Exercise 1: Maximum-Likelihood Estimation (5 + 5 + 5 + 5 P)

We consider the problem of estimating using the maximum-likelihood approach the parameters $\lambda, \eta > 0$ of the probability distribution:

$$p(x, y) = \lambda\eta e^{-\lambda x - \eta y}$$

supported on $\mathbb{R}_+^2$. We consider a dataset $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N))$ composed of N independent draws from this distribution.

(a) *Show* that x and y are independent.

$$p(x,y) = \lambda e^{-\lambda x} \cdot y e^{-yy} = p(x) \cdot p(y)$$

where $p(x) = \lambda e^{-\lambda x}$, $p(y) = y e^{-yy}$

(b) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$.

$$\log p(\mathcal{D}|\lambda) = \log \prod_{i=1}^{N} \lambda y e^{-\lambda x_i - y y_i}$$

$$= \sum_{i=1}^{N} \log\lambda + \log y - \lambda x_i - y y_i$$

$$= N(\log\lambda + \log y) - \sum_{i=1}^{N} (\lambda x_i + y y_i)$$

$$\nabla_\theta \log p(\mathcal{D}|\lambda) = N \cdot \frac{1}{\lambda} - \sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \lambda = \frac{N}{\sum_{i=1}^{N} x_i}$$

(c) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1/\lambda$.

$$\log p(\mathcal{D}|\lambda) = N(\log\lambda - \log\lambda) - \sum_{i=1}^{N} (\lambda x_i + \frac{1}{\lambda} y_i)$$

$$\nabla_\theta \log p(\mathcal{D}|\lambda) = -\sum_{i=1}^{N} (x_i - \frac{y_i}{\lambda^2}) = 0$$

$$\Rightarrow \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \frac{y_i}{\lambda^2}$$

$$\Rightarrow \lambda^2 = \sum_{i=1}^{N} \frac{y_i}{x_i}$$

$$\Rightarrow \lambda = \sqrt{\sum_{i=1}^{N} \frac{y_i}{x_i}}$$

(d) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1 - \lambda$.

$$\log p(\mathcal{D}|\lambda) = N(\log\lambda + \log y) - \sum_{i=1}^{N}(\lambda x_i + y y_i)$$

$$= N\left(\log\lambda + \log(1-\lambda)\right) - \sum_{i=1}^{N}(\lambda x_i + y_i - \lambda y_i)$$

$$\nabla_\theta \log p(\mathcal{D}|\lambda) = N\left(\frac{1}{\lambda} - \frac{1}{1-\lambda}\right) - \sum_{i=1}^{N}(x_i - y_i) = 0$$

$$\Rightarrow \quad \frac{1-2\lambda}{\lambda-\lambda^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i) = t$$

$$\Rightarrow \quad t\lambda^2 - (2+t)\lambda + 1 = 0$$

$$\Rightarrow \quad \lambda = \frac{2+t+\sqrt{t^2+4}}{2t}$$

Exercise 2: Maximum Likelihood vs. Bayes (5 + 10 + 15 P)

An unfair coin is tossed seven times and the event (head or tail) is recorded at each iteration. The observed sequence of events is

$$\mathcal{D} = (x_1, x_2, \ldots, x_7) = (\text{head}, \text{head}, \text{tail}, \text{tail}, \text{head}, \text{head}, \text{head}).$$

We assume that all tosses $x_1, x_2, \ldots$ have been generated independently following the Bernoulli probability distribution

$$P(x \mid \theta) = \begin{cases} \theta & \text{if} \quad x = \text{head} \\ 1-\theta & \text{if} \quad x = \text{tail}, \end{cases}$$

where $\theta \in [0, 1]$ is an unknown parameter.

(a) *State* the likelihood function $P(\mathcal{D}|\theta)$, that depends on the parameter θ.

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{7} p(x_i|\theta) = \theta^5(1-\theta)^2$$

(b) *Compute* the maximum likelihood solution $\hat{\theta}$, and *evaluate* for this parameter the probability that the next two tosses are "head", that is, evaluate $P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \hat{\theta})$.

$$\log p(\mathcal{D}|\theta) = \log \theta^5(1-\theta)^2 = 5\cdot\log\theta + 2\log(1-\theta)$$

$$\nabla_\theta \log p(\mathcal{D}|\theta) = \frac{5}{\theta} - \frac{2}{1-\theta} = 0$$

$$\Rightarrow \quad \frac{5-5\theta-2\theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow \quad \theta = \frac{5}{7}$$

$$\Rightarrow P(X_8 = \text{head},\ X_9 = \text{head}\ |\ \theta) = P(X_8 = \text{head}\ |\ \theta) \cdot P(X_9 = \text{head}\ |\ \theta)$$
$$= \theta \cdot \theta$$
$$= \frac{25}{49} \pm 0.51$$

(c) We now adopt a Bayesian view on this problem, where we assume a prior distribution for the parameter θ defined as:

$$p(\theta) = \begin{cases} 1 & \text{if} \quad 0 \le \theta \le 1 \\ 0 & \text{else.} \end{cases}$$

Compute the posterior distribution $p(\theta|\mathcal{D})$, and *evaluate* the probability that the next two tosses are head, that is,

$$\int P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \theta)\, p(\theta|\mathcal{D})\, d\theta.$$

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{\int P(D|\theta) \cdot P(\theta)\, d\theta} = \frac{\theta^5 (1-\theta)^2}{\int_0^1 \theta^5 (1-\theta)^2\, d\theta}, \quad \text{for} \quad 0 \le \theta \le 1$$

$$= \frac{\theta^5 (1-\theta)^2}{\int_0^1 \theta^5 - 2\theta^6 + \theta^7\, d\theta} = \frac{\theta^5 (1-\theta)^2}{\left[\frac{1}{6}\theta^6 - \frac{2}{7}\theta^7 + \frac{1}{8}\theta^8\right]_0^1} = 168\, \theta^5 (1-\theta)^2$$

$$P(X_8 = \text{head},\ X_9 = \text{head}\ |\ \theta, D) = \int P(X_8 = \text{head},\ X_9 = \text{head}\ |\ \theta) \cdot P(\theta|D)\, d\theta$$
$$= \int_0^1 \theta^2 \cdot 168\, \theta^5 (1-\theta)^2\, d\theta$$
$$= 168 \int_0^1 \theta^7 (1 - 2\theta + \theta^2)\, d\theta$$
$$= 168 \int_0^1 \theta^7 - 2\theta^8 + \theta^9\, d\theta$$
$$= 168 \left[\frac{1}{8}\theta^8 - \frac{2}{9}\theta^9 + \frac{1}{10}\theta^{10}\right]_0^1$$
$$= \frac{168}{360} \doteq 0.467$$

Exercise 3: Convergence of Bayes Parameter Estimation (5 + 5 P)

We consider Section 3.4.1 of Duda et al., where the data is generated according to the univariate probability density $p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known and where μ is unknown with prior distribution $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Having sampled a dataset $\mathcal{D}$ from the data-generating distribution, the posterior probability distribution over the unknown parameter μ becomes $p(\mu|\mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, where

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

(a) *Show* that the variance of the posterior can be upper-bounded as $\sigma_n^2 \leq \min(\sigma^2/n\,,\, \sigma_0^2)$, that is, the variance of the posterior is contained both by the uncertainty of the data mean and of the prior.

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \Rightarrow \frac{1}{\sigma_n^2} \geq \max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)$$

$$\Rightarrow \quad \sigma_n^2 \leq \frac{1}{\max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)}$$

$$\Rightarrow \quad \sigma_n^2 \leq \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right)$$

(b) *Show* that the mean of the posterior can be lower- and upper-bounded as $\min(\hat{\mu}_n, \mu_0) \leq \mu_n \leq \max(\hat{\mu}_n, \mu_0)$, that is, the mean of the posterior distribution lies somewhere on the segment between the mean of the prior distribution and the sample mean.

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \leq \frac{n}{\sigma^2}\max(\hat{\mu}_n, \mu_0) + \frac{1}{\sigma_0^2}\max(\hat{\mu}_n, \mu_0)$$

$$\leq \frac{1}{\sigma_n^2}\max(\hat{\mu}_n, \mu_0)$$

$$\Rightarrow \quad \mu_n \leq \max(\hat{\mu}_n, \mu_0)$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \geq \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\min(\hat{\mu}_n, \mu_0)$$

$$\geq \frac{1}{\sigma_n^2}\min(\hat{\mu}_n, \mu_0)$$

$$\Rightarrow \quad \mu_n \geq \min(\hat{\mu}_n, \mu_0)$$

$$\Rightarrow \quad \min(\hat{\mu}_n, \mu_0) \leq \mu_n \leq \max(\hat{\mu}_n, \mu_0)$$