
\begin{aligned}
p(r|y_N) &\propto p(y_N|t)p(r) \\
p(r|y_N) &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} \times \begin{pmatrix} N \\ y_N \end{pmatrix} r^{y_N}(1-r)^{N-y_N}
\end{aligned}
$$

We don't need the constant parts that don't involve r so we have :

$$
\begin{aligned}
p(r|y_N) &\propto r^{\alpha-1}(1-r)^{\beta-1} \times r^{y_N}(1-r)^{N-y_N} \\
p(r|y_N) &\propto r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1}
\end{aligned}
$$

We can see that the posterior is the same shape as a beta distribution :

$$
\begin{aligned}
p(r|y_N) &\propto Beta(\delta, \gamma) \\
p(r|y_N) &\propto r^{\delta-1}(1-r)^{\gamma-1}
\end{aligned}
$$

With $\delta = \alpha + y_N$ and $\gamma = \beta + N - y_N$

a)

In this first case, $\alpha = \beta = 1$ which means our new parameters are :

$$
\begin{aligned}
\delta &= 1 + y_N \\
\gamma &= 1 + N + y_N
\end{aligned}
$$

The posterior density is then :

$$
p(r|y_N) \propto r^{y_N}(1-r)^{N+y_N}
$$

b)

First, let's find the parameters α and β for the prior :

$$
\begin{aligned}
p(r) = 2r \Leftrightarrow r^{\alpha-1}(1-r)^{\beta-1} &\propto 2r \\
&\propto 2r^1(1-r)^0 \\
&\propto r^1(1-r)^0
\end{aligned}
$$

We can then solve the following equations :

$$
\begin{cases} \alpha - 1 &= 1 \\ \beta - 1 &= 0 \end{cases} \iff \begin{cases} \alpha &= 2 \\ \beta &= 1 \end{cases}
$$

We can now find the parameters δ and γ for the posterior :

$$
\begin{aligned}
p(r|y_N) &\propto p(y_N|t)p(r) \\
p(r|y_N) &\propto r^{y_N}(1-r)^{N-y_N} \times 2r \\
p(r|y_N) &= 2r^{y_N+1}(1-r)^{N-y_N} \propto r^{y_N+1}(1-r)^{N-y_N}
\end{aligned}
$$

The prior is still a beta distribution, so the posterior is also a beta distribution. Therefore, we can write it as so :

$$
p(r|y_N) \propto r^{\delta-1}(1-r)^{\gamma-1}
$$

with :

$$
\begin{cases} \delta - 1 &= y_N + 1 \\ \gamma - 1 &= N - y_N \end{cases} \iff \begin{cases} \delta &= y_N + 2 \\ \gamma &= N - y_N + 1 \end{cases}
$$

c)

First, let's find the parameters α and β for the prior :

$$\begin{aligned} p(r) = 3r^2 \Leftrightarrow r^{\alpha-1}(1-r)^{\beta-1} &\propto 2r \\ &\propto 3r^2(1-r)^0 \\ &\propto r^2(1-r)^0 \end{aligned}$$

We can then solve the following equations :

$$\begin{cases} \alpha - 1 & = 2 \\ \beta - 1 & = 0 \end{cases} \iff \begin{cases} \alpha & = 3 \\ \beta & = 1 \end{cases}$$

We can now find the parameters δ and γ for the posterior :

$$\begin{aligned} &p(r|y_N) \propto p(y_N|t)p(r) \\ &p(r|y_N) \propto r^{y_N}(1-r)^{N-y_N} \times 3r2 \\ &p(r|y_N) = 3r^{y_N+2}(1-r)^{N-y_N} \propto r^{y_N+2}(1-r)^{N-y_N} \end{aligned}$$

The prior is still a beta distribution, so the posterior is also a beta distribution. Therefore, we can write it as so :

$$p(r|y_N) \propto r^{\delta-1}(1-r)^{\gamma-1}$$

with :

$$\begin{cases} \delta - 1 & = y_N + 2 \\ \gamma - 1 & = N - y_N \end{cases} \iff \begin{cases} \delta & = y_N + 3 \\ \gamma & = N - y_N + 1 \end{cases}$$

Exercise 4

a)

As, $p(t|X, w, \sigma^2)$, the likelihood function, we can rewrite it as the product of all the likelihood for each the data points.

$$p(t|X, w, \sigma^2) = \prod_{n=1}^{N} p(t_n|X_n, w, \sigma^2)$$

And here, since our noise in our generative model, is an i.i.d a normal variable we can rewrite this product as :

$$p(t|X, w, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(w^T X_n, \sigma^2) = \mathcal{N}(Xw, \sigma^2 I) = \mathcal{N}(Xw, 10I)$$

b)

$$p(w|t, X, \sigma^2) = \frac{p(t|X, w, \sigma^2)p(w)}{p(t|X, \sigma^2)}$$

We don't need to keep the normalization because $p(w|t,X,\sigma^2)$ is fixed and therefore $p(t|X,\sigma^2)$ is a constant. If we omit that term, we have :

$$p(w|t,X,\sigma^2) \propto p(t|X,w,\sigma^2)p(w)$$

$$p(w|t,X,\sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2 I}}\exp(-\frac{1}{2}(\frac{(t-Xw)^T(t-Xxw)}{\sigma^2 I}) \times \frac{1}{\sqrt{2\pi\Sigma_0}}\exp(-\frac{1}{2}\frac{(w-\mu_0)^T(w-\mu_0)}{\Sigma_0})$$

Here the constant parts of the expression $\frac{1}{\sqrt{2\pi\sigma^2 I}}$ and $\frac{1}{\sqrt{2\pi\Sigma_0}}$ can be omitted as well as they don't involve w :

$$\begin{aligned}
p(w|t,X,\sigma^2) &\propto \exp(-\frac{1}{2}(\frac{(t-Xw)^T(t-Xxw)}{\sigma^2 I}) \times \exp(-\frac{1}{2}\frac{(w-\mu_0)^T(w-\mu_0)}{\Sigma_0}) \\
&\propto \exp(-\frac{1}{2}(\frac{(t-Xw)^T(t-Xxw)}{\sigma^2 I} + \frac{(w-\mu_0)^T(w-\mu_0)}{\Sigma_0}) \\
&\propto \exp(-\frac{1}{2}(\frac{-t^TXw + t^Tt - X^Tw^Tt + X^Tw^TXw}{\sigma^2 I} + \frac{w^Tw - w^T\mu_0 - \mu_0^Tw - \mu_0^T\mu_0}{\Sigma_0})
\end{aligned}$$

Again, we don't need the terms that don't involve w :

$$p(w|t,X,\sigma^2) \propto \exp(-\frac{1}{2}(-\frac{1}{\sigma^2}2t^TXw + X^Tw^TXw + \frac{w^Tw - 2\mu_0^Tw}{\Sigma_0})$$

Since the prior distribution is normal, we know that the posterior is also a Gaussian. We can therefore write :

$$\begin{aligned}
p(w|t,X,\sigma^2) &= \mathcal{N}(\mu,\Sigma) \\
p(w|t,X,\sigma^2) &\propto \exp(-\frac{1}{2}\frac{(w-\mu)^T(w-\mu)}{\Sigma}) \\
p(w|t,X,\sigma^2) &\propto \exp(-\frac{1}{2}\frac{(w^Tw - w^T\mu - \mu^Tw + \mu^T\mu}{\Sigma})
\end{aligned}$$

We don't need terms without w in them :

$$p(w|t,X,\sigma^2) \propto \exp(-\frac{1}{2}\frac{(w^Tw - 2\mu^Tw}{\Sigma})$$

Now that we have a similar expression to the one computed above, we can find the value of μ and Σ by identity :

$$\begin{aligned}
\frac{w^Tw}{\Sigma} &= \frac{1}{\sigma^2}w^TX^TXw + \frac{w^Tw}{\Sigma_0} \\
\frac{w^Tw}{\Sigma} &= w^T(\frac{1}{\sigma^2}X^TX + \frac{1}{\Sigma_0})w \\
\frac{1}{\Sigma} &= \frac{1}{\sigma^2}X^TX + \frac{1}{\Sigma_0} \\
\Sigma &= (\frac{1}{\sigma^2}X^TX + \frac{1}{\Sigma_0})^{-1}
\end{aligned}$$

$$\begin{aligned}
-\frac{2\mu^Tw}{\Sigma} &= -\frac{1}{\sigma^2}2t^TXw - \frac{2\mu_0^Tw}{\Sigma_0} \\
\mu^Tw &= (\frac{1}{\sigma^2}t^TXw + \frac{\mu_0^Tw}{\Sigma_0})\Sigma \\
\mu^T &= (\frac{1}{\sigma^2}t^TX + \frac{\mu_0^T}{\Sigma_0})\Sigma \\
\mu &= \Sigma(\frac{1}{\sigma^2}X^Tt + \frac{\mu_0}{\Sigma_0})
\end{aligned}$$

Therefore, the posterior distribution is :

$$p(w|t, X, \sigma^2) = \mathcal{N}(\mu, \Sigma)$$

with $\Sigma = (\frac{1}{\sigma^2}X^TX + \frac{1}{\Sigma_0})^{-1}$ and $\mu = \Sigma(\frac{1}{\sigma^2}X^Tt + \frac{\mu_0}{\Sigma_0})$

c)

Python

```python
    # **Exercise 2c)**
sigma_square = 10
mu_0 = np.array([[0], [0]])
Sigma_0 = np.array([[100, 0], [0, 5]])

def posterior(X, t, Sigma_0, mu_0):
    Sigma = np.linalg.inv(1 / sigma_square * X.T @ X + np.linalg.inv(Sigma_0))
    mu = Sigma @ (1 / sigma_square * X.T @ t + np.linalg.inv(Sigma_0) @ mu_0)
    return mu, Sigma
```

d)

Python

```python
>>> muw = [[10.99417141]
           [-0.04578724]]
>>> Sigmaw = [[ 1.31233642 -0.06718612]
              [-0.06718612  0.00478513]]
```

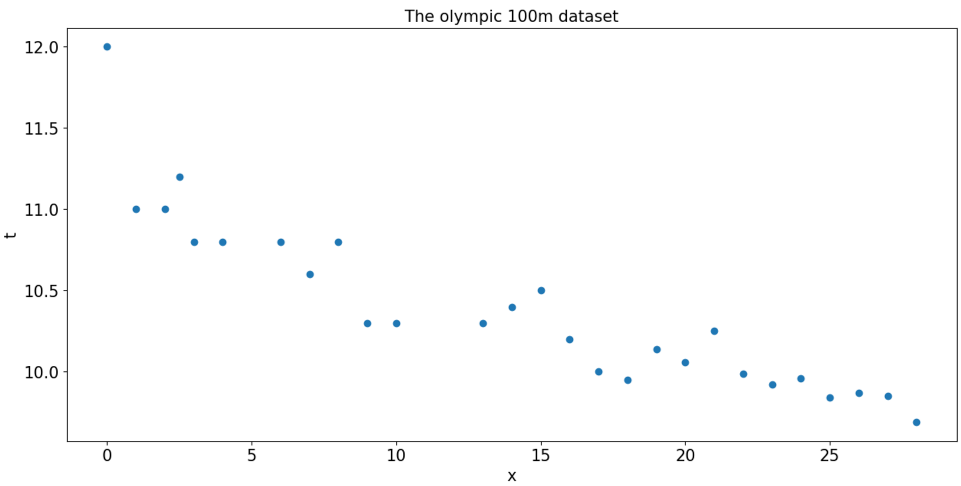

Figure 1 – Olympic 100m data set

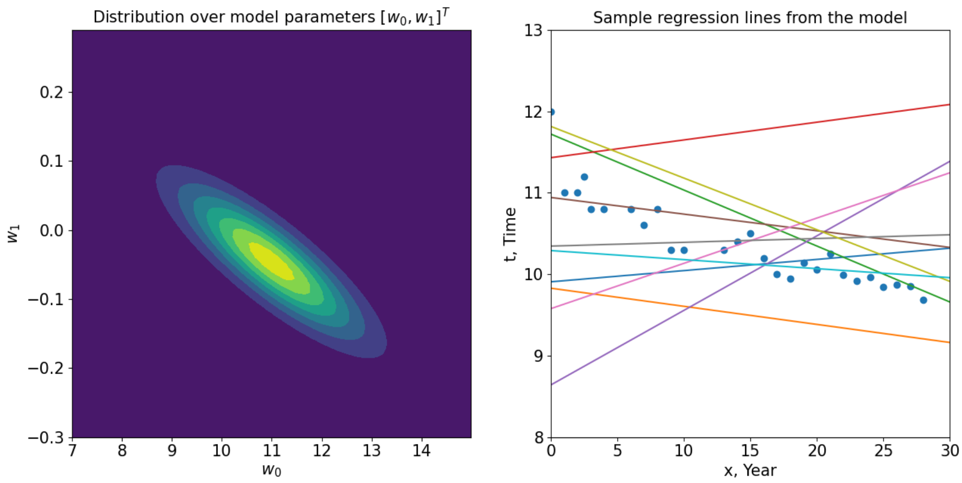

Figure 2 – Posterior visualization

In the visualization model function, ten samples are made randomly, so the lines on the graph can appear to be chaotic.