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.

$$p(x,y) = p(x)p(y)$$

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

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

$$J(\lambda) = \log p(\mathcal{D} \mid \lambda, \eta)$$

$$= \sum_{k=1}^{N} \log(p(x_k, y_k \mid \lambda, \eta))$$

$$= \sum_k \log\lambda + \log\eta - \lambda x_k - \eta y_k$$

$$= N \cdot (\log\lambda + \log\eta - \lambda\bar{x} - \eta\bar{y})$$

$$\frac{\partial J}{\partial\lambda} = N \cdot \left(\frac{1}{\lambda} - \bar{x}\right) \overset{!}{=} 0 \implies \hat{\lambda} = \frac{1}{\bar{x}}$$

$$\bar{x} = \frac{1}{N}\sum_k x_k \Leftrightarrow \sum_k x_k = N \cdot \bar{x}$$

concave ✓

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

$$J(\lambda) = N\left(\cancel{\log\lambda} + \cancel{\log\frac{1}{\lambda}} - \lambda\bar{x} - \frac{1}{\lambda}\bar{y}\right)$$

$$= N\left(-\lambda\bar{x} - \frac{1}{\lambda}\bar{y}\right)$$

concave ✓

$$\frac{\partial J}{\partial\lambda} = -\bar{x} + \frac{1}{\lambda^2}\bar{y} \overset{!}{=} 0 \implies \hat{\lambda} = +\sqrt{\frac{\bar{y}}{\bar{x}}}$$

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

$$(\lambda > 0) \wedge (\underbrace{1-\lambda}_{\eta} > 0) \implies 0 < \lambda < 1$$

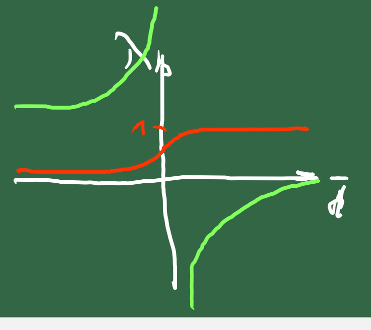
$$J(\lambda) = N \cdot (\log\lambda + \log(1-\lambda) - \lambda\bar{x} - (1-\lambda)\bar{y})$$

concave ✓

$$= N \cdot \left(\log(\lambda - \lambda^2) + \lambda\underbrace{(\bar{y} - \bar{x})}_{\bar{d}} + \text{c.}\right)$$

$$\frac{\partial J}{\partial\lambda} = N \cdot \left(\frac{1 - 2\lambda}{\lambda - \lambda^2} + \bar{d}\right) \overset{!}{=} 0$$

$$\hat{\lambda} = \frac{(\bar{d} - 2) \pm \sqrt{\bar{d}^2 + 4}}{2\bar{d}}$$



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 $P(\mathcal{D} \mid \theta)$, that depends on the parameter θ.

$$p(\mathcal{D} \mid \theta) = \prod_{k=1}^{7} p(x_k \mid \theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta = \theta^5 \cdot (1-\theta)^2$$

(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 $P(x_8 = \text{head}\,,\; x_9 = \text{head} \mid \hat{\theta})$.

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

concave ✓

$$\frac{\partial \log p(\mathcal{D} \mid \theta)}{\partial\theta} = \frac{5}{\theta} - \frac{2}{1-\theta} \overset{!}{=} 0 \implies \hat{\theta} = \frac{5}{7}$$

$$p(\text{head} \mid \hat{\theta}) \cdot p(\text{head} \mid \hat{\theta}) = \frac{5}{7} \cdot \frac{5}{7} = \frac{25}{49}$$

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

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

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

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

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{\int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta} = \frac{\theta^5 (1-\theta)^2 \cdot 1}{\underbrace{\int \theta^5 \cdot (1-\theta)^2 \cdot 1\, d\theta}_{1/168}} = 168 \cdot \theta^5 \cdot (1-\theta)^2$$

$$\int_0^1 \theta^2 \cdot 168 \cdot \theta^5 \cdot (1-\theta)^2\, d\theta = \frac{168}{360} = \frac{7}{15}$$

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 $p(x \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$, 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 $p(\mu \mid \mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, 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.

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

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

(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.

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

$$\leq \cancel{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} \cdot \max(\mu_0, \hat{\mu}_n)$$

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